HELLENIC REPUBLIC

UNIVERSITY OF IOANNINA

SCHOOL OF ENGINEERING

DEPARTMENT OF MATERIALS SCIENCE AND ENGINEERING

Analysis and processing of medical and other related big data

Vasileios C. Pezoulas

PhD Thesis

IOANNINA 2022

ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΕΠΙΣΤΗΜΗΣ ΥΛΙΚΩΝ

Επεξεργασία και ανάλυση ιατρικών και άλλων συναφών δεδομένων μεγάλου όγκου

Βασίλειος Χ. Πεζούλας

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΙΩΑΝΝΙΝΑ 2022

«Η έγκριση της διδακτορικής διατριβής από το Τμήμα Μηχανικών Επιστήμης Υλικών της Πολυτεχνικής Σχολής του Πανεπιστημίου Ιωαννίνων δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα Ν. 5343/32, άρθρο 202, παράγραφος 2».

Date of Application of Mr. Vasileios Pezoulas: 25/09/2017

Date of Appointment of PhD Advisory Committee: 03/10/2017

Members of the 3 Member Advisory Committee:

Thesis Advisor

Dimitrios Fotiadis, Professor, Department of Materials Science and Engineering, School of Engineering, University of Ioannina

Members

Athanasios Tzioufas, Professor, Department of Pathophysiology, School of Medicine, National and Kapodistrian University of Athens

Leonidas Gergidis, Associate Professor, Department of Materials Science and Engineering, School of Engineering, University of Ioannina

Date of Thesis Subject Definition: 03/10/2017

PhD Thesis Title: Analysis and processing of medical and other related big data

Date of Appointment of the 7-member Examination Committee: 27/07/2022

Dimitrios Fotiadis	Professor, Department of Materials Science and Engineering, School of Engineering, University of Ioannina
Leonidas Gergidis	Associate Professor, Department of Materials Science and Engineering, School of Engineering, University of Ioannina
Athanasios Tzioufas	Professor, Department of Pathophysiology, School of Medicine, National and Kapodistrian University of Athens
Andreas Goules	Assistant Professor, Department of Pathophysiology, School of Medicine, National and Kapodistrian University of Athens
Konstantinos Papaloukas	Professor, Department of Biological Application and Technology, School of Health Sciences, University of Ioannina
Themistoklis Exarchos	Assistant Professor, Department of Informatics, School of Informatics, Ionian University
Metin Akay	Professor, Department of Biomedical Engineering, University of Houston, USA

The PhD thesis is approved, with «EXCELLENT» on 30/08/2022

The Chairman of the Department **The Secretary of the Department**

Apostolos Avgeropoulos
Professor

Maria Kontou

Ημερομηνία Αίτησης του κ. Βασίλη Πεζούλα: 25/09/2017

Ημερομηνία Ορισμού Τριμελούς Συμβουλευτικής Επιτροπής: 03/10/2017

Μέλη Τριμελούς Συμβουλευτικής Επιτροπής:

Επιβλέπων

Δημήτριος Φωτιάδης, Καθηγητής, Τμήμα Μηχανικών Επιστήμης Υλικών, Πολυτεχνική Σχολή, Πανεπιστήμιο Ιωαννίνων

Μέλη

Αθανάσιος Τζιούφας, Καθηγητής, Κλινική Παθολογικής Φυσιολογίας, Ιατρική Σχολή, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

Λεωνίδας Γεργίδης, Αναπληρωτής Καθηγητής, Τμήμα Μηχανικών Επιστήμης Υλικών, Πολυτεχνική Σχολή, Πανεπιστήμιο Ιωαννίνων

Ημερομηνία Ορισμού Θέματος: 03/10/2017

Θέμα Διατριβής: Επεξεργασία και ανάλυση ιατρικών και άλλων συναφών δεδομένων μεγάλου όγκου

Διορισμός Επταμελούς Εξεταστικής Επιτροπής: 27/07/2022

Δημήτριος Φωτιάδης	Καθηγητής, Τμήμα Μηχανικών Επιστήμης Υλικών, Πολυτεχνική Σχολή, Πανεπιστήμιο Ιωαννίνων
Λεωνίδας Γεργίδης	Αναπληρωτής Καθηγητής, Τμήμα Μηχανικών Επιστήμης Υλικών, Πολυτεχνική Σχολή, Πανεπιστήμιο Ιωαννίνων
Αθανάσιος Τζιούφας	Καθηγητής, Κλινική Παθολογικής Φυσιολογίας, Ιατρική Σχολή, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
Ανδρέας Γουλές	Επίκουρος Καθηγητής, Κλινική Παθολογικής Φυσιολογίας, Ιατρική Σχολή, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
Κωνσταντίνος Παπαλουκάς	Καθηγητής, Τμήμα Βιολογικών Εφαρμογών & Τεχνολογιών, Σχολή Επιστημών Υγείας, Πανεπιστήμιο Ιωαννίνων
Θεμιστοκλής Έξαρχος	Επίκουρος Καθηγητής, Τμήμα Πληροφορικής, Σχολή Επιστήμης της Πληροφορίας & Πληροφορικής, Ιόνιο Πανεπιστήμιο
Metin Akay	Professor, Department of Biomedical Engineering, University of Houston, USA

Έγκριση Διδακτορικής Διατριβής με βαθμό «ΑΡΙΣΤΑ» στις 30/08/2022

Ο Πρόεδρος του Τμήματος **Η Γραμματέας του Τμήματος**

Απόστολος Αυγερόπουλος
Καθηγητής

Μαρία Κόντου

Dedication

To my family

Acknowledgements

First of all, I would like to cordially thank and express my sincere gratitude to my supervisor, Professor Dimitrios I. Fotiadis, for all his support and guidance he has given to me throughout my PhD studies. His mentoring inspired me to advance and widen my research interests and allowed me to grow as a data scientist. I would also like to thank the members of my advisory committee, Associate Professor Leonidas Gergidis and Professor Athanasios Tzioufas, for their continuous support and inspiration to advancing my background in machine learning and applied mathematics.

I would like to express my sincere thanks to Assistant Professor Andreas Goules and Professor Konstantinos Papaloukas for their excellent guidance in conducting the final research step in this PhD thesis. I would also like to express my sincere appreciation to Professor Metin Akay for his participation in my examination committee.

I would like to cordially thank Assistant Professor Themis Exarchos for his great assistance and the excellent collaboration we had throughout these years. I would also like to warmly thank all those colleagues and friends who supported me from the Unit of Medical Technology and Intelligent Information Systems for the harmonious cooperation we had, their great support and invaluable advice.

Last but not least, I would like to express my deepest gratitude and thanks to my parents and family for their constant unending support and encouragement throughout my PhD journey, and all of my friends who helped me in order to complete this thesis.

This research was part funded by the European Commission (HarmonicSS Project - GA: 731944, SILICOFCM Project - GA: 777204, TO_AITION Project - GA: 848146, IMMUNAID Project - GA: 779295), as well as, through the **Hellenic Foundation for Research and Innovation (HFRI)** under the HFRI PhD Fellowship grant (Fellowship Number: 1357).

Table of Contents

List of Figures

List of Tables

List of Supplementary Figures

List of Supplementary Tables

List of Abbreviations

Acronym	Description
AD	Autoimmune diseases
AI	Artificial Intelligence
ANN	Artificial Neural Networks
API	Application Program Interface
BGMM	Bayesian Gaussian Mixture Models
BGMM-OCE	BGMM with Optimal Components Estimation
BN	Bayesian Networks
CCE	Cloud Computing Engine
CCN	Central Computing Node
CD	Correlation Difference
CN	Computing Node
cV	coefficient of Variation
DART	Multiple additive regression trees with dropout rates
FAIR	Findable, Accessible, Interoperable, Reusable guiding principles
FHBF	Federated Hybrid Boosted Forests
FHIR	Fast Healthcare Interoperability Resources
GBT	Gradient Boosting Trees
GDPR	General Data Protection Regulation
GEE	Gaussian Elliptic Envelopes
GMM	Gaussian Mixture Models
GOF	Goodness of fit
HCM	Hypertrophic Cardiomyopathy
HIPAA	Health Insurance Portability and Accountability Act
HL7	Health Level 7
IQR	Interquartile range
IF	Isolation Forests
KL	Kullback-Leibler
LOF	Local Outlier Factor

LSA	Latent Semantic Analysis
MART	Multiple Additive Regression Trees
ML	Machine Learning
MLP	Multi-layer Perceptron
MNB	Multinomial Naïve Bayes
MVND	Multivariate Normal Distribution
OCVD	Ontology for cardiovascular diseases
OHDSI	Observational Health Data Sciences and Informatics
OMD	Ontology for mental disorders
pSS	Primary Sjogren's Syndrome
RBF	Radial Basis Function
REST	REpresentational State Transfer
SAIDs	Systemic Autoinflammatory Diseases
SGD	Stochastic Gradient Descent
SHAP	Shapley additive explanation analysis
SOMs	Self-Organizing Maps
SSAD	Standardized Squared Absolute Difference
STE	Supervised Tree Ensembles
SVM	Support Vector Machines
UTE	Unsupervised Tree Ensembles
XAI	eXplainable Artificial Intelligence

Abstract

The reduced quality and the increased structural and conceptual heterogeneity of the clinical databases combined with the presence of data silos obscure the sharing and analysis of medical data. These open issues in healthcare leverage the development and secure deployment of robust and unbiased AI (Artificial Intelligence) workflows to address clinical unmet needs, including: (i) the development of robust disease classification and risk stratification models, (ii) the detection of new biomarkers, and (iii) the discovery of targeted therapies, among others. In this thesis, we aim to address the open issues and unmet needs in healthcare through the development of beyond the state of the art methods which are built on top of four main innovation areas: (i) Innovation Area 1 - data curation, where we propose a fully automated, efficient and scalable medical data curation workflow to enhance the quality of the diverse medical data including clinical and genetic data across multiple time-points, (ii) Innovation Area 2 - data harmonization, where we propose a hybrid, fully automated data harmonization workflow combining lexical and semantic analysis based on word embeddings which is built on top of external knowledge bases to overcome structural heterogeneities across clinical databases, (iii) Innovation Area 3 - synthetic data generation, where we propose a large-scale synthetic data generator to significantly enhance the statistical power of clinical databases with insufficient population size in order to enable the simulation of clinical trials, as well as, to enhance the classification performance of the existing AI models through data augmentation, and (iv) Innovation Area 4 – federated/distributed learning, where we propose a federated AI deployment framework which removes the need for the installation of local servers or any type of software in each site through the adoption of a federated AI modeling engine supporting a large family of federated AI algorithms yielding interpretable and explainable AI models. The proposed four stage workflow was evaluated across six different clinical domains, including autoimmune diseases (AD) and particularly in primary Sjogren's Syndrome (pSS), hypertrophic cardiomyopathy (HCM), cardiovascular diseases (CVD), mental disorders (MD), systemic autoinflammatory diseases (SAIDs), and particularly Kawasaki disease (KD), and Coronavirus disease (COVID-19). The applicability of the proposed workflow was successfully demonstrated by: (i) enhancing the quality of the

clinical and laboratory data in pSS, HCM, COVID-19, CVD, MD, KD, (ii) reducing the levels of structural and conceptual heterogeneity among the clinical and laboratory data in pSS, CVD, MD and at the same time enabling the evaluation of cross-domain data harmonization, (iii) producing high quality and large scale synthetic data for *in silico* clinical trials in HCM, (iv) augmenting the existing lymphoma classification models in pSS and HCM risk stratification models, and (v) producing robust AI models for lymphoma classification in pSS, the detection of biomarkers for lymphomagenesis, the detection of biomarkers for Kawasaki disease, HCM risk stratification, ICU admission and mortality classification in COVID-19.

Περίληψη

Η μειωμένη ποιότητα και η αυξημένη δομική και εννοιολογική ετερογένεια των κλινικών βάσεων δεδομένων παγκοσμίως σε συνδυασμό με την παρουσία silo δεδομένων δυσκολεύουν τον διαμοιρασμό, την διασύνδεση και την επικείμενη ανάλυση των ιατρικών δεδομένων. Αυτά τα ανοιχτά ζητήματα στον τομέα της υγείας αναδεικνύουν την ανάγκη τον σχεδιασμό και την ανάπτυξη ασφαλών και αμερόληπτων ροών εργασίας AI (Τεχνητή Νοημοσύνη) για την αντιμετώπιση κλινικών ανεκπλήρωτων αναγκών, όπως: (i) η ανάπτυξη ισχυρών μοντέλων ταξινόμησης ασθενειών και διαστρωμάτωσης κινδύνου, (ii) η ανίχνευση νέων βιοδεικτών, και (iii) η ανακάλυψη στοχευμένων θεραπειών, μεταξύ άλλων. Σε αυτή τη διατριβή, στοχεύουμε να αντιμετωπίσουμε τα ανοιχτά ζητήματα και τις ανεκπλήρωτες ανάγκες στον τομέα της υγείας μέσω της ανάπτυξης καινοτόμων μεθόδων και ροών εργασίας, οι οποίες δομήθηκαν γύρω από τέσσερις κύριους τομείς καινοτομίας: (i) Περιοχή Καινοτομίας 1 - Εξυγίανση δεδομένων (data curation), όπου προτείνουμε μια πλήρως αυτοματοποιημένη, αποτελεσματική και επεκτάσιμη ροή εργασιών εξυγίανσης των ιατρικών δεδομένων για τη βελτίωση της ποιότητας των ιατρικών δεδομένων, συμπεριλαμβανομένων των κλινικών και γενετικών δεδομένων σε πολλαπλά χρονικά σημεία, (ii) Τομέας Καινοτομίας 2 - εναρμόνιση δεδομένων (data harmonization), όπου προτείνουμε μια υβριδική και πλήρως αυτοματοποιημένη μέθοδο εναρμόνισης δεδομένων που συνδυάζει την λεκτική και την σημασιολογική ανάλυση βασισμένη σε ενσωματώσεις λέξεων, η οποία δομήθηκε γύρω από εξωτερικές βάσεις γνώσεων για να ξεπεραστούν οι δομικές και εννοιολογικές ετερογένειες σε κλινικές βάσεις δεδομένων, (iii) Τομέας Καινοτομίας 3 - παραγωγή συνθετικών δεδομένων (synthetic data generation), όπου προτείνουμε μια γεννήτρια μεγάλης κλίμακας συνθετικών δεδομένων με στόχο να ενισχύσει σημαντικά τη στατιστική ισχύ των κλινικών βάσεων δεδομένων με ανεπαρκές μέγεθος πληθυσμού, προκειμένου να καταστεί δυνατή η προσομοίωση κλινικών δοκιμών, καθώς και για τη βελτίωση της απόδοσης της ταξινόμησης των υφιστάμενων μοντέλων τεχνητής νοημοσύνης μέσω της επαύξησης δεδομένων και (iv) Τομέας Καινοτομίας 4 – κατανεμημένη μάθηση εντός και εκτός του νέφους (Federated/distributed learning), όπου προτείνουμε ένα πλαίσιο ανάπτυξης κατανεμημένων μοντέλων τεχνητής νοημοσύνης που καταργεί την ανάγκη

εγκατάστασης τοπικών διακομιστών και την εγκατάσταση οποιουδήποτε είδους λογισμικού σε κάθε silo δεδομένων μέσω της υιοθέτησης μιας κατανεμημένης μηχανής μοντελοποίησης AI που υποστηρίζει μια μεγάλη οικογένεια κατανεμημένων αλγορίθμων τεχνητής νοημοσύνης που παράγουν ερμηνεύσιμα και επεξηγήσιμα μοντέλα τεχνητής νοημοσύνης. Η προτεινόμενη μεθοδολογία τεσσάρων σταδίων αξιολογήθηκε σε έξι διαφορετικούς κλινικούς τομείς, συμπεριλαμβανομένων των αυτοάνοσων νοσημάτων (AD) και συγκεκριμένα στο πρωτοπαθές σύνδρομο Sjögren (pSS), την υπερτροφική μυοκαρδιοπάθεια (HCM), τις καρδιαγγειακές παθήσεις (CVD), τις ψυχικές διαταραχές (MD), τις συστημικές αυτοφλεγμονώδεις νόσους (SAIDs) και συγκεκριμένα της νόσου Kawasaki (KD) και τέλος του COVID-19. Η κλινική και τεχνική απήχηση της προτεινόμενης μεθοδολογίας αποδείχθηκε επιτυχής δεδομένου ότι οδήγησε: (i) στην βελτίωση της ποιότητας των κλινικών και εργαστηριακών δεδομένων στις ασθένειες pSS, HCM, COVID-19, CVD, MD, KD, (ii) στην μείωση των επιπέδων δομικής και εννοιολογικής ετερογένειας μεταξύ κλινικών και εργαστηριακών δεδομένα στις ασθένειες pSS, CVD, MD και ταυτόχρονα επιτρέποντας την αξιολόγηση της εναρμόνισης δεδομένων μεταξύ τομέων, (iii) στην παραγωγή συνθετικών δεδομένων υψηλής ποιότητας και μεγάλης κλίμακας για κλινικές δοκιμές πυριτίου στην HCM, (iv) στην βελτίωση της απόδοσης των υπαρχόντων μοντέλων ταξινόμησης λεμφώματος και διαστρωμάτωσης κινδύνου στις ασθένειες pSS και HCM μέσω της τεχνικής επαύξησης των δεδομένων, και (v) στην παραγωγή ισχυρών μοντέλων AI για ταξινόμηση λεμφώματος σε ασθενείς με pSS, ανίχνευση βιοδεικτών για λεμφογένεση σε ασθενείς με pSS, στην ανίχνευση βιοδεικτών για τη νόσο Kawasaki, στην διαστρωμάτωση κινδύνου σε ασθενείς με HCM, στην πρόβλεψη εισαγωγής ασθενών με COVID-19 στη ΜΕΘ και στην πρόβλεψη της θνησιμότητας αυτών.

CHAPTER 1. INTRODUCTION

1.1. Big data in healthcare

In our rapidly advancing technological area, the large volumes of accumulated data, on a daily basis, yield many benefits in different areas of our everyday lives including finance, medicine, and industry, among others [1]–[3]. These large-scale datasets are referred to as big data. The big data are characterized by four dimensions, namely the volume, the velocity, the veracity, and the variety [4]–[6]. The speed of the daily generated data, the amounts of collected data, the different types of collected data, and the biases which are introduced during the data collection process are the fundamental characteristics of the big data against the traditional datasets, which are only characterized by one dimension, i.e., their volume.

The big data in medicine can improve the patient care through the enhancement of the clinical decision-making process, as well as, enhance the statistical power of the clinical research studies yielding more accurate outcomes and powerful prediction models [1]–[6]. Furthermore, the big data can further enhance the development of effective patient stratification methods towards the identification of sensitive population subgroups, as well as, to provide better insights on large population groups towards the development of new public health policies and targeted therapeutic treatments.

According to Figure 1 there are many types of big data in medicine. These types of data vary from biosignals and medical images to laboratory tests and omics data. The biosignals are produced by the electrical activity that arises from the biological function of the organs in the human body. Examples of the most common types of biosignals include the Electrocardiogram (ECG) [7] which records the electrical activity as a result of the heart's depolarization and repolarization function, the Electroencephalogram (EEG) [8] which records the changes in the electrical activity as a result of the neural activation (i.e., the electrical field from the extracellular currents), along with the Magnetoencephalogram (MEG) [9] which measures the changes in the ensuing magnetic field (from the intracellular currents), the Electromyogram (EMG) [10] which records the changes in the electrical activity as a result of the muscles contraction, the Electrooculogram (EOG) [11] which records the corneo-retinal potential as a result of the eye movement, etc. The biosignals yield high temporal information regarding a disease's onset and progress, with numerous applications in medical conditions and diseases that vary from cognitive deficiencies, schizophrenia, to heart failure, and Parkinson's disease [12]–[15].

The medical images comprise another type of medical data with significant importance in clinical diagnosis and screening procedures. Computerized tomography (CT) [16] scans, and magnetic resonance imaging (MRI) [17] scans, can provide detailed insight on the anatomic and tissue characteristics of different body parts, yielding high spatial information, and are useful in the detection of malignancies and other disorders. Furthermore, the positron emission tomography (PET) [18] scans, the single-photon emission computerized tomography (SPECT) [19] scans, and the functional magnetic resonance imaging (fMRI) [20] scans provide additional information regarding the biological and physiological operations, i.e., the metabolic processes, at a molecular level, as well as, the brain activations under specific physical and mental tasks. Furthermore, ultrasound [21] and photoacoustic [22] images are fast, non-ionizing, real-time methods which are based on acoustic properties, having numerous applications in echocardiography, obstetric ultrasonography, intravascular ultrasonography, and duplex ultrasonography, among others. Spectroscopy-based methods, such as, the functional near-infrared spectroscopy (fNIRS) [23] can shed light into the measurement of the metabolic rate of oxygen consumption which indicates a neural activation, like the fMRI.

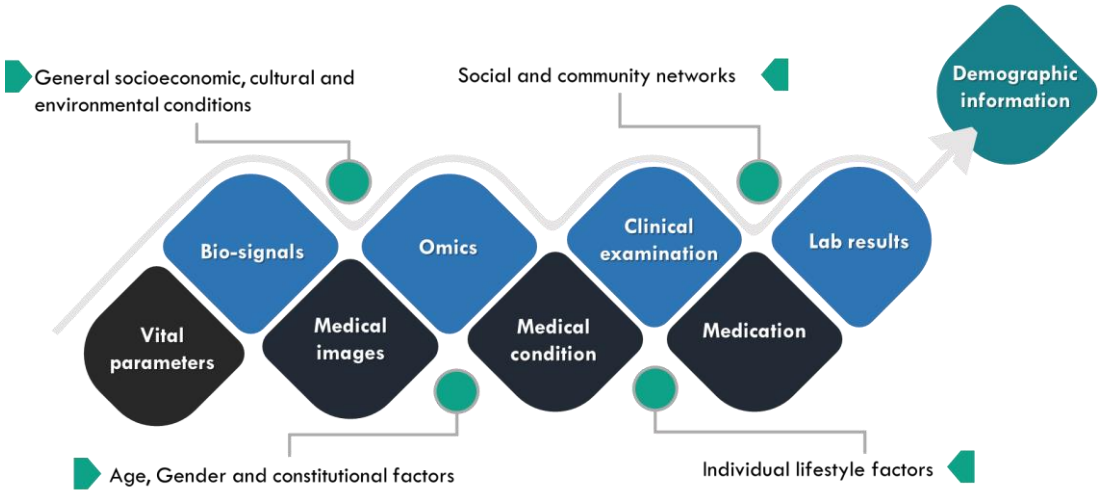


Figure 1. Big data in healthcare and related factors.

The field of omics constitutes another vast domain of medical data with numerous sub-fields, such as, the fields of genomics [24], lipidomics [25], proteomics [26], metabolomics [27], microbiomics, epigenomics [28], [29], and transcriptomics [30], among others. The omics data can be generated from high-throughput next-generation (NGS) technologies [31], such as, RNA-sequence analysis [32], mass spectrometry (MS) [25], [27], and thin layer chromatography (TLC) [33], which are able to analyze the proteins, the lipids, the transcriptomes, the metabolic profiles of the biological cells, the microorganisms in the tissues, pathological factors, and even the whole human genome. The RNA-sequence analyzers are able to capture all the single cell-based (or even group-based) RNA molecules (i.e., the whole transcriptome). In addition, the mass spectrometry technology is able to reveal the structural and functional characteristics of proteins, as well as, identify the lipids and their involvements in cell functionality. Omics can be used to study a variety of molecular-level functions, including the examination of bacteria and fungi on the tissues and organs, the interactions between the proteins, the detection of pathological factors and metabolic effects in degenerative and chronic diseases, gene expression analysis, among others.

The laboratory tests along with the medical claims and the subscribed medications can offer a powerful basis for understanding the underlying mechanisms of a virus and detecting various pathological conditions in tissue specimens. The most common laboratory tests include the hematological tests, the serological tests, the skin tests, the histopathological tests, the immunological tests, the endocrine function tests, and the coagulation tests, among others. Straightforward methods, such as, microscopic

analysis [34], fluoroscopy [35], immunocytochemistry [36], and immunohistochemistry [37], are used to analyze the tissue and blood samples. Each test offers a unique insight on a medical condition or a disease towards the detection of blood clotting disorders, tumors, anemia, diabetes, fungal infections, autoimmune disorders, skin cancer, allergies, inflammatory disorders, and endocrine dysfunctions, among many others.

The sources of medical data are many. With the growing number of large volumes of daily generated data from health sensors, medical images, laboratory tests, electronic patient records, patient registries, clinical and pharmaceutical claims, genome registries, the estimated amount of data is expected to overcome the zettabyte (10^{21} gigabytes) and even the yottabyte (10^{24} gigabytes) [38], [39]. The medical data acquisition process is often conducted according to international standards and protocols for each type of medical data. For example, in signal acquisition, well-known international standards are used for the placement of surface electrodes, such as, the 12-lead placement [40] for ECG signal acquisition, and the International "10-20" system (and "10-5" system) [41] for EEG signal acquisition. In laboratory tests, hemodynamic, coagulation, serological, and immunoassay analyzers are most commonly used for measuring biochemical (e.g., blood pressure, blood clotting time) and pathological factors (e.g., the presence of antigens in the antibodies), as well as, analyzing tissue specimens (e.g., for skin cancer, endocrine disorders), under different measurement units.

Medical image acquisition protocols are also used for the reconstruction of MRI, CT, fMRI, PET, and SPECT images, such as, the filtered backprojection (FBP) algorithm [42], and the family of the iterative reconstruction algorithms, such as, the algebraic reconstruction algorithm [43], and the iterative sparse asymptotic minimum variance (SAMV) algorithm [44], as well as, the universal backprojection algorithm [45] for photoacoustic imaging reconstruction, towards the examination of tissues and organs for tumors and other disorders. In the field of omics, standard methods, such as, the microscopic analysis [34], the RNA-sequencing analysis [32], the mass spectrometry (MS) [25], [27], the thin layer chromatography (TLC) [33], along with the high throughput next generation sequencing (NGS) technology [31], are widely used in omics to study the proteins interactions, the genetic profiles and metabolic effects of

different viruses, the lipids, the whole transcriptome, the genetic profiles of the human microbiome, among many others.

A research-oriented source of medical data with high clinical significance are the cohorts. Cohort studies are special types of observational studies [46] which are used to examine a disease's origins and the effects of the population characteristics. The longitudinal cohort studies are observational studies that involve the repetitive collection of patient data over long (or short) periods of time and are able to provide deeper insight on the disease progress over time with increased accuracy, overcoming recall biases [47]. In general, a cohort study can either use retrospective or prospective data. The retrospective cohort studies make use of data that have been already collected with the purpose of identifying the association between the causes (symptoms) and the disease's outcomes.

On the other hand, the temporal dimension that is introduced by the prospective cohort studies (i.e., the follow-up data) can reveal significant associations between the disease's outcomes and the causes of the disease, as well as, the effects of various prognostic factors on the outcomes, over time. The risk ratio and the hazard ratio are mainly used to quantify the associations between the drug exposure and the outcomes, as well as, the frequency of death, as a ratio between the exposed group and reference (or control) group [48], [49]. The former includes the subjects that are exposed on a specific drug whereas the latter consists of healthy individuals. Cohort studies are able to overcome several limitations that are present in traditional clinical trial studies by: (i) measuring patient-specific outcomes from large population groups, (ii) keeping track of follow-up patient data, and (iii) being less-expensive that large-scale clinical trials [50]. An example of the clinical importance of a cohort study lies on the fact that it can address the unmet needs in the special case where the exposure is a rare condition, such as, an autoimmune disease. In practice, a well-designed cohort study can provide deep insight into the underlying mechanisms of a disease's onset and progress.

1.2. Types and sources of big medical data

In medical research, cohort, case-control, and cross-sectional studies are three special types of observational studies (Figure 2) [51]. A clinical cohort study is comprised of data from a group of people that share common disease occurrences, medical conditions

(e.g., experience a common type of a chronic disease) and are useful for measuring the disease occurrence and progress [51]. A cohort study design can be either prospective or retrospective. In a prospective study, the cohort data are expected to be updated within the duration of the study whereas in a retrospective study, the patient data are predefined. In prospective studies, the existence of individual follow-up time points is necessary to keep track of the upcoming data. On the other hand, cross-sectional studies measure the disease occurrence at one time point and thus are not able to capture the relationship between the occurrence and the progress of a disease. To understand the meaning of a cohort it is necessary to understand the fundamental types and sources of medical data.

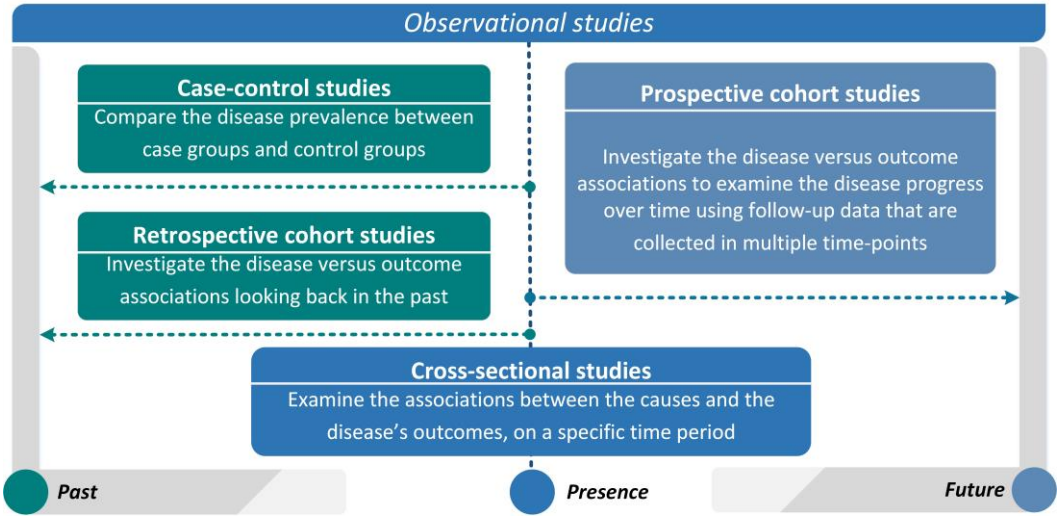


Figure 2. The fundamental types of observational studies [52].

According to Figure 3 sources of medical data are many. Laboratory results comprise a widely known source of medical data. Laboratory tests include a large number of biochemical tests [53], such as, (i) hematological tests which measure the oxygen levels in the blood flow, urine tests which are usually used to detect kidney, liver disease and diabetes, (ii) serological tests which are blood tests that seek for antibodies (e.g., to detect rubella, fungal infections), (iii) coagulation tests which are used to detect thrombophilia and hemophilia, (iv) histological tests which are employed in order to examine different types of tissues (e.g., muscle, nervous, epithelial), etc. Laboratory results combined with valuable information from medical conditions and medications can offer a powerful basis for: (i) understanding the progress of a disease, (ii) dividing sensitive populations into sub-groups (i.e., patient stratification), and (iii) evaluating existing and/or proposing new treatments, in large scale population studies. Other

common parameters that can often be found in clinical datasets include demographic information (e.g., age, gender, socioeconomic factors), vital parameters (e.g., heart rate, blood pressure), medications (e.g., antibiotics, antiseptics) and medical conditions (e.g., Alzheimer, Parkinson), physical and mental conditions, nutrition habits, environmental and lifestyle factors [54], among others.

Other sources of medical data include medical images which are obtained by a variety of diagnostic imaging modalities or systems, such as, computed tomography, magnetic resonance, optical topography, ultrasound, positron emission tomography, single-photon emission computed tomography, etc. Advances in surface-rendering and volume-rendering methods have led to 3D medical image visualization which has significantly improved the quality of image interpretation. Moreover, the rapidly increasing spatial resolution of such systems combined with the technical advances in medical image processing (e.g., reconstruction, fusion) can significantly enhance the diagnostic accuracy and the consistency of the image interpretation by doctors in a variety of diseases ranging from heart failure, osteoporosis and diabetes to Alzheimer's disease and cancer [55]. Undoubtedly, computer-aided diagnosis is one of the major computer-assisted technologies for medical diagnostics.

Biosignals comprise another domain of medical data including a variety of biomedical signals, such as, (i) electroencephalography (EEG) and (iii) electrocorticography (ECoG) which capture the electrical fields that are produced by the activity of the brain cells, (ii) magnetoencephalography (MEG) which captures the magnetic fields that are produced by the electrical activity of the brain cells, (iv) electrocardiography (ECG) which records the electrical activity that arises from the depolarization and repolarization activity of the heart, (v) electromyography (EMG) which records the electric potential that is generated by the muscle cells, (vi) electrooculography (EOG) which records the electric potential generated by the cornea and the retinal activity, etc. Biosignals provide high temporal information about a disease's onset and progress and have been employed in a variety of diseases ranging from epilepsy, schizophrenia, to heart failure and muscle atrophy [56]–[58]. Biomedical signals are usually combined with medical imaging systems (e.g., EEG and MRI) to provide both high spatial and temporal information for more effective diagnosis and treatment. The advances in biomedical signal processing have made signal manipulation much easier.

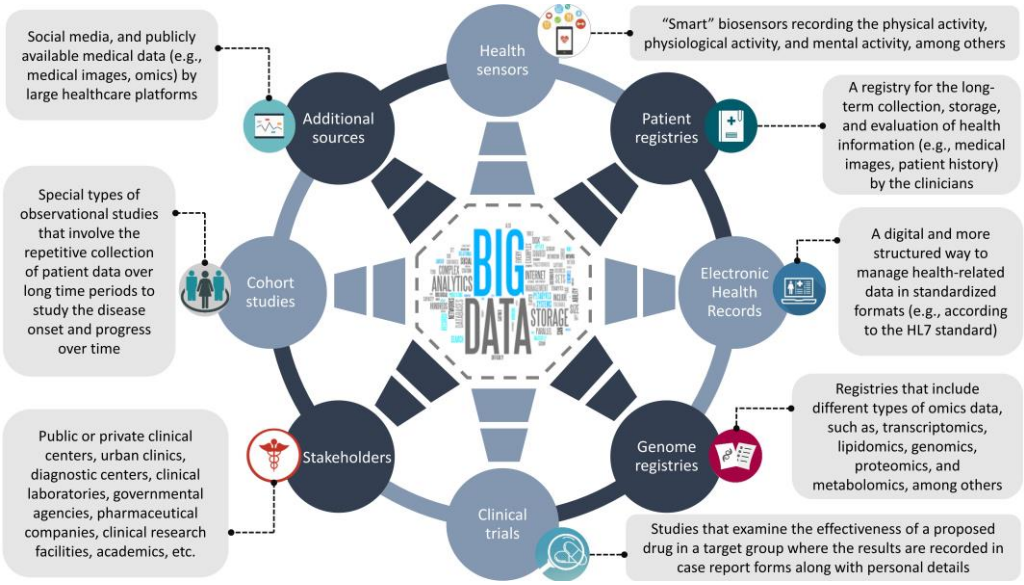


Figure 3. Sources of big medical data *[52]*.

The field of genetics constitutes a vast domain of medical data. Genetic data can be generated from high-throughput (next-generation) DNA and RNA sequences. The outrageous number of these sequences has created the well-known field of genomics. Genetic data are generally of more complex form than the types of medical data since they require the use of multiple processing pipelines with unique input. This complexity arises from the different formats of the genetic data, such as, the fastq files used for RNA sequence analysis, the haplotypes for haplotype analysis, etc. In the last decade, genetic data generated from genome-wide association studies (GWAS) have led to thousands of robust associations between common single-nucleotide polymorphisms (SNPs) and common diseases ranging from auto-immune diseases to psychiatric disorders, quantitative traits, genomic traits [59].

The recent advances in omics technologies [60], [61], such as, genomics (the study of genomic information), transcriptomics (the study of all the RNA transcripts of an organism), proteomics (the study of proteins and their interactions), lipidomics (the study of lipids, i.e., biomolecules with structural diversity and complexity), and metabolomics (the study of the multitude of metabolites) has increased the demand for properly annotated and well-preserved biospecimens, which has led to the development of the biobanks . Biobanking involves the: (i) collection, (ii) processing, (iii) storage, and (iv) quality control of the biological samples along with their associated clinical information [62]. Biobanks have been widely used for meeting scientific goals in genetic and molecular biology due to their long-term sustainability [62].

1.3. Big data is about sharing

Why is medical data sharing so important? Imagine what you could do in a medical area if you had access to almost all the medical data in this area. To answer this question in a realistic way, we will focus on presenting four clinical needs which have been identified as of great importance in several cohort studies: (i) patient stratification, (ii) identification of new biomarkers and/or validation of existing ones, (iii) new therapy treatments, and (iv) development of new health policies. Each of these needs highlights the necessity of medical data sharing in promoting research worldwide. As it was already mentioned in 1.2, cohort studies can resolve crucial scientific questions related to the predictive modelling of a disease's onset and progress, the clinical significance of genetic variants, and the adequate identification of high-risk individuals. Although data sharing is valuable for the public, the ignorance of knowing what the denominators and the requirements are in a study, leads to contradictory findings.

Medical data sharing involves all those mechanisms concerning the protection of patient's rights and privacy. It comprises the core of a federated platform since it enables the interlinking of medical cohorts worldwide [52], [63]. A data sharing framework is responsible for two major functionalities: (i) the assessment of whether the data origin and acquisition, as well as, the processes that are undertaken in a federated platform fulfill the guidelines posed by the corresponding data protection regulations (i.e., the legal aspect), and (ii) the assessment of the quality and completeness of medical data (i.e., the data quality aspect) by taking into consideration existing clinical domain knowledge and related public health policies. The latter is usually referred to as data governance and is related to: (i) the evaluation of data quality metrics, (ii) the inspection of the data organizational structure, and (iii) the overall information management [64].

The data sharing framework constitutes the primary stage prior to the development and application of the federated data analytics services. From a legal point of view, a clinical center that wishes to share clinical data to a federated platform must provide all the necessary ethical and legal documents, prior to any further data manipulation. These documents depend on the data protection regulations posed by each party (e.g., according to the GDPR guidelines in Europe or the HIPAA guidelines in USA) and usually include: (i) precise definition of legitimate interests, (ii) complete data

protection impact assessments, (iii) exact purposes of processing, (iv) signed consent forms for the processing of personal data from the data subjects, (v) purposes of transferring to third parties, (vi) data protection guarantees, and (vii) notifications to the data subject about the processing, among many others.

A federated platform that is responsible for data sharing must first provide complete definitions for the primary data collectors and the secondary analysts. Informed consent forms for pooled data analysis are also necessary for data analysis through a process which is currently referred to as "handshaking" [65]. Ethical issues for data collection introduced by different countries inside and outside the EU must also be taken into consideration. Moreover, the fear for data abuse and losing the control of the data is a crucial barrier towards data sharing. Secure data management and data de-identification is thus mandatory for privacy preserving to enable the sharing of sensitive data.

From the data quality point of view, under the data governance part of the data sharing framework lies a fundamental procedure, known as data quality assessment [66]–[70], which aims to improve the quality of the data in terms of consistency, accuracy, relevance, completeness, etc. Data cleansing [71], [72], also referred to as data curation, is a multidisciplinary process which comprises the core of the data quality assessment procedure and deals with duplicate fields, outliers, compatibility issues, missing values, etc., within a raw clinical dataset. Nowadays, automated data curation is a crucial technical challenge for data analysts and researchers worldwide who wish to manage and clean huge amounts of data.

For this reason, emphasis must be given on the development of tools for realizing such a concept. In addition, it is important to define a common format for the clinical datasets, i.e., a template of pre-defined variables, data ranges, and types, for a specific domain, which serves as a model that can be used to develop rules for: (i) matching variables across heterogeneous datasets, and (ii) normalizing them where necessary. The former is an intermediate step of data harmonization [52], [73], [74] and the latter is known as data standardization [75] (CHAPTER 4).

Several data sharing initiatives have been launched towards the integrity of clinical research data [76]–[79]. These initiatives aim at providing frameworks and guidelines for sharing medical and other related data. They mainly focus on the transparency of

the data collection protocols and the patients' de-identification process to promote medical research worldwide. Most of these initiatives develop publicly available gateways in the form of data warehouses, which host data from thousands of highly qualified clinical studies worldwide, including prospective and retrospective data from clinical trials, case-studies, genome-wide association studies, etc., with the purpose of providing access to large amounts of data for scientific purposes. Powerful cloud-based systems have been launched, with all the processes (registration, de-identification, quality control) being conducted automatically through the web. Thus, the meaningful interpretation of the outcomes of studies that make use of such data is reassured due to the increased statistical power they offer. Centralized patient databases, however, are often prone to data breach and sometimes unable to comply with data protection regulations [80].

A promising solution to this can be accomplished using multisite databases which serve as remote data warehouses that communicate in a distributed manner, giving emphasis on the "*sharing of information from the data instead of sharing the data themselves"* [52], [80], [81]. This approach overcomes several data sharing barriers discussed previously since the fear for data abuse can be controlled through distributed firewalls and individual data monitoring mechanisms. Moreover, the need to transfer sensitive data is nullified since an individual researcher can work independently on each site through coordinating systems that distribute the commands per site.

A federated platform should take into consideration several technical challenges. Treating patients with respect is a key-factor towards its establishment. Emphasis must also be given to the cost scalability over security which is a crucial trade-off, as well as, on software and copyright licenses for all the tools that will be employed in the platform. Big data monitoring, validation, storage, multidimensional interoperability (legal, regulatory) are a few examples of such challenges.

1.4. Open issues and unmet needs in healthcare

Under the ages of our rapidly advancing technological era, the vast amount of daily generated digital data has led to a scientific breakthrough with huge benefits in many fields of our everyday lives including finance, medicine, and industry [4], [82]–[87]. The term "big data" has been extensively used to characterize these massively

accumulated data sets which are mainly characterized by the four-V's [4], [82]–[87]: (i) volume, (ii) velocity, (iii) veracity, and (iv) variety, where each dimension has a unique scope. The "volume" dimension refers to the massive amounts of collected data elements whereas the "velocity" refers to the speed of the continuously generated data flows. The "veracity" dimension refers to the biases that are introduced during the data collection process and the "variety" refers to the different types of data sources (e.g., different formats, structural differences). The definition of big data overcomes the classic definition of the ordinary datasets, which is only limited to their size (i.e., volume). Big data is a promising tool that provides broader and more comprehensive insight from large data elements, a fact that greatly enhances their impact in various scientific and research areas, especially in healthcare. However, the size of the collected data and the speed of the data generation process, combined with the different types and the complexity of the big data are crucial challenges that still need to be addressed by the scientific community.

Currently, there are many types of big data in healthcare. More specifically, medical big data can be found in the medical imaging domain, where the thin-slice technology that has already been adopted by the modern diagnostic imaging scanners (e.g., CT, MRI, OCT) is able to capture thousands (>2000 slices) of high-quality (in terms of spatial and temporal resolution) slices of different body parts, in a very small amount of time. In addition, in the field of genomic analysis, the advances in high-throughput sequencing (HTS) technology has led to the next-generation or second-generation sequencing (NGS) technology which is able to capture the entire human genome (which consists of 30000 to 35000 genes), producing millions of DNA and RNA sequences. Moreover, in the field of biomedical signal analysis, the continuous, high-resolution, monitoring (e.g., for days or even weeks) of a patient's physiological, mental, or physical activity produces large amounts of recorded waveforms which consist of thousands of samples per second.

The application of big data in healthcare is promising and with many benefits. The vast amount of generated data can improve the statistical power of the conventional methods for data analytics including data mining and predictive modeling. Furthermore, the big data can boost the clinical decision-making process and yield clinical outcomes with higher statistical power (i.e., higher scientific impact due to the large number of

participating subjects) and improved accuracy [4], [82]–[87]. As a result, the patient care will be greatly improved since the patients will avoid the risk of unnecessary surgery operations, as well as, the negative implications of unnecessary (or even false) drug administration.

Big data can enhance the performance of the conventional machine learning methods and workflows, giving rise to a field that is known in computer science as deep learning [1-6] ˙ a modern technique that makes use of multi-layer neural networks that are able to capture valuable patterns and associations that are hidden between the large data elements, such as, in multi-slice medical images, omics, biomedical signals, etc. For example, the high-resolution, multi-dimensional, four-dimensional (4D) medical images, such as, the PET images which can capture additional information regarding the metabolic effects of a radioisotope apart from the anatomic structure, or the fMRI images which can depict the brain activations under a specific physical activity, can greatly enhance the accuracy of clinical diagnosis.

The application of big data in medicine can also enhance the patient stratification process according to which straightforward machine learning methods can be applied to identify and discriminate high-risk individuals from large populations [4], [82]–[87], i.e., groups of patients having high-risk for the development of a type of malignancy, such as, lymphoma. This will also yield significant improvements in personalized medicine for the selection of appropriate therapies by taking into consideration molecular-level health information. Furthermore, the multivariate analysis of large population groups can also reveal significant statistical associations between the disease's manifestations and different demographic factors, such as, age, gender, etc., and other medication related factors. Moreover, the outcomes from large-scale clinical trials and clinical research studies that make use of big data from large populations can enable the development of new, low-cost, targeted therapies for chronic and rare diseases, as well as, the development of new public health policies towards a global and sustainable healthcare system.

The software advancements towards the development of methods for big data analytics is an emerging field. The current software advancements in neuroimaging have led to the voxel-wise analysis of hundreds of thousands of voxels (>100000) within the human brain yielding large-scale similarity matrices, i.e., brain networks, which are able to

simulate the brain activations across different regions of interest (ROIs), yielding millions of connections between the voxels [88], [89]. These large-scale networks have been widely used to study the brain activation patterns during resting-state or under specific physical, as well as, mental tasks [90].

Furthermore, the analysis of omics big data using high-computing resources can reveal important clinical information concerning the genetic variants and cellular functionalities in different types of diseases, as well as, assist the development of effective drugs with reduced implications in the participating subjects. A great example can be found in the field of interactomics [91], where the protein-protein interaction (PPI) networks are constructed, on a cellular basis, to study the stable and transient interactions among proteins [92]. In addition, in biosignal analysis, the applications of deep learning methods for the prediction of disease outcomes have shown significant performance yielding high sensitivity and specificity scores in numerous cases, such as, the prediction of epileptic events [56].

Understanding big data is a difficult and demanding task for researchers and data analysts. With the growing number of large volumes of daily generated data from health sensors, social media posts, medical images, laboratory tests, electronic patient records, blogs, and web pages, the estimated amount of data is expected to overcome the zettabyte (10^{21} gigabytes) and even the yottabyte (10^{24} gigabytes) [38], [39]. Therefore, the development of straightforward software architectures along with hardware components and computer-aided tools and systems towards the efficient storage, management, quality assessment, high-performance computing analysis, and visualization of big data, is a constant and increasing demand. For example, in medical imaging, emphasis must be given to the development of methods for big data compression (e.g., image compression), registration and mapping of thousands of slices, and methods for segmentation of anatomical structures across these slices.

A scientific researcher who can understand the nature (e.g., the patterns) of big data can discover new opportunities for the development of new methods for big data analytics. There is no doubt that the benefits of big data in healthcare are many. However, there are several technical and clinical challenges that still need to be addressed. The main challenge is the fact that the sources of big data are disparate, heterogeneous, and costly, a fact that increases the computational complexity of handling large volumes of data,

as well as, hampers the application of traditional statistical and machine learning methods for big data analytics. In addition, the big data are often incomplete with several discrepancies due to the lack of a global protocol for big data acquisition. As a result, data standardization methods need to be adopted to overcome this structural heterogeneity. Moreover, the big data are difficult to manage due to their size and structural complexity. Furthermore, the risk of data misuse is increased in big data with the data quality assessment process being a significant challenge along with the lack of the researcher's skills that might hamper the quality of the data yielding unreliable outcomes. The big data are often prone to the existence of missing values and measurement errors throughout their context which pose significant obstacles towards their effective analysis. As a result, the irrational use of machine learning methods for predictive modeling in large datasets might lead to false outcomes, with no clinical importance at all.

There are also privacy issues that lurk behind the use of big data [38], [80]. Ethical and legal issues must be carefully taken into consideration during the collection and processing of big medical data from multiple data sources. Since the big data are large collections of patient data, it is difficult and even impossible to obtain signed informed consent forms from every single patient. In addition, the large volume of medical data shall not be stored in centralized databases since the risk for data abuse is greatly increased. Therefore, the data should be stored in cloud environments which are compliant with data protection regulations and should be collected under appropriate data protection agreements based on international privacy and protection standards. The researchers and data analysts must be fully aware of the data protection regulations during the collection and processing of the data. Furthermore, there is an increased necessity towards the development of machine learning methods for analyzing data that are distributed in multiple sites, a fact that remains a great challenge (see CHAPTER 6 for methods that deal with data analysis across federated databases).

1.4.1. Data security and data protection

1.4.1.1. Legal and ethical barriers

The backbone of data governance lies on the legal and ethical compliance (with the requirements that are posed by the existing data protection legislation) that any entity

(e.g., an organization), which wishes to get involved in the processing of personal data, must meet. A federated platform however faces significant ethical and legal compliance challenges which directly affect the privacy of personal data. Towards this direction, strict legal and ethical requirements [93]–[98] must be adopted by the data protection laws to ensure the privacy during the inner (i.e., within the entities of a country) and outer (i.e., between the entities of a country and the entities of a third country) personal data flows, including the following:

- The individuals' personal data must be processed with respect to the individual's rights and freedoms.
- Individual consent forms must be obtained by anyone who wishes to process personal data according to the purposes of processing.
- The individual must be informed about all types of processing which involve his/her personal data and provide his/her informed consent according to the consequences (i.e., the risks) that might arise because of the processing of his/her personal data.
- This comprises a strict requirement which involves the participation of the individuals in the processing of their data.
- The individuals must be given the right to: (i) access, rectify, and erase their personal data, (ii) object and restrict the processing of their data, and (iii) request to obtain their data when they wish to do so.
- The risks behind the processing of the individual data (i.e., risk assessment) must be clearly stated.
- Any cross-border data flows involving sensitive data must be subject to international legal requirements and data protection principles which require the co-operation of international supervised authorities.
- The sensitive data must not be transferred to third countries (parties) without the fulfillment of adequate data protection requirements and principles under the international data protection regulations.

- The existence of any third parties must be clearly defined in the related contracts and rules of conduct, along with any natural or legal person who is authorized to collect the data and further manipulate them.

All the entities must describe any measures to be taken to comply with the above requirements. In any other case, strict legal sanctions and ethical ramifications will be issued against such entities.

1.4.1.2. Patient privacy issues

The term personal data includes a variety of personal identifiers that can either directly or indirectly lead to the identification of the individual by any processing entity. Nowadays, the number of personal identifiers has been greatly increased due to the rapid digital advancements.

These personal identifies not only include, names, telephone numbers, license numbers and social security numbers but also email addresses, biometric identifiers, bank account numbers, Internet Protocol addresses and other unique digital identifiers [93]–[100]. The following privacy issues [93]–[98] shall be considered to protect the individual's identity:

- Personal data must be de-identified by either pseudonymizing or anonymizing them. Anonymization involves the complete removal of any information that can lead to the identification of the individual whereas pseudonymization involves the partial removal of the individual data with an additional storage of information that can indirectly lead to the identification of the individual (e.g., an identifier).

- Only a small portion of the individuals' data must be processed according to the purposes of processing.

- Common international standards and definitions must be introduced for the terms data anonymization and data pseudonymization to avoid any confusion during data collection and data processing.

- It must be clearly defined who is responsible for data collection (i.e., the primary data collectors) and data processing (i.e., the secondary analysts), as well as, the existence of any involved third parties.

- Researchers and analysts must be well-qualified with appropriate expertise in data protection to avoid data embezzlement and data misuse that might harm the individual.

- Audit trails are necessary so that the individuals can see who accessed their medical records. In the case of a patient privacy breach, the involved patients must be directly informed by the related authorities.

- All the data processing operations must be transparent and fair according to the individuals' rights.

- Strict data protection protocols are needed to avoid unauthorized surveillance and prevent data breach.

- Law enforcement agencies need to be involved in the out-of-border tracking of personal data flows.

1.4.1.3. Technical limitations

The technical limitations are like the technical challenges of a federated platform in the basis of data sharing and data protection [101], [102]. Those limitations include the following:

- Secure mechanisms for user access management and multiple-factor user authentication services.

- Secure and encrypted communication mechanisms for the collection and transmission of personal data.

- Secure private data layers within the cloud for the storage of personal data (in remote private spaces).

- Effective de-identification mechanisms through the construction of unique identifiers per patient.

- Efficient methods for reducing the information that is needed during the processing of personal data.

- The “bring the analysis to the data” design where the sensitive data are stored in remote private spaces.

- Batch-based processing mechanisms (i.e., distributed methods for data analytics), especially when the personal data are stored in de-centralized databases where secure communication is necessary.

- The data shall be made always available which is a fundamental principle of data sharing in federated platforms.

- Data availability, i.e., the reuse of personal data, promotes scientific research, worldwide.

- Automated error recovery mechanisms when the operating system fails to respond to any kind of functionality and especially when the operating system loses the control of the data, e.g., in the case of a data breach. In the latter case, supervised government authorities must be properly informed.

- Automated mechanisms for assessing and crosschecking the quality of the data (i.e., data curation). The data must be accurate, up-to-date, relevant, adequate, complete and in a readable form.

- Digital forms to upload personal consent forms through highly remote secure systems.

- Continuous monitoring of the data input and export processes along with the logging and processing.

- Scalability and interoperability of the federated platform that accounts for data sharing and international data protection regulations.

    - The interoperability factor includes legal, regulatory and application issues. The scalability factor involves efficient resource management (e.g., IT infrastructure).

    - Pooled (centralized) analysis must be supported only when informed consent forms are employed.

1.4.1.4. Other aspects

The significant challenges that a federated platform can face towards its compliance with the data protection regulations involves additional multidimensional aspects, including:

- The heterogeneity of the data protection laws across different countries, i.e., the existence of legal and ethical inequalities across developed and developing countries, as well as, ethical issues during the data collection process which is introduced by different countries.

- The heterogeneity of the data protection protocols across international laboratories and institutions. Different entities have different legal and ethical regulations regarding the data collection process.

- Additional bioethical regulations in the case of genome-wide studies must be taken into consideration. The health policies regarding the processing of genetic data are usually stricter and harder to follow.

- The negative implications of big data in privacy protection (e.g., the use of big data for the identification of individuals using information from the social media or any other information from the internet).

- The negative effect of centralized data warehouses in the case of data breach. It is easier for the hackers to breach centralized data repositories instead of distributed data repositories where the access to the rest of the repositories can be blocked in the case of data breach in a specific repository. On the other hand, distributed data repositories pose significant computational challenges.

- The existence of potential data obscuration/aggregation mechanisms in the form of malicious software. These mechanisms can be uploaded in the form of an ordinary software and cause serious leaks.

- The early-detection and prevention of personal data information leaks in large-scale platforms. Large-scale platforms might be hard to breach but a successful attempt can have serious consequences.

Ineffective tracking of entities which falsely claim to be compliant with the data protection regulations. This is a serious issue which led to the repeal of the Safe Harbor [103] data protection agreement between the EU-U.S., in 2016, regarding the existence of secret unauthorized surveillance programs and the lack of data protection during the transatlantic data flows from the EU to the entities that lie within the U.S.

1.4.2. Lack of data quality

Data quality assessment has been characterized as a key factor for achieving sustainable data of great quality in various domains varying from finance to healthcare [66], [67], [69], [71], [104], [105]. Lacking data quality results in bad data manipulation which makes data useless and has numerous negative effects on further processing. Thus, emphasis must be given on the development of proper mechanisms for data quality assessment. The latter lies under the well-known data governance part of a data sharing system. Data cleansing, also referred to as data curation [66], [67], [69], [71], [104], [105], comprises the core of the data quality assessment procedure. It aims to transform a dataset into a new one that meets specific criteria according to pre-defined quality measures. Examples of data quality measures include: (i) accuracy, (ii) completeness, (iii) consistency, (iv) interpretability, (v) relevancy, and (vi) ease of manipulation, among many others [106]. Data curation can be also used as a diagnostic tool for marking problematic attributes that exhibit incompatibilities (e.g., unknown data types, missing values, outliers). In this way, data curation can guide the clinician for fixing missing clinical misinterpretations which are not easy to be automatically detected, especially when fixing missing values.

Automated data curation overcomes the complexity of processing huge amounts of medical data and can be easily scalable in contrast with traditional manual data curation which is not feasible in the case of big data management. However, clinical evaluation is necessary to ensure the reliability and applicability of automation. Data curation can be seen as a sequential process, i.e., a series of methodological steps, which involves functionalities for curating both prospective and retrospective data. Mechanisms for curating retrospective data include: (i) the detection and elimination of duplicate fields (i.e., de-duplication), (ii) the characterization of data according to their context (i.e., data annotation), (iii) the identification of duplicate fields with highly similar distributions (i.e., similarity detection), (iv) the transformation of data into standardized

formats (i.e., standardization), (v) dealing with missing values (i.e., data imputation), and (vi) outlier detection for detecting values that deviate from the standard data range. Mechanisms for curating prospective data can be incorporated in the form of check constraints.

So far, it is clear that data sharing is indeed a benefit for the public good since it enables the interlinking of out-of-border medical and other related data, as well as, the reuse of these data and thus promotes scientific research worldwide. The strong demand for biomedical research and innovation, as well as, the existence of a smart healthcare system for disease surveillance and prevention are a few of the clinical unmet needs that data sharing has been proven to fulfill. However, apart from the fear for data abuse and the privacy laws, which constitute the two main significant barriers towards data sharing, there is still one significant concern that can make data sharing harmful; and that is data misuse [107]. The misuse of shared data has bad consequences and is many-sided. In this section, we will discuss the reasons behind the misuse of shared data, as well as, propose solutions for overcoming the fear regarding the misuse of shared data.

- *Absence of real evidence*: The researcher must make clear the reason behind data sharing, as well as, state the ensuing opportunities. The absence of real evidence hampers the data sharing process and produces the exact opposite outcomes. Thus, emphasis must be given to the purpose of data sharing.

- *Lack of data quality control*: Prior to the analysis of the data, it is of primary concern to assess the quality of the data, i.e., to curate the data. However, the misuse of methods for data curation introduces biases during the analysis which yields false outcomes. Two of data curation's important functionalities are the outlier detection and the data imputation. If a researcher performs data imputation prior to outlier detection, the dataset is very likely to be contaminated with false values (outliers) and thus will become useless. On the other hand, the outlier detection methods might identify mathematically correct extreme values but without any clinical interpretation. Therefore, the clinician's guidance is necessary not only to validate these findings but also to deal with missing values so as to avoid data contamination.

- *Lack of the researcher's skills*: The lack of knowledge regarding the hypothesis of a study makes the study pointless. A researcher must first state the hypothesis under

examination and then develop tools towards this direction. In addition, the researcher must be well aware of the scientific advances in the domain of interest, as well as, the software and tools that meet the specifications set by the hypothesis. Only high-quality researchers who are aware of data quality problems and causal inference methodologies are more likely to produce reliable outcomes [108]. In addition, the public health policy makers and decision makers might be too credulous sometimes, especially when the outcomes of a study involve large databases. As a matter of fact, the availability of big data does not always guarantee correct study outcomes, which yields another question here: Is bigger data always better?

- *Ignorance of the data collection protocols*: Not knowing the population characteristics of a study introduces many biases during the analysis procedure and produces false outcomes. In general, there are three types of biases which affect observational studies: (i) the selection bias, (ii) the confounding bias, and (iii) the measurement bias. The selection bias appears when the selected group of individuals for a particular study is not representative of the overall patient population [107]. Another appearance of this bias can be met in causal-effect studies, i.e., studies that involve the validation of a drug's treatment (benefit or harm effect on individuals). In this type of study, if a variable has a common effect on both the treatment/exposure factor and the outcome factor, it is considered as a collider-bias, which is also known as "M-bias" [107]. An example of this type of bias occurs when a patient's follow-up data is lost either because the patient's treatment is harmful (treatment factor) or because the patient's treatment is good (outcome factor). The lack of such information yields false statistical associations between these two factors and introduces distortions on the true causal effect [107]. Selection bias introduces distortions (e.g., false positives) in the outcome measures which hampers the disease prevalence and the risk exposure yielding false data models for patient stratification. It is, thus, important to appropriately adjust these types of variables during statistical analyses for obtaining true causal estimations.

Confounding bias, which is also met in causal-effect studies, is even worse than selection bias. A confounding variable is a variable which has a common cause on both the treatment/exposure and the outcome rather than a common effect [107]. A typical

example of confounding occurs when a clinician's decision is affected by a patient's disease severity or duration which in turn affects the treatment's outcome. Patients at an earlier stage of a disease receive different treatment than those in a later stage of the same disease whereas sicker patients may have worse treatment outcomes than the healthy ones. In this example, the confounding variable is the degree of sickness that is exhibited by the patients that receive different treatments. Such types of variables must be identified and properly adjusted.

Finally, the measurement bias is a widely known bias which arises from errors during the data measurement and collection process. The main reasons behind measurement bias are the following: (i) improper calibration of the measurement systems, (ii) lack of the measurement system's sensitivity, (iii) lack of the physician's expertise during the data measurement process, (iv) lack of a patient's trust and confidence during a questionnaire competence, and (v) patient's medical state (e.g., dementia).

- *Ignorance of the privacy laws and ethics policies*: The lack of knowledge regarding the data protection legislations has severe consequences concerning the patients' privacy and obscures data sharing. This factor has nothing to do with the biases in the outcomes of a study or the strategy used for data analytics rather than the privacy legislations breached by the study. The patient data must be first de-identified and qualified by appropriate scientific advisory boards. The de-identified data must be maintained in secure databases with private networks undergoing strict authorization procedures.

- *Poor use of the available data*: This has to do again with the skills and expertise of the researcher. The lack of data management and domain knowledge from the researcher's point of view results to misconceived analyses with extremely harmful results for the public.

- *Different interpretations of the same outcome*: This is a common mistake which underestimates the findings of a study. Clinical centers and laboratories worldwide, make use of different measurement systems and units for characterizing a patient's laboratory test. For example, a typical hemoglobin test may be recorded by a clinical center A in "mg/mL" whereas a clinical center B might record it in "g/dL". Moreover, the thresholds for characterizing the test's outcome might vary, e.g., the

clinical center A may consider a hemoglobin value of 15.5 “g/dL” as the threshold above which the hemoglobin levels are abnormal whereas clinical center B may consider a value of 17.5 “g/dL”. A solution to this is to include a new variable which states whether the hemoglobin levels are normal or abnormal. Standardization is thus important for the normalization of common terms across heterogeneous data.

1.4.3. Heterogeneity across medical data

The heterogeneity of data among biobanks, cohorts, and other sources of medical data is a critical scientific limitation which poses significant obstacles in the effective analysis of such data, yielding clinical studies with poor statistical power and, thus, inaccurate disease outcomes [52], [73], [109]–[111]. In computer science, data harmonization is an emerging technique which aims to overcome the structural heterogeneities that are present among the medical data derived from multiple sources by producing homogenized versions of the heterogeneous data that share a common medical domain (context). The overall idea of data harmonization is to transform the heterogeneous data into a common format with the exact same parameters and range values, using data-driven, and other computational approaches, such as, lexical, and semantic matching, to enable the integrative analysis of the heterogeneous data and therefore, enhance the statistical power of the clinical studies which make use of such data. To this end, data harmonization can enable the interlinking and subsequent integration of clinical data to deal with the unmet needs in various diseases.

The lack of a standard reference model often obscures the harmonization process, making the adoption of most of the data harmonization methods difficult to be implemented. In addition, the medical terms and the acronyms that are often adopted by the majority of the clinical centers during the data collection process are difficult to be parsed and sometimes unable to be matched with standard medical terms and indices (e.g., the use of the acronym “HGB” instead of “hemoglobin” or “haemoglobin” or any other use of acronyms during the definition of the attributes), a fact that obscures the accuracy of the harmonization process due to the underlying information loss. A solution to this would involve the clinician’s effort during the terminology mapping process so that he/she would be able to verify the validity of the terms that were marked as homogeneous. On the other hand, the absence of timestamps during the collection of prospective data, as well as, the existence of erroneously parsed fields during the data

collection process are additional factors that obscure the structural alignment process. Therefore, the application of a data curation workflow is an important pre-harmonization requirement to fix problematic fields (e.g., outliers, incompatibilities, unknown symbols) that are present within the original data.

The lack of terminology descriptions along with the absence of information regarding the meaning of the range values on each attribute, especially in the case of attributes with categorical values, hamper the data standardization process. For example, a clinical center may record the state of a medical condition using the coding term "normal" or "abnormal". Another clinical center can record the same condition using the binary values 0 and 1, respectively. This knowledge should be clearly indicated prior to the harmonization process. A similar example occurs when a clinical center records the levels of a laboratory measure as "low", "normal" or "high", whereas another center may use the values 1, 2 and 3, respectively, to indicate these measurement levels and another one may use the terms "low", or "high", skipping the "normal" level. These again are important factors that should be taken into consideration prior to the harmonization process. As for the attributes with continuous values, the measurement units (or normalized units) should be clearly stated. For example, a clinical center may record a laboratory measure in "mg/mL" whereas another clinical center may record the same value in "mg/dL", "μmol/L" or "g/L".

So, what if the parameters which are present in the standard template are not representative or limited to only a small portion of the domain's knowledge? This is a critical limitation that enhances the loss of information during the harmonization process and specifically during the terminology mapping process. For example, a retrospective dataset may include a set of 100 attributes whereas the standard model may only include a set of 50 related parameters, where the relevance of an attribute is trivial since a medical condition (e.g., cryoglobulinemia) can be followed by a set of related symptomatology (e.g., fever, weight loss). One way to reduce this type of information loss is to define a semantic representation of the standard model, where each parameter is assigned to a category (or class). In the previous example, the parameters "fever" and "weight loss" can be assigned to the category "symptomatology". A similar example occurs in the case where a clinical dataset includes more than one demographic-related parameters (e.g., education level) and/or

laboratory tests (e.g., blood tests) that could be parsed in the categories “Demographics” and “Laboratory tests”, respectively. If any detected symptomatology is assigned in the homonymous category, instead of being ignored, then the overall information loss would be greatly reduced.

The majority of these barriers can be overcome in the case of the prospective data collection process where the prospective data can be recorded through appropriate digital data entry forms that already include these standard measurement units and range values, as well as, the terminologies for each type of attribute. The data entry form can be used as a standard template like the one which is used in the case of the retrospective data harmonization process. In all cases, the scope of harmonization should be well-defined. However, apart from the technical challenges that are met during data harmonization, the most prominent factor that facilitates data harmonization is the establishment of a legitimate environment that enables the sharing of data from multiple data sources. Data harmonization is in line with data sharing and, thus, the lack of a legitimate data sharing mechanism would make data harmonization pointless.

The platform must support efficient web communication for faster data transfer. Most importantly, the platform must offer effective de-identification mechanisms through the construction of unique identifiers per patient (e.g., hash keys). In addition, the metadata must be followed by an expiration date for security reasons. The data must be stored in secure private repositories and the access should be controller through multiple-factor authentication systems. Informed consent forms must be requested in the case of pooled data analysis in order to overcome the fear for data abuse. Any data transfer within the platform must be tracked down by proper system monitoring and log mechanisms including audit tables which record the date, the time, and additional information regarding the user’s access to the stored data.

All operations within the cloud must take place in secure virtual private networks for ensuring the confidentiality during the transfer of sensitive data. Data security comprises the biggest barrier of cloud computing. The lack of data security results to data leakage, data abuse, loss of data integrity and control over the hosted data and the cloud applications. To deal with such issues, OAUTH-type authorization frameworks must be adopted to ensure secure user authentication and access management and

secure access to the information and services of the platform. The flow of sensitive information outside the platform (e.g., user credentials) must be encrypted and decrypted through Secure Sockets Layer (SSL)/Transport Layer Security (TLS) protocols using public decryption keys and private encryption keys. The inner information flows are performed through secure firewalls and virtual private networks which enhance the reliability of the platform and ensure a highly secure information transfer.

1.4.4. Lack of population size

Nowadays, the lack of access to open, secure, interoperable, and transparent health data hubs poses significant obstacles to researchers and innovators, such as, SMEs, and healthcare stakeholders towards the deployment of trustworthy data analytics workflows for synthetic data generation, data anonymization and AI modeling to promote wellbeing, diagnosis, disease prevention, progression, and treatment. This has led to an emerging need for the development of high-quality synthetic data generators. The reduced amount of available training data [112], particularly in rare diseases (e.g., primary Sjögren's Syndrome), where the population size is inadequate and the quality of data is low [113] highlights the emerging need for the development of high-quality synthetic data and robust generative models to address the challenges of today, such as, data confidentiality and data augmentation. Moreover, the financial burden of expensive drugs leverages the orchestration of viable Phase II/III clinical trials (CTs) [114]–[116], as well as, the identification of predictors for disease prevention, diagnosis, progression, treatment, decision-making in common diseases, such as, type-2 diabetes and Alzheimer's disease.

Furthermore, the in-applicability of secure, cryptographic techniques [117] that can facilitate the interconnection of decentralized clinical data registries and cohorts obscure the successful deployment of (AI)-powered workflows. As a matter of fact, the aforementioned factors have a significant negative impact in the capacity of the existing healthcare systems, where the costs and delays for treatment and re-admission are already high. In addition, although the existing data anonymization algorithms incur high levels of information loss, patient privacy is not guaranteed given that the protection of sensitive patient data is considered a fundamental right [118]. Moreover, since the most common strategy for knowledge distillation is based on integrative data

analysis from multiple dispersed clinical registries and cohorts [119], the collection of sensitive data out of premises is not feasible, due to GDPR (General Data Protection Regulation) violations during the sharing of patient data [120].

1.4.5. Data silos undermining the deployment of AI models

Centralized databases are less complex and convenient. The data management and maintenance process are easier since the data are gathered in a unified form. However, centralized databases are not reliable since all the data can be compromised and abused in the case of an attack and thus lacks crucial security measures. This security issue is reduced in distributed databases, since an attack to a specific local DBMS node can cause the rest of the local DBMS nodes to lock down the access to their connected databases and, thus, provide a moderate data security level. This is also present in de-centralized databases, where a malicious attack on a single node does not compromise any of the rest nodes and thus accomplishes high security levels. In addition, in the case of an error in a single node, the data can be removed from the problematic node to another node, for safety purposes. This can be also applied in distributed databases but at a higher level since the data from the single nodes can only be moved through their corresponding local DBMS nodes.

Furthermore, the recovery rate in a centralized database is small since a query (e.g., a search operation) must be executed on the whole database whereas in distributed networks, the query is distributed to smaller portions of data (subsets or batches) where the execution is faster. In addition, a centralized database cannot be easily expanded due to their low scalability whereas the distributed and de-centralized databases can be easily expanded due to the high scalability they offer. On the other hand, distributed databases are more complex than centralized databases since the former require continuous communication with the local DBMS nodes to send the queries and receive the results. This complexity is largely increased in de-centralized networks where the communication needs grow exponentially.

1.4.6. AI model explainability and interpretability

Nowadays, there is an emerging need to provide explainable and trustworthy AI models which envisage to shed light into the backbone of the decision-making process and the interpretability of the identified risk factors rather than focusing only on the

classification performance of the AI models in terms of accuracy, sensitivity, specificity, and area under the curve, among others.

An AI model must be designed to fulfill a set of seven fundamental requirements to prove its trustworthiness in terms of compliance with the four ethical principles of [121]–[123]: (i) respect for human autonomy (the AI systems should be designed to empower human congestive and social skills), (ii) prevention of harm (the AI systems should be designed to protect the human dignity by being safe and secure), (iii) fairness (the AI systems must be developed and deployed in such a way to increase societal fairness), and (iv) explicability (the processes that are implemented by the AI system must be transparent in terms of traceability and auditability). The seven requirements include the following [121]–[123]: (i) accountability, (ii) privacy and data governance, (iii) societal and environmental wellbeing, (iv) technical robustness and safety, (v) human agency and oversight, (vi) diversity, non-discrimination and fairness, and (vii) transparency. These principles are further explained below.

- **Accountability** has to do with the responsibility of the outcomes of the AI system during their development and after their deployment in terms of auditability, risk minimization and respect to the fundamental rights.
- **Privacy and data governance** has to do with the quality of the data in terms of relevance, completeness, integrity, as well as, the fulfillment of the data protection legal and ethical requirements for data sharing.
- **Societal and environmental wellbeing** involves the development of a sustainable and environmentally friendly AI system.
- **Technical robustness and safety** involve the prevention or risks and the minimization of any unacceptable harm.
- **Human agency and oversight** involve the support of the human autonomy and oversight through decision-making.
- **Diversity, non-discrimination, and fairness** involves the avoidance of biases and the adoption of a global design for accessibility.
- **Transparency** ensures the traceability, explainability and human interaction of the AI system. The current thesis considers all the relevant advances in AI model explainability. The SHapley Additive exPlanations (SHAP) method [124] is widely used to quantify the contribution of each feature to the classification outcome.

The Shapley values provide contrastive explanations of the classification outcomes which can be utilized to reveal local interpretations for the given features. These explanations are based on the classification outcomes from specific training and testing instances. Moreover, the Shapley values preserve the properties of efficiency, symmetry, and additivity which are the fundamental properties during the evaluation of any feature importance score.

1.5. Contribution of the thesis

The current dissertation is dedicated to the design, development, and deployment of beyond the state-of-the-art workflows that can be used to address open issues and unmet needs in healthcare. Examples of open issues in healthcare, include:

- the lack of data quality,
- the underlying data heterogeneity and complexity,
- the development of semantic data models that can reflect the domain knowledge,
- the lack of sufficient population for disease modeling, particularly in rare diseases,
- the inability to integrate data into a centralized repository for AI modeling, and
- the design of federated AI workflows for AI modeling across multiple databases.

Examples of unmet needs in healthcare, include:

- the early detection of high-risk patients,
- the accurate prediction of an event (e.g., a disease or a condition),
- the discovery of new digital biomarkers,
- the explainability of the identified digital biomarkers,
- the explainability of the AI models in healthcare.

By taking into consideration these issues and needs in healthcare, the current thesis has been built on top of four fundamental pillars which shift the current state of the art in data curation, data harmonization, synthetic data generation and federated/distributed learning, as presented next.

- **Pillar 1: Data sharing and data curation.** We propose a fully automated framework for medical data curation to enhance the quality of the data in terms of completeness and conformity. The framework serves as a diagnostic tool for

managing incomplete terminologies, irrelevant terms, outliers, missing values, data categorization, and duplicated terms. We developed a "smart" data imputation approach based on optimal virtual profile matching to address data incompleteness across complex clinical data structures. In this work, we extend data standardization as a pre-harmonization process to make data harmonization easier and faster. More specifically, we use lexical matching combined with model-based rules and external sources, i.e., vocabularies, to match and classify terms according to a pre-defined reference model which is a set of parameters which describe the requirements (variables with their types and ranges) of the clinical domain of interest. Through this procedure, we attempt to produce semantic relations between the fields of the raw dataset with those from a reference model and therefore enhance the semantic matching process for data harmonization. The proposed framework accounts also for data standardization since it can produce a set of semantic relations through a rule-driven approach that is developed based on a pre-defined reference model and captures important semantic relations which enable faster data harmonization. In addition, the framework can be easily adjusted with new rules according to a provided reference model that describes the clinical domain of interest.

- **Pillar 2: Data harmonization.** We propose a hybrid data harmonization workflow which adopts an automated strategy that combines lexical analysis with semantic models (ontologies) to identify terminologies with lexical and conceptual overlap. The proposed approach is based on the definition of a reference semantic data model (reference ontology) for the domain of interest. A medical corpus is then defined by interlinking FHIR compliant terminologies from the SNOMED-CT and the ICD-10/11 under the OHDSI Athena vocabulary. Synonyms of the reference ontology are also harnessed by the NLTK toolkit to further enhance the medical corpus with ontology-oriented terminologies. The lexical and semantic analyzers are applied on top of the medical corpus to automatically align the terminologies of the raw data with those from medical corpus, at a metadata level. The coherence is calculated to quantify the lexical and semantic overlap of the identified terminologies.
- **Pillar 3: Synthetic data generation.** We propose a hybrid synthetic data generator which focuses on the optimal estimation of the Gaussian components in the BGMM algorithm to yield concrete estimations of the VI at reduced computational complexity for large-scale synthetic data generation (we refer to this approach as BGMM with Optimal Components Estimation: BGMM-OCE). To do so, we first

apply spectral clustering based on the Locally Optimal Block Preconditioned Conjugate Gradient (LOBPCG) method to identify the best clustering solution as the one with the highest Davies Bouldin score (DBS) at small complexity. Then, we set the optimal number of clusters as the number of Gaussian components, and we define an exponentially decaying gamma value.

- **Pillar 4: Development and deployment of machine learning workflows across distributed/federated environments.** We propose a federated AI model deployment framework to ensure the trustworthiness of the AI modeling process across federated databases. We extended conventional supervised machine learning implementations to support federated learning, such as, the federated multinomial naïve bayes (FMNB), the federated stochastic gradient descent-based algorithms (FSGD) (e.g., logistic regression, support vector machines), and the federated gradient boosting trees (FGBT) with dropout rates (FDART). We developed the federated hybrid boosted forests (FHBF) algorithm which implements a hybrid weight update approach to deal with ill-posed problems that arise from overfitting effects during the training across complex and highly imbalance data in federated databases. A scale parameter is introduced to control the shape of the hybrid loss function based on the dropout rate to avoid overfitting effects. The FHBF currently supports both the hybrid FGBT (HFGBT) and the hybrid FDART (HFDART) as boosters. Class imbalance handling functionalities are incorporated to develop clusters of HFDARTs, where each cluster is formulated based on a random subset of the federated training instances. Then, a log loss score is used to isolate the weak sets of regression trees to further boost the classification performance of the algorithm and increase its resilience against weak decisions. We placed particular emphasis on SHAP analysis to yield explainable outcomes. Distributed implementations of these algorithms are also supported.

Data sharing, data curation and data harmonization are three fundamental pillars that envisage to break down data silos and promote data interoperability. One of the aims of the current dissertation is to shed light into these pillars to understand the clinical need for medical data sharing, data curation and harmonization along with the related technical challenges lurking around it. Resolving privacy issues is an important part of this effort which emphasizes the need to safeguard patients' rights before any data manipulation. In addition, the present dissertation presents beyond the state-of-the-art

methods to enhance the quality of raw data and overcome data heterogeneity between complex clinical data structures. Data sharing, combined with data harmonization, is a powerful tool that provides the ability to collect and harness knowledge from heterogeneous data across different clinical databases to extract homogeneous data structures that can increase the statistical power of the clinical studies that make use of such data. The lack of population size combined with the lack of open data and the reduced predictive performance of the existing AI models can be addressed using high-quality syntenic data with increased convergence with the real data under the aegis of data augmentation.

Furthermore, the recent technological advances in federated/distributed learning algorithms in conjunction with existing data mining algorithms will provide the scientific community with the opportunity to better understand the fundamental basis, clinical needs, and functionalities of a federated architecture to break down data silos and promote research. These unmet needs are related to patient stratification for the early identification of high-risk individuals considering various clinical laboratory factors, as well as, the identification of risk factors that can serve as potential predictors for disease progression.

To this end, the current dissertation aims to answer the following eight research questions. These questions have been thoroughly investigated across six core case studies, in six different domains: (i) in autoimmune diseases (AD) for the primary Sjogren's Syndrome, (ii) in hypertrophic cardiomyopathy (HCM), (iii) in systemic autoinflammatory diseases (SAIDs) for the Kawasaki disease, (iv) in COVID-19, (v) in cardiovascular diseases (CVD), and (vi) in mental disorders (MD), to promote the technical and clinical impact of this thesis.

Q1.How can we share sensitive patient data from multiple dispersed databases?

A1. **Related Pillar(s):** 1. **Case Study:** AD.

Q2.How can we automatically improve the quality of medical data?

A2. **Related Pillar(s):** 1. **Case Studies:** AD, HCM, COVID-19, SAIDs, CVD, MD.

Q3.How can we sufficiently describe the domain knowledge of a disease of interest?

A3. **Related Pillar(s):** 2. **Case Studies:** AD, CVD, MD.

Q4. How can we automatically deal with data heterogeneity across diverse databases?

A4. **Related Pillar(s):** 1, 2. **Case Studies:** AD, CVD, MD.

Q5. How can we enhance the population size of a database?

A5. **Related Pillar(s):** 3. **Case Studies:** AD, HCM.

Q6. Is data augmentation effective in disease modeling?

A6. **Related Pillar(s):** 3. **Case Studies:** AD, HCM.

Q7. How can we apply federated learning with resilience against overfitting effects?

A7. **Related Pillar(s):** 4. **Case Study:** AD.

Q8. How can we enhance the explainability of the AI models and identify high-risk subgroups?

A8. **Related Pillar(s):** 4. **Case Studies:** AD, SAIDs, COVID-19.

1.6. Structure

The thesis is structured as follows:

Chapter 1 is an introductory chapter which aims to familiarize the reader with the fundamental principles and concepts behind the value of big data in healthcare, the types and sources of big data and the importance of data sharing in healthcare. In addition, the chapter places particular emphasis on the open issues and unmet needs in healthcare, including the lack of data quality, data heterogeneity, and lack of statistical power, among others. The chapter concludes with the contribution of the current thesis.

Chapter 2 presents the proposed workflow to address the unmet needs in six different thematic areas (clinical domains) along with the current technical and clinical state of the art regarding data curation, harmonization, synthetic data generation and federated/distributed learning. The six thematic areas include: (i) the autoimmune diseases, (ii) the hypertrophic cardiomyopathy, (iii) the systemic autoinflammatory

diseases, (iv) the coronavirus disease, (v) the cardiovascular diseases, and (vi) the mental disorders.

Chapter 3 provides a concrete view on the functionalities of the beyond the state-of-the-art data curation service which has been developed under the aegis of this thesis. The functionalities of the data curator involve metadata extraction and data annotation, outlier detection, de-duplication, and data imputation using a variety of univariate, multivariate and advanced machine learning based methods. Emphasis is also given on the description of the primary outcomes of the data curation workflow, including the data evaluation report, the curated dataset, and the clean curated dataset.

Chapter 4 presents the proposed data harmonization workflow and its' beyond the state-of-the-art functionalities to overcome the structural heterogeneity across complex clinical data structures. Lexical matching methods are fist described including the Levenshtein distance, the Jaro distance and the Jaro Winkler distance. Then, emphasis is given on the rationale of semantic matching including a complete view on data modelling, varying from relational modeling, and ontologies to HL7-standards and web ontology languages. The concept of word embeddings is introduced to be incorporated into a hybrid data harmonizer combining lexical and semantic matching methods with knowledge bases with international medical indices and word embeddings to identify terminology overlaps across heterogeneous clinical data.

Chapter 5 focuses on the description of the proposed large scale, computationally efficient and high-quality synthetic data generator and it's beyond the state-of-the-art functionalities in the context of *in silico* clinical trials to promote drug research. For comparison purposes, both statistical and machine learning based generators are also presented. Then, the proposed method for the robust initialization of the Gaussian components in the Bayesian Gaussian Mixture Model (BGMM) algorithm is presented along with additional hyperparameters, like the weight concentration parameter. The model training and sampling process is described along with widely used synthetic data quality metrics. Emphasis is given on data augmentation and its vision towards the improvement of the existing AI models' predictive performance.

Chapter 6 offers the basis for understanding the issues with centralized data analysis and the rationale of federated/distributed learning. The main learning schemas are first

presented, including online learning, meta-learning, and incremental learning. Then, the proposed federated AI framework is presented followed by federated implementations of popular supervised learning algorithms. Emphasis is given on the proposed federated hybrid boosted forests classifier for solving intensive supervised learning problems across highly imbalanced data structures in the cloud. Additional emphasis is given on distributed implementations and the design of hybrid loss functions to avoid overfitting effects. The chapter concludes with the importance of explainability analysis and its clinical impact on the decision-making process.

Chapter 7 is dedicated to the detailed evaluation of the proposed data curation, data harmonization, synthetic data generation, and federated AI modeling workflows which are presented in Chapters 3-6, respectively, across the six clinical domains from Chapter 2.

Chapter 8 summarizes the key points of the previous chapters and presents the latest trends in the rapidly evolving fields of data curation, data harmonization, synthetic data generation and federated learning.

Chapter 9 summarizes the major points of this thesis and the future work.

CHAPTER 2. STATE OF THE ART

2.1. An overview of the proposed workflow

The proposed workflow to address the existing open issues and unmet needs in healthcare (Section 1.4) is depicted in Figure 4.

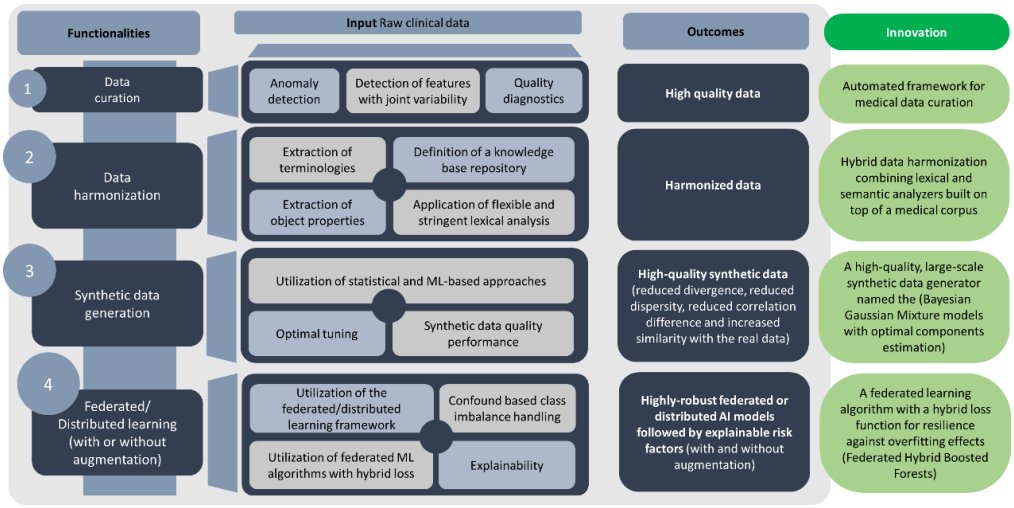


Figure 4. An illustration of the proposed workflow (the contribution in each innovation area – functionality is highlighted with green color).

According to Figure 4, the current thesis aims at delineating the open issues and clinical unmet needs in healthcare by offering beyond the state-of-the-art methods which have been developed around four innovation areas that are described next.

- **Innovation Area 1 - Data curation (Chapter 3):** Unlike conventional approaches which focus on the application of semi-automated or qualitative methods for data

curation, we focused on the design and successful development of a fully automated and highly scalable medical data curation service, which offers a suite of functionalities for:

- the precise detection of outliers (anomalies) across structured data using both univariate and multivariate methods,
- the accurate detection of highly correlated pairs of features,
- the robust detection of lexically similar pairs of features as potential duplicates,
- the effective application of a "smart" data imputer which utilizes a profile matching score to identify the best matching virtual patients for each real patient,
- the generation of re-usable data quality report highlighting important metadata along with descriptive statistics and feature-level quality status,
- the generation of a diagnostic report (often referred to as curated dataset) where the data incompatibilities are highlighted using proper color coding, and
- the efficient initialization of the memory requirements across complex gene expression microarray datasets.

Core publications: [66], [125].

- **Innovation Area 2 - Data harmonization (Chapter 4):** Unlike conventional approaches which focus on the application of semi-automated methods like lexical matching or on the manual definition of pairing rules for semantic matching, we focused on the development of a hybrid data harmonization workflow which utilizes both lexical and semantic matching methods on top of a medical corpus harnessing knowledge from word embeddings, reference ontologies and international knowledge bases to automatically identify terminologies with lexical and semantic overlap. The hybrid data harmonizer offers a suite of straightforward functionalities for:
  - the automated extraction of semantic information from reference ontologies,
  - the extraction of word embeddings for a given set of terminologies,
  - the extraction of NLTK-based synonyms for a given set of terminologies,
  - the extraction of FHIR-compliant terminologies from SNOMED-CT, LOINC, etc.,

- the application of string-matching methods aiming to solve the edit distance problem (e.g., Levenshtein distance) between two strings (terminologies),
- the definition of a complete medical corpus,
- the inclusion of semantic knowledge in the corpus, and
- the generation of a re-usable data harmonization report including the matched terminologies along with a coherence score and the most frequent terminologies.

Core publications: [126]–[129].

- **Innovation Area 3 - Synthetic data generation and augmentation (Chapter 5):** Unlike conventional approaches which focus on the application of statistical or less-efficient, large-scale machine learning based synthetic data generators, we focused on the development of cost-effective, large-scale synthetic data generator named Bayesian Gaussian Mixture Models with Optimal Components Estimation (BGMM-OCE) algorithm which supports:
  - the robust estimation of the number of Gaussian components,
  - the unbiased definition of the weight concentration parameter (or gamma),
  - the generation of high-quality synthetic data in terms of reduced goodness of fit, Kullback-Leibler divergence, coefficient of variation and correlation difference,
  - the generation of large-scale synthetic data with small computational complexity.

  A computational workflow for data augmentation was also developed to:
  - populate clinical databases with small population size (e.g., in rare diseases),
  - aggregate high-quality synthetic data with real data,
  - to train AI algorithms on the aggregated (real and virtual) data,
  - to evaluate the classification performance of the AI model on real subsets,
  - to compare the classification performance of the AI models trained on the real and on the aggregated data (to evaluate the impact of augmentation).

  Core publications: [130]–[134].
- **Innovation Area 4 - Federated/distributed learning (Chapter 6):** Unlike conventional approaches which focus on the deployment of local nodes on the premises of each hospital, we proposed a federated AI model deployment system, on the cloud, where the nodes are replaced with private cloud spaces. In addition, we propose a hybrid federated learning algorithm named federated hybrid boosted

forest (FHBF) which is resilient against overfitting effects during the training across federated databases with increased class imbalance. To this end, the proposed FHBF approach offering:

- an adjustable hybrid loss topology based on the dropout rate,
- resilience against overfitting effects through a hybrid loss function,
- two types of hybrid boosters for training,
- explainability analysis, and
- confound-based class imbalance handling.

Core publications: [65], [128], [129], [135].

2.2. Technical point of view

2.2.1. Data curation

Table 1 presents the objectives, methods, and outcomes of the current state of the art studies focusing on data quality control in healthcare. More specifically, in [136] a data quality assessment framework is proposed to enhance the completeness, correctness, concordance, plausibility and currency of medical data by computing the percentage of matched variables across records, and matched records across patients, as well as, the type of records per patient, the presence of selected variables, and the frequency of records per patient over time. In addition, in [137] a draft set of harmonized terms is presented. The set of terms was defined by the experts and was organized into three quality categories (i.e., conformance, completeness, and plausibility) to be compared with ten existing data quality terminologies in the context of electronic health record data, where the comparison was based in terms of coverage in the EHR domain.

Furthermore, in [138] a framework that deals with the completeness, consistency, correctness, non-redundancy, and timeliness of medical data in a semi-automated way where the user defines the quality mapping criteria (e.g., completeness) and the data quality levels (e.g., acceptable) for each data source. Moreover, in [139] a multi-dimensional data storage solution in a semi-structured data curation engine, which provides foundational support for archiving heterogeneous medical data and achieving partial data interoperability in the healthcare domain. The ExeTera software [70] provides functionality that enables a data curation pipeline incorporating data curation methods for COVID-19. The pipeline includes preliminary data cleaning and filtering

using semantic information, and generation of meta-analytics for daily assessment. Finally, in [140] the raw data are curated to a common data model (C-Surv). The C-Surv ontology was designed to simplify the analytic challenge of working across multiple datasets and multiple modalities by providing standard structure, variable naming, and value labelling conventions.

Table 1. Current state-of-the-art methods, tools, and frameworks for data curation.

Study	Objective	Method	Outcomes
[136]	Assess the quality of electronic medical records (EMR)	A data quality assessment framework that aims to deal with the following data quality dimensions: (i) completeness, (ii) correctness, (iii) concordance, (iv) plausibility and (v) currency.	A conceptual data quality assessment framework based on six qualitative metrics.
[137]	Assess the quality of electronic health record (EHR) data	A draft set of harmonized terms organized into three quality categories to be compared with ten existing data quality terminologies in the context of EHRs.	A set of harmonized terms for EHR quality assessment.
[138]	Present a framework for data quality management in health care institutions	A semi-automated framework that deals with the completeness, consistency, correctness, non-redundancy, and timeliness of medical data.	A semi-automated framework for data quality management in health care institutions.
[139]	Crate data curation services for storing healthcare data, creating, and storing the semantic reconciliation knowledge base.	A multi-dimensional data storage solution in a semi-structured data curation engine to archive heterogeneous medical data in the healthcare domain.	An interoperable data storage framework in the form of a semantic knowledge repository for quality control.
[70]	To present an open-source data curation software designed to	The ExeTera software provides functionality that enables a data curation pipeline incorporating	A data curation software with qualitative methods

Study	Objective	Method	Outcomes
	address scalability across the COVID Symptom Study dataset.	data curation methods for COVID-19.	mainly for data quality control based on pre-defined semantic knowledge.
[140]	Present a platform for data curation, data discovery, access brokerage, data analysis and knowledge preservation	The raw data are curated to a common data model (C-Surv) which is designed to simplify the analytic challenge of working across multiple datasets.	A platform that considers data quality criteria which are manually defined for each individual data source.

2.2.2. Data harmonization

According to the literature, a variety of computational methods for medical data harmonization has been proposed so far [141]–[146]. A robust data harmonization method involves the application of lexical and semantic matching algorithms. A lexical matching algorithm uses string similarity techniques [52], [143], [146] to identify common terminologies (i.e., exact sequences or similar block sequences) that are present between the terms of the standard model and those of the original dataset. External vocabularies can also be used to enrich the clinical domain knowledge and thus enhance the accuracy of the overall lexical matching process through the identification of homonyms or synonyms.

On the other hand, the semantic matching method [52], [142], [144], [145] uses semantic relationships that exist between the terminologies, apart from the lexical matching process that is already included, to reduce the information loss and enhance the overall data harmonization process. This can be accomplished through the construction of ontologies which represent the clinical domain knowledge of interest in the form of entities (e.g., classes), and object properties (e.g., "includes", "has", "consists of"). Semantic matching uses a standard (or reference) model which is usually expressed in the form of an ontology, where the classes are considered as categories, e.g., "Clinical tests", that might consist of further sub-classes, e.g., "Blood Tests", etc. Each class can include a set of variables which are related to the class they belong to in terms of common meaning or concept. For example, the class "Blood tests" includes

the variables “age”, “gender”, “hemoglobin levels”, etc. This can lead to the semantic matching of the variables which might not be lexically identical but share a common concept. According to Table 2 several data harmonization frameworks have been launched to enable the integrative data analysis of heterogeneous medical data, such as, clinical, and genomic data, the majority of which is presented below, including the DataSHaPER (DataSchema and Harmonization Platform for Epidemiological Research) [141], [147] framework, the SORTA (System for Ontology-based Re-coding and Technical Annotation of biomedical phenotype data) tool [142], the BiobankConnect software tool [143], the S-match semantic matching framework [144], and the FOAM (framework for ontology alignment and matching) [145].

Table 2. A summary of the fundamental frameworks for data harmonization.

Tool	Data harmonization strategy	Application
DataSHaPER [141]	Uses a DataSchema as a reference model to harmonize heterogeneous data schemas according to the user-defined DataSchema through the development of pairing rules.	A 36% compatibility for creating a harmonized database across 53 of the world’s largest longitudinal population-based epidemiological studies [147].
SORTA [142]	Uses lexical matching to align phenotype data from heterogeneous biobanks according to international coding systems.	Matched 5,210 entries in the LifeLines biobank [148] (97% recall) and 315 entries in the DUMR (58% recall) [142].
BiobankConnect software [143]	Uses lexical and semantic matching to align heterogeneous biobanks according to a desired set of pre-defined elements.	An average precision 74.5% towards the harmonization of data across six biobanks (7,461 terms) with 32 desired elements [143].
S-Match [144]	Uses semantic matching to quantify the semantic relations that exist between the elements of two light-weight ontologies into 4 different categories.	A 46% precision on the correct identification of semantic matches in the TaxMe2 dataset [149] as part of the Ontology Alignment Evaluation Initiative [150].
FOAM [145]	Trains HMMs on sequence profiles that exist in international	A functional ontology that includes a set of more than 70,000 trained

Tool	Data harmonization strategy	Application
	registries to align heterogeneous sequence profiles.	HMMs targeting 2,870 different KOs [145].
BiobankUniverse [146]	Integrates lexical comparison, Unified Medical Language System ontology tagging and semantic query expansion.	A fast matchmaking service for biobanks and researchers, where Users can quickly explore matching potential and search for biobanks/data elements matching their research.

The DataSHaPER (DataSchema and Harmonization Platform for Epidemiological Research) framework [141], [147] was developed under the BioSHaRE project [151] to enable the harmonization of heterogeneous biobanks. The DataSHaPER uses a DataSchema (which is also referred to as "Generic DataSchema") as a reference model which consists of a core set of more than 180 variables that belong to 45 domains and 13 themes [151]. The DataSchema also includes 3 modules which are more abstract entities that include the domains, the themes, and the variables. The DataSHaPER framework supports the harmonization of both prospective and retrospective studies only when the reference model is exclusively defined in a DataSchema format. The harmonization strategy involves the execution of a three-step procedure which involves [141], [147], [151]:

- the development of pairing rules that quantify the ability of each variable in the study to generate the variables of the "DataSchema",
- the application of the pairing rules to classify each variable in the study into "complete", "partial", or "impossible" according to its ability to generate the variables of the "Generic DataSchema", and
- the development of a processing algorithm that can automatically generate the variables of the "DataSchema" from the variables that have been marked as potentially matched.

The processing algorithm that enables data harmonization is executed through the Opal software [152] as soon as the harmonized DataSchema is constructed and distributed to the biobanks. The harmonized DataSchema includes a core set of variables which is

related with the domain of interest and are selected according to seven criteria [141], [147]. The biobank data are stored in dedicated Opal servers [152] which process the data to compute the harmonized DataSchema specifications. Then, the harmonized DataSchema specifications from each individual study are converted to the variables in the DataSchema format. The individual harmonized datasets are stored in the Opal servers of each biobank and delivered in remote clients through the Mica web portal [152]. The overall procedure takes place under a secure cloud infrastructure known as DataSHIELD (Data Aggregation Through Anonymous Summary-statistics from Harmonized Individual-levEL Databases) [153], [154]. The DataSHaPER approach has been applied in [151] to harmonize eight retrospective cohort studies with more than 200,000 individuals under the Healthy Obese Project (HOP), yielding a 70% harmonization accuracy towards the identification of variables with shared terminology.

The BiobankConnect software [143] was launched under the aegis of the BioSHaRE EU-funded project [151] to enable the integration of heterogeneous biobanks. The software uses lexical matching combined with semantic matching to enable the semiautomated harmonization of terms in heterogeneous biobanks. The software quantifies the potential of harmonizing the terms among the heterogeneous data with the desired elements that are defined by the user by searching for lexical matches between the terms of the raw data with existing terms along with semantic information which is derived by the subclasses and the object properties of the ontologies. The user annotates the desired elements, i.e., the core set of elements, through the Bio-portal [143] which serves as a widely used repository for biomedical ontologies. The software then seeks for lexical matches between the terms of the raw data with those from the existing ontologies and provides a matching score that reflects the percentage of relevant matches. The BiobankConnect software tool has been used to harmonize data across six biobanks including 7,461 terms based on a schema of 32 desired elements yielding 0.74 and 0.75 precision rates in two ranks that were defined by the experts [143].

The System for ontology-based re-coding and technical annotation of biomedical phenotype data (SORTA) tool [142] is an example of another tool that was developed under the BioSHaRE EU-funded project [151], towards the integration of phenotypes

across heterogeneous biobanks. While the BiobankConnect software involves the mapping of the heterogeneous data structures (elements) into a target schema, the SORTA tool deals with the heterogeneity of the data in a more high-level manner through the coding (or recoding) of the data values into international coding systems, such as, the SNOMED-CT [155], the ICD-11 [156] and the Human Phenotype Ontology (HPO) [157]. More specifically, the SORTA tool seeks for matches between the data values and a target, user-specified, coding system. The coding system can be defined in terms of an ontology or a .csv document. The SORTA tool then uses lexical matching algorithms, such as, the n-gram, to quantify the lexical similarity between the data values to provide a sorted list of candidate elements per data value. SORTA has been used to match 5210 unique entries in the LifeLines biobank [142], [148] and 315 unique entries in the Dutch Uniform Multicenter Registration system for genetic disorders and malformation syndromes [7, 60] in line with the HPO, yielding 97% and 58% recall ratios, respectively.

The S-Match tool [144] is an open-source semantic matching framework which provides a rigorous solution for the semantic interoperability problem using lightweight ontologies. A lightweight ontology is defined as a taxonomy, where the natural elements are described in a formal way using formal expressions in the form of tree like structures, where each node in the tree has a specific meaning or concept [144]. The S-Match tool uses semantic matching algorithms that quantify the semantic relations between the elements (nodes) of two lightweight ontologies into four categories, namely: (i) equivalent, (ii) less general, (iv) more general, and (v) disjointless. First, the algorithm computes the concepts and the meaning of each label and then it computes the relations between the concepts of the labels and the nodes. The nodes are constructed in the form of trees and the meaning of each “child” node is related with the meaning of its “parent” node. Once a new ontology is given as input the algorithm seeks for an existing conceptual relation between the concepts of the labels. If no semantic relations exist, the algorithm uses syntactic level matchers, like the Edit distance, to seek for lexical matches. The algorithm outputs a file with the identified semantic and syntactic relations between the concepts of the labels of the two lightweight ontologies. The performance of the tool was tested on the TaxMe2 dataset [149] which is a benchmark under the Ontology Alignment Evaluation Initiative, yielding 46% precision towards the correct identification of the semantic relations.

The Functional Ontology Assignments for Metagenomes (FOAM) [145] is an example of a computational framework that focuses on the classification of gene functions which are related with microorganisms using Hidden Markov Models (HMMs). The HMMs are trained on the sequence profiles that exist on the Kyoto Encyclopedia of Genes and Genomes (KOGG) orthologs (KOs), which are part of the Gene Ontology [158], to enhance the accuracy of the classification process towards the successful alignment of the sequence profiles with the targeted KOs. In fact, the KOGG is used as an external vocabulary to enrich the existing information from the sequence profiles. The whole process is semiautomated since the user needs to first define the target space for the alignment. FOAM also provides a functional ontology that describes the hierarchy of the groups during the training of the HMMs.

Another family of data harmonization methods is statistical harmonization which involves the application of linear and non-linear statistical models to investigate the effect of different latent factors on a set of one or more items [159]–[162]. In statistical theory, the items can be considered as all types of informative variables (e.g., depression) that are observed and the latent factors as variables that are not directly observed but are rather inferred by the items. The purpose of statistical data harmonization is to homogenize scales that measure the same item and transform them into a common metric of the same scale, where the types of the items might vary from discrete and ordinal to continuous [159]–[162]. Statistical approaches, such as, multiparameter logistic Item Response Theory (IRT) analysis were deployed in [159]–[162] to examine how a set of items (e.g., psychiatric phenotypes) is affected by other factors for scale homogenization.

For example, a clinical center might record the cholesterol levels using the scale low, medium, high, whereas another clinical center may record the same levels using the scale 0, 1, and 2. Thus, statistical harmonization tries to recode the variables that belong to the same construct so that they are commensurately scaled at the end [159]–[162]. Of course, the detection of variables that express different (or common) scales and belong to the same construct is challenging since there is no prior knowledge regarding the names of the items like in lexical or semantic matching [159]–[162]. Even if two variables (items) describe the same construct, it is not always proper to match these two

variables since there might be differences in the population characteristics between the clinical studies, e.g., differences in the education level, ethnicity, gender, etc.

Such differences need to be controlled during the statistical analysis process and thus the complexity of the harmonization process is greatly enhanced. Moreover, the types of the items directly affect the type of the statistical model to be used. Towards this direction, a variety of statistical methods has been proposed so far, especially for the harmonization of psychometric and cognitive items with different measurement scales across clinical data, including [159]–[162]: (i) simple linear factor analysis (LFA) for continuous items, (ii) 2-parameter and multi-parameter logistic Item Response Theory (IRT) analysis for binary items, (iii) generalized linear factor analysis (GLFA) for mixtures of continuous and discrete items, and (iv) moderated non-linear factor analysis (MNFA) for mixtures of continuous and discrete items with non-linear dependencies, among others.

2.2.3. Synthetic data generation

The current state-of-the-art methods for virtual population/synthetic data generation can be classified into two major categories; the parametric methods which resample instances and generate new feature combinations from an existing clinical dataset, and the non-parametric methods where virtual patients are produced by randomly selecting patients from a clinical dataset. Examples of parametric methods include the multivariate normal distribution (MVND) and its variant the multivariate log-normal distribution originally proposed in [163] towards the generation of virtual patients based on real clinical data. The MVND was also deployed in [164] to create plausible virtual populations. A similar approach has been also introduced in [165], where the generated cohort data were able to match the observed data without the need for feature weighting. In [155], multinomial logistic models to model sequence count data with complex covariance structure. The fundamental assumption of the MVND though is that it assumes that the real data are normally distributed and thus poses significant biases in the synthetic data distribution although alternatives like the log-MVND have been proposed in the literature to better simulate the normality of the distribution [167].

Apart from the conventional statistical methods though, machine learning based methods have been also proposed. In [168] a package named “deal” was developed in

R, which includes Bayesian networks for virtual population generation by taking into consideration the conditional probabilities among the features, supporting both discrete and continuous type of data. Bayesian networks (BN) have been used in [169], [170] for the generation of virtual distributions based on the modeling of conditional probabilities across diverse network topologies. Bayesian resampling techniques were also described in [171] for the generation of virtual populations in the context of Pharmacokinetic (PPBK) and pharmacokinetic modeling. In the BN-based approaches, however, the conditional probabilities are modeled using assumptions on the prior distribution of the features, where the network topology is not pre-defined.

To this end, ML based generators like the artificial neural networks (ANNs) with radial basis functions, the supervised tree ensembles (STE), the unsupervised tree ensembles (UTE), have been proposed in the literature, yielding favorable performance against the probabilistic approaches. More specifically, Robnik-Šikonja [172], [173] utilized tree ensembles and artificial neural networks with radial basis functions (RBFs) as activation functions to detect hidden patterns among the features in the real data by either including or excluding a target feature yielding virtual data with decreased divergence with the real one. However, a major weakness of these data generators is that they are not computationally efficient since they require increased training and testing time, as well as, a target feature which introduces biases in the correlation patterns among the synthetic data.

Thus, there is an emerging need for the development of computationally efficient and unbiased virtual data generators remains a technical challenge, particularly in the case of large-scale clinical trials, where the computational complexity is important. Gaussian Mixture Models (GMM) were used to generate virtual data (imaging, omics and clinical) based on Dirichlet processes in [174]–[176] as an efficient data generator. Since GMM maximizes only the likelihood based on the expectation maximization (EM) approach, it might yield specific structures that might or might not apply to the data. A solution to this is to use variational inference (VI) [177]–[179] which maximizes a lower bound on the model evidence instead of the data likelihood like in the EM to reduce the computational complexity compared against the MVND, BN, UTE, STE, and ANN algorithms.

Table 3. A summary of the state-of-the-art methods and related applications for synthetic data/virtual population generation.

Study	Strategy	Application
[163]	Continuous and categorical covariate distribution modeling using multivariate statistical functions.	Their findings demonstrate that the MVND generates covariate correlations that are realistic and representative of the general population. The simulations included cohorts with 3500, 1000 and 100. The covariates presented in the matrix are respectively: smoking status, gender, disease severity, weight, height, and age.
[164]	Multivariate and discrete re-sampling techniques to account for covariate effects within the target population during the generation of virtual data.	This work presents a mechanism for quickly creating and selecting virtual patients to match clinical population-level statistics, which advances earlier methodologies. With all virtual patients weighted equally, the final fitted populations closely resemble empirical data, avoiding the possibility of overweighting specific solutions and distorting simulation outcomes present in some prior algorithms.
[165]	A technique for efficiently generating virtual patients that best fit the observed data using multivariate log-normal distribution (log-MVND).	Both evaluation techniques correctly produced the target population's proportions and summary data. In general, the Continuous method outperformed the Discrete method, apart from the few clinically relevant examples where the subgroups, defined by categorical value, had significantly different continuous covariate means. Instead of analyzing various subgroups, the Continuous technique permits examination of the entire population, which lowers the number of analyses required and boosts efficiency.
[155]	Application of multinomial logistic normal models for virtual population generation.	Through application to multinomial logistic-normal (MLN) models, the authors demonstrate that their inference scheme is both highly accurate and often 4-5 orders of magnitude faster than Markov Chain Monte Carlo (MCMC) methods.

Study	Strategy	Application
[158]	The authors propose the PCHC as a new Bayesian network learning approach that may be used with categorical or continuous data.	Max-Min Hill Climbing (MMHC) is far slower than PC Hill Climbing (PCHC), which can handle millions of observations in just a few minutes while maintaining accuracy levels that are on par with or better than MMHC. Additionally, PCHC has the excellent scalability attribute that its computing cost scales well with the size of the data sample.
[159]	The combination with deep learning techniques, specifically autoencoder networks, to reduce the dimensionality of our data. It further enables the application of BN structure learning to data of realistic sample size at reasonable computational cost.	Using this approach, the authors demonstrate that their simulated Alzheimer's (AD) and Parkinson's Disease (PD) VCs cannot be reliably discriminated from real patients in ADNI and PPMI. Furthermore, their method can be used to simulate a VC for a situation that has not been observed in the real data, e.g. a less cognitively impaired AD cohort.
[161]	The proposed generator is based on radial basis function networks, which learn sets of Gaussian kernels.	A large-scale empirical evaluation was performed using 51 data sets from UCI repository with great variability in the number of attributes, types of attributes, and number of class values. The results show a considerable similarity between the original and generated data and indicate that the method can be useful in several development and simulation scenarios.
[162]	Introduces the supervised and the unsupervised tree ensembles as high-quality synthetic data generators.	The proposed workflows were evaluated on random splits of several datasets and by comparing original datasets with datasets produced by a generator of semi-artificial data. The results show that the proposed workflows can reveal relevant similarity information about datasets needed in many data mining scenarios.

2.2.4. Federated/distributed learning

2.2.4.1. Algorithms

The conventional data integration strategy, where patients' data from different clinical centers are integrated into a centralized database is not always feasible either viable due to legal violations or security compromise attempts that will expose the patient data. A more technical and rather legal and ethical challenge in federated environments lies on the training of federated ML workflows across diverse data which are stored in federated databases [81], [112], [180]. Towards this direction, batch processing methods, such as, online learning and meta-learning [181], [182] have been proposed, where the former [170] uses stochastic optimization to update an existing estimator on upcoming training batches, whereas the latter [171] focuses on the aggregation of outcomes from models which are trained on each federated database. Meta-learning methods, however, limit the "horizon" of the training process since the individual ML models are trained on individual subsets [52]. Online learning methods, on the other hand, are restricted to the additive update of the weights of an existing ML model on new "online" training instances.

A solution to this is to use incremental learning [52], [129], [135], [183]–[185] which trains a classifier on an initial database, and then incrementally adjusts the weights of the classifier on a series of existing databases. Towards this direction, many incremental learning algorithms have been proposed including the family of the multiple additive regression trees (MART) [129], [186], the Support Vector Machines (SVM) [135], [187], the Multinomial Naïve Bayes (MNB) [129], [135], [188], [189], and additional stochastic gradient descent (SGD) based implementations [190]–[192]. In this case, the databases must have the same structure along with a common set of variables (features).

A common problem with the MART family of algorithms, however, is the fact that trees added early in the ensemble, at a particular stage, tend to have a higher impact during the decision-making process than those added later [193]. To this end, dropouts have been used [193] to deal with this issue, by scaling the most prominent trees in the ensemble, with a specific rate of rejected trees. A summary of the advantages and disadvantages of the above incremental learning algorithms is presented in Table 4. According to Table 4, the SGD-based approaches are computationally efficient but with

less classification performance and are prone to overfitting. In a similar, the MNB is also computationally efficient but also immune to overfitting.

Table 4. A summary of incremental learning implementations of existing ML schemas.

Incremental ML schema	**Advantages**	**Disadvantages**
SGD-based (SVM, regression)	Low computational complexity, simple to implement and deploy in federated environments	During federated training, poor classification performance and a tendency to overfit
Multinomial Naïve Bayes (MNB)	Low computational complexity, immune to overfitting, and simple to deploy in federated setups	Biases are introduced into the outcomes of a probabilistic approach due to a number of assumptions made about the independence of the collection of input features
Multiple additive regression trees (MART)	Favorable classification performance due to boosting's ability to reduce error, easy deployment in federated setups, scalability, and depth-first approach's ability to start pruning trees backward	Low to medium computational complexity, particularly with more boosting rounds
Multiple additive regression trees with dropout rates (DART)	Favorable classification performance because of the error reduction provided by boosting, simple deployment in federated setups, scalability, and the ability to use dropout rates, which can greatly improve the performance	Arbitrarily defined dropout rates and low to medium computational complexity, particularly with more boosting steps, can cause overfitting and cause the model's performance to be neglected

On the other hand, the independence assumptions introduce biases in the model. The MART algorithm incorporates a boosting stage which reduces errors and thus increases its classification performance particularly in demanding tasks. On the other hand, the

MART is more computationally complex especially in the cases of multiple boosting stages (rounds). It is important to note that the DART schema can further improve the classification performance of the MART due to the dropout rate which further boosts the decision-making process but on the other hand the dropout rate is arbitrarily defined and can yield significant overfitting effects during the training process.

2.2.4.2. Frameworks/platforms

Several distributed/federated learning platforms have been proposed in the literature [180], [194]–[205]. The euroCAT platform [202], [204] offers a distributed learning framework for the development of multi-centric models through the installation of local servers on the hospital's premises. The local databases and local learning connection are hosted on a server that is dedicated to the euroCAT network by a particular institution (a site) (Varian Medical Systems, Palo Alto, USA). The universal learning environment (Varian Learning Portal) connects the learning connections inside the sites to a central server (the master) outside the sites' IT architecture. Asynchronous messaging that is file-based connects the master and sites. Through a web browser-based interface, the user interacts with the learning environment and can upload learning applications (MATLAB, MathWorks, Natick, MA, USA) and start machine learning runs. Every learning application consists of two components: a site algorithm that operates inside the infrastructure of the sites and communicates with the learning connector, and a master algorithm that operates in the overall learning environment and can communicate with the site algorithms. A technical expert visited each premise to ensure the quality of the data in terms of inconsistencies and mistakes. The data stored at several sites is processed simultaneously and individually during each iteration. Then, each site transmits updated model parameters to the master. An algorithm at the master compares the model parameters and makes additional modifications. The algorithm also determines if the learning process has adequately converged (according to pre-set convergence criteria). The master sends the parameters back to each of the sites if the convergence criteria have not yet been satisfied. This completes one cycle of iterations. Up until the convergence conditions are met, the learning iterations are continued.

The Personal Health Train (PHT) platform [203] adopts a similar methodology for distributed analysis through the training of distributed logistic regression models with

adequate performance. With the help of the PHT, which connects FAIR (Findable, Accessible, Interoperable, Reusable) data sources, distributed data analysis and machine learning are made possible. A healthcare facility never loses access to patient data. Cohort discovery is the goal of the first application group. To determine and disseminate general statistics (counts) of the data that are available in the FAIR data station, an application is issued to each site. A SPARQL Protocol and RDF Query Language (SPARQL) query that may be run against the graph database is included in this cohort discovery application. A master application running at the VLP receives reports from each site application on its site statistics, which are subsequently forwarded to the researcher who created the application. This application group was used in a variety of ways to produce summary data for patient subgroups. A logistic regression (LR) model is intended to be trained by the second application group. Given a SPARQL query, each LR site application can train an LR model using the local dataset. The master application then iteratively comes to a consensus using the regression coefficients of each site LR model and patient counts. A particular LR model is validated on the sites by the third application group. To calculate model performance metrics (RMSE, ROC curve, AUC, calibration plots), an application is sent to each site. These metrics are then transferred back to the master application, which combines them and sends them to the researcher.

The Open Federated Learning (OpenFL) [205] is a software platform for federated learning (FL) that was first created as a component of a joint research project between Intel Labs and the University of Pennsylvania on FL for healthcare. Intel and the open-source community in GitHub continue to develop OpenFL for general-purpose real-world applications. Although healthcare was the initial use case, the OpenFL project is made to be independent of use cases, industries, and ML frameworks. The open-source code, which is primarily written in Python, is delivered through pip, conda, and Docker packages. The solution enables programmers to train ML models on remote data owners' nodes (aka collaborators). On the hardware at the collaborator node, the ML model is trained. Artificial neural networks trained using either TensorFlow [199] or PyTorch [206] are current examples. Through an extendable methodology, additional ML model libraries and neural network training frameworks can be supported. Only the model weight updates and metrics are communicated to the model owner via the aggregator node; the training data is always kept at the collaborator node. The setup

and workflow are described in a FL plan. All nodes in the federation share this FL plan, which outlines the federation's regulations. The OpenFL design, which was first created in the Intel Labs Security and Privacy Research lab, prioritizes important security ideas like limited interfaces, code reuse, open-source code, streamlined information security reviews, and code design suitable for running on trusted computer hardware, like a trusted execution environment (TEE). The Federated Tumor Segmentation (FeTS) initiative is a current project of the largest international federation of healthcare organizations with the goal of learning about tumor boundary identification from vast and varied patient populations without disclosing any patient data. A dedicated open-source platform with an intuitive graphical user interface was created to support this initiative. Its goals are to: I bring cutting-edge pre-trained segmentation models of numerous algorithms and label fusion approaches closer to clinical experts and researchers, enabling easy quantification of new radiologic scans and comparative evaluation of new algorithms; and ii) enable multi-institutional collaborations via FL by level. FeTS has been originally used to segment brain tumor sub-regions across n = 56 clinical locations located all over the world.

PySyft [201] is an open-source multilingual library that enables safe and private machine learning by transparently wrapping and expanding well-known deep learning frameworks like PyTorch. Its goals are to be extensible so that new Federated Learning (FL), Multi-Party Computation, or Differential Privacy methods can be flexibly and easily implemented and integrated, as well as to help make privacy-preserving techniques in machine learning as accessible as possible via Python bindings and common tools familiar to researchers and data scientists. The methods offered by the PySyft library will be introduced in this chapter, along with details on how they are implemented. Then, using a convolutional neural network training example, we will present a proof-of-concept demonstration of a FL procedure. The application of PySyft in academic literature is then reviewed, and prospective use cases and development strategies are covered. We highlight Duet, a solution for simpler FL for scientists and data owners.

FedML [200] is an open research library and benchmark that makes it easier to create FL algorithms and compare their performances fairly. Three computing paradigms are supported by FedML: distributed computing, single-machine simulation, and on-device

training for edge devices. Additionally, FedML encourages a variety of algorithmic research through the architecture of flexible, generic APIs and thorough implementations of reference baselines (optimizer, models, and datasets). We are hopeful that FedML will offer a productive and repeatable way for FL researchers to create and assess FL algorithms. We look after the user community, documents, and source code.

Paddle Federated Learning (PFL) [197] is an Apache 2.0-licensed FL framework that is available as open source. It makes use of the deep learning (DL) platform PaddlePaddle [207]. PFL can handle both vertically and horizontally partitioned data. Each type of data splitting has its own set of implemented algorithms. Use the MPC package for processing vertical data partitions and the paddle fl package for processing horizontal data partitions. The NN and LR models are part of the paddle fl package. To construct them, PFL employs the FedAvg [208], SecAgg [209], and differentially private stochastic gradient descent (DPSGD) methods. A differentially private method called DPSGD protects the privacy of data. Processing of horizontal data partitions is carried out using a centralized system. PFL supports both simulation and federated deployment modes, and it is advised that Docker containers be used in both cases. PFL needs at least 6 GB of RAM and 100 GB of HDD space to function effectively. The use of PFL in IoT systems is constrained by these criteria.

A deep learning framework for decentralized data called TensorFlow Federated (TFF) is open source [199]. The most recent version of TFF (0.17.0) learns, estimates, and uses NNs using TensorFlow (TF) of version 2.3. The use of GPUs is not supported, nevertheless. It is made available using the Apache 2.0 license. TFF implements base classes for the FedAvg and FedSGD algorithms [210], a straightforward federated evaluation implementation, and federated personalized evaluation. TFF includes a core API for the development of new federated algorithms. It is made up of classes that define templates for stateful activities such as value aggregation, estimate computation, and metric production. Using them, an analyst may create their own custom analytical procedures. The current TFF 0.17.0 version is incomplete and still needs key crucial components needed for the framework to be used in real applications: Only a differential privacy method is employed; the federated mode of operation is not

implemented; vertical and hybrid data splitting is not supported; the decentralized architecture of the system is not supported.

A smoother transition from experimental study in simulation to system research on a large cohort of actual edge devices is made possible by Flower, a revolutionary end-to-end federated learning platform [195]. In terms of simulation and real-world devices, Flower offers individual strengths in both areas. It also provides the flexibility for experimental implementations to move between the two extremes as needed during exploration and development. Flower is a cutting-edge FL framework that allows large-cohort training and assessment on single-node or multi-node compute clusters as well as on actual edge devices. As a result, scalable algorithmic investigation of real-world system conditions, such as constrained computational resources that are typical for FL workloads, becomes possible.

LEAF, a framework for measuring modular learning in federated environments [194]. A collection of open-source federated datasets, a stringent evaluation system, and several reference implementations are all included in LEAF with the goal of capturing the complexities and challenges of real-world federated contexts. It comprises of a set of reference implementations, a collection of open-source datasets, and a range of statistical and system metrics. LEAF's modular construction enables these three elements to be quickly added to a variety of experimental pipelines.

Table 5. Description of the existing frameworks/platforms for federated learning.

Framework/ platform	**Description**	**Application**
euroCAT [202], [204]	A distributed learning framework for the development of multi-centric models through the installation of local servers on the hospital's premises.	Bayesian networks and Support Vector Machines which were trained across 3 centers to predict dyspnea yielding modest prediction performance.
PHT [203]	Adopts a similar methodology for distributed analysis through the training of distributed logistic regression models.	A distributed logistic regression model was trained across 8 sites to predict post-treatment with adequate performance.

Framework/ platform	Description	Application
OpenFL [205]	A software platform for federated learning (FL) supporting ANNs for demanding applications like image segmentation tasks based on deep learning.	Enabled multi-institutional collaborations via FL by leveraging OpenFL to improve these pre-trained models without sharing patient data, thereby overcoming legal, privacy, and data-ownership challenges, FeTS has been initially deployed towards the task of brain tumor sub-region segmentation by partnering with n = 56 clinical sites spread all around the world.
PySyft [201]	An open-source multilingual library that enables safe and private machine learning by transparently wrapping and expanding well-known deep learning frameworks like PyTorch.	FL system designers can make good use of our performance model to analyze the performance under their FL scenarios and establish an efficient and balanced FL system without trial-n error cost.
FedML [200]	An open research library and benchmark that makes it easier to create FL algorithms and compare their performances fairly.	The obtained accuracy at round R was equal to 0.77. The accuracy of the model increases linearly as defined by linear interpolation. The strategy was able to get an accuracy of 0.68 in 10 rounds, increasing the number of epochs from round 8 in the Fashion-MNIST dataset.
PFL [197]	An Apache 2.0-licensed FL framework that is available as open source. It makes use of the deep learning (DL) platform PaddlePaddle [15] and can handle both vertically and horizontally partitioned data.	Not reported. Using elastic scheduling of training job on Kubernetes and large-scale distributed training of PaddlePaddle's, paddle FL can be easily deployed on full stack open sourced software.
TFF [199]	An open-source deep learning framework for decentralized data.	On a 5-client distributed dataset, the best character accuracy is achieved

Framework/ platform	Description	Application
	The most recent version of TFF (0.17.0) learns, estimates, and uses NNs using TensorFlow (TF).	by TFF at 49.20%. Extensive experiments are also conducted to evaluate the effect of distributed data storage over the performance of trained models. TFF again achieved a maximum character precision of 54.33% with non-distributed dataset.
Flower [195]	A cutting-edge FL framework that allows large-cohort training and assessment on single-node or multi-node compute clusters as well as on actual edge devices.	Flower was deployed on 10 Android clients to train a model with 2 convolutional layers and 3 fully-connected layers (Flower, 2021) on the CIFAR10 dataset achieving 0.67 accuracy in 80.32 ms. w Flower can perform FL experiments up to 15M in client size using only a pair of high-end GPUs.
LEAF [194]	LEAF, a framework for measuring modular learning in federated environments.	LEAF's modularity was assessed across three datasets yielding accuracy 89.64% in CelebA, 71.89% in Synthetic, and 74.72% in FEMNIST.

2.3. Clinical point of view

In this section we will present the current clinical state of the art regarding the existing open issues and unmet needs across the six clinical domains which will be investigated under the aegis of this thesis: (i) the autoimmune diseases (AD), (ii) the hypertrophic cardiomyopathy (HCM), (iii) the systemic autoinflammatory diseases (SAIDs), (iv) the coronavirus disease (COVID-19), (v) the cardiovascular diseases (CVD), and (vi) the mental disorders (MD).

2.3.1. Autoimmune diseases (AD)

The lack of data quality is a major threat in the domain of autoimmune diseases, where the rarity of the endogenous disease subtypes poses significant obstacles in the

application of AI-based approaches towards the development of robust patient risk stratification models, and the identification of biomarkers, among others. Examples of subtypes of autoimmune diseases along with classification/prediction applications, include: (i) multiple sclerosis [211], [212] for diagnosis and prognosis, (ii) rheumatoid arthritis [213] for risk assessment, (iii) systemic lupus erythematosus [214] for variations of prognosis, (iv) psoriasis [215] for diagnosis and disease severity, (v) systemic sclerosis [216] for diagnosis, treatment and prognosis, (vi) thyroid diseases for diagnosis [217], and (vii) autoimmune liver diseases for prognosis [218], among others, containing no more than 600 samples. Indicative data types that were used in the previous domains include clinical, survey, gene expression, gait, proteomic, microbiome, and peptide, among others. This emerging need for data quality has been extensively highlighted in [125].

Furthermore, the underlying data heterogeneity obscures the co-analysis of diverse autoimmune diseases data sources and therefore the existing studies in the field focus on the analysis of low-quality datasets with small statistical power and significant assumptions regarding the independence of the variables in the conventional multivariate regression analysis (these issues have been highlighted in [129]). On the other hand, the lack of sufficient population size in the domain of autoimmune diseases is emerging and poses a significant barrier which affects the statistical power of the outcomes from the analysis of small populations [129]. Moreover, the existing data silos hamper the application of AI-empowered workflows since they leverage the sharing and interlinking of data across multiple centers [219], [220] and consequently have a negative effect on the explainability of the AI models for disease progression and treatment in all the above subtypes of autoimmune diseases. In addition, they undermine efforts to characterize, predict, and mitigate missing person incidents [221].

Primary Sjögrens Syndrome (pSS) is one of the most common chronic systemic autoimmune diseases, affecting the lacrimal, salivary glands and other exocrine glands, such as, the larynx, trachea, skin, and vagina [222]–[224]. In fact, pSS is unique not only due to its clinical impact but also as one of the few disease “models” linking autoimmunity with cancer and especially lymphoproliferative disorders. Its main difference with the secondary Sjögren Syndrome (sSS) lies on the fact that in the latter, the patient also exhibits other rheumatic diseases, such as, rheumatoid arthritis (RA),

and systemic lupus erythematosus (SLE) [222]–[224]. The annual incidence of Sjögren's Syndrome (SS) among North and South European populations has been estimated from 200 - 3000 per 100000 individuals, with the corresponding figures for RA and SLE being 200 - 900 and 20-70 individuals, respectively [225].

According to the literature [226]–[230], pSS has the most unbalanced gender ratio with almost 10 females affected per 1 male while the development of B-cell non-Hodgkin lymphoma (NHL) complicates about 5% of patients during the disease course [226]–[230]. Female preponderance, *peri*-epithelial lymphocytic infiltration of the affected organs, B-cell hyperactivity manifested as hypergammaglobulinemia, as well as, activation of interferon and B-cell activating factor pathways are considered hallmarks of the disease. Although the cause of pSS remains unknown, the disease develops in the context of genetic, environmental, and immune factors. Previously suggested histopathological, as well as, clinical laboratory risk factors for lymphoma development in terms of prognostic and diagnostic purposes include the salivary gland enlargement (SGE), the rheumatoid factor (RF), the cryoglobulinemia, the germinal centers that are present during salivary gland biopsy, the C4 hypocomplementemia, and the purpura, among others [223], [224], [231]–[234].

Additional risk factors that have been extensively reported in several clinical studies [235]–[240], as prominent determinants for various phenotypic infiltrations related to lymphoma development, include the anti-Ro/SSA and/or anti-La/SSB autoantibodies, the lymphadenopathy, the monoclonal gammopathy, and the Raynaud phenomenon. As in other systemic autoimmune or neoplastic diseases, the lack of patient stratification models: (i) increases the risk of producing unsatisfactory or sub-optimal results in clinical trials employing novel and expensive drugs, and (ii) hampers the definition of evidence-based health policies. These two issues are related with the unmet needs in pSS which involve the development of robust lymphoma classification models and the extraction of biomarkers. The clinical unmet needs in pSS include the development of lymphoma classification and lymphomagenesis models, as well as, the extraction of prominent indicators for lymphoma development. Besides, the challenge of harmonization and integration of cohorts in pSS has been highlighted in [241] along with the necessity of the validation of the existing biomarkers and the discovery of new biomarkers in large-scale cohort studies in [242].

Only a few relevant studies have been reported in the literature concerning the design and application of lymphoma classification models, as well as, the discovery of biomarkers for lymphoma development and progression. Most of these studies adopt univariate and multivariate statistical methods (including time-to-event models) [235], [243]–[247] to identify independent risk factors for lymphoma development which in turn are utilized as independent variables for regression analysis with the dependent variable usually being set to lymphoma. A more straightforward method for the detection of risk factors was presented in [233], where the fast correlation-based filter selection (FCBF) method was deployed to identify robust independent factors for lymphoma development, following a logistic regression analysis. Furthermore, supervised machine learning methods [129], [135], [248]–[250], such as, the supervised tree ensembles, the Support Vector Machines (SVMs), and the artificial neural networks (ANNs) have been utilized in the literature for the development of robust lymphoma classification models in pSS with adequate performance. However, these studies have poor statistical power due to the reduced population size, since they adopt either a single cohort analysis approach in [226], [234]–[239], [241] or a small-scale but straightforward analysis involving no more than four cohorts in [129]. The reduced quality, and the structural heterogeneity of the existing cohorts along with the lack of data curation pipelines obscure the development of robust AI models and the detection of biomarkers.

2.3.2. Hypertrophic cardiomyopathy (HCM)

The leading cause of death worldwide is cardiovascular disease (CVD) [251], [252]. One in 500 people in the general population have hypertrophic cardiomyopathy (HCM), a frequent subtype of cardiovascular illness [253]–[256]. HCM is an inherited illness since it has a genetic component. The most prevalent hereditary cardiovascular condition is this one [257], [258]. It can result from any one of 1,400 mutations in 11 or more genes that code for cardiac sarcomere-related proteins [259]. Due to the lack of any noticeable symptoms, HCM is one of the most common causes of sudden cardiac death among young people and sports [260]. Most patients with the illness are unaware that they have it [261].

Only 15 nations are involved in the execution of clinical studies with an average duration of 3 years for hypertrophic cardiomyopathy (HCM), despite 122 of the 191

countries in the world showing a disease burden [262]. In fact, the creation of cost-effective treatments is hampered by the small population size mixed with the poor data quality. The necessity for drug development for in silico clinical trials is growing because they are expensive and necessitate a large enough population. Where the urgent need for the creation of risk stratification models and biomarkers for the development and progression of HCM is pressing, [262], [263] have highlighted the lack of population size.

Pharmacological therapy hasn't progressed beyond its initial goals of straightforward symptom relief and functional capacity improvement, despite major advancements in the management of the condition with interventional procedures, device installation, and surgery [264]. In fact, only 45 trials (i.e., less than 1 per year) with a total of 2,121 HCM patients were found in a recent review of all literature relevant to any pharmaceutical regimen ever used to treat HCM [265]. No pharmaceutical (medical) intervention has so far been shown to lower the risk of sudden cardiac death or increase patient survival [264]. The development of novel therapy regimens is essential for improving survival outcomes, clinical outcomes, and clinical benefits. The development of novel therapy regimens is essential for improving survival outcomes, clinical outcomes, and clinical benefits. Therefore, it is crucial to conduct clinical trials to examine and assess new treatments and to give HCM patients the chance to take part in those trials. By doing this, we will be able to produce important data about HCM management and cut back on spending on unproductive management strategies.

2.3.3. Systemic autoinflammatory diseases (SAIDs)

Systemic autoinflammatory diseases (SAIDs) are a set of evolving groups of conditions sharing a core of phenotypical similarities [266], [267]. They encompass several rare disorders which have been characterized by extensive clinical and biological inflammation, with no specific age or gender distribution in the human population. Genetic mutations that may cause dysregulation of the innate immune system underlie the etiology of some SAIDs. Although they were proposed to constitute a continuum of disorders with potential overlap, SAIDs should not be confused with the autoimmune family of diseases, related to adaptive immune system dysfunction and response to self-antigen(s) [268]. Primary physical manifestations of SAIDs typically involve fever, rash, joint involvement, lymphadenopathy, and musculoskeletal symptoms. Due to the

numerous symptoms observed in the different SAID-related conditions and their lack of specificity, diagnosis is challenging. Unlike autoimmune diseases whose autoantibodies are a tool for ascertaining the diagnosis, there is no known constitutive and disease-defining biomarker for SAIDs. Although inflammasome activation is thought to be a common pathophysiological pathway, the complex network of cytokine cascades together with multiple cell type activation makes difficult the use of these features as diagnostic or classification markers for SAIDs.

2.3.4. Coronavirus disease 2019 (COVID-19)

Among the infected individuals with SARS-CoV-2 [269], it is estimated that 1/3 of them never develop symptoms [270], [271] and those who will develop symptoms may have a mild to moderate self-limiting disease [271]. In contrary, the severity of symptomatic infection ranges from mild to critical, and most individuals will develop a non-severe illness [272]. The progression of the disease and the risk of severe illness varies by age, underlying comorbidities, and risk factors for disease progression, such as, cardiovascular diseases (CVD), diabetes mellitus (DM), chronic obstructive pulmonary disease (COPD), cancer (e.g., hematologic malignancies, lung cancer), chronic kidney disease, solid organ or hematopoietic stem cell transplantation, obesity, and smoking [273]. According to the official report of the Centers for Disease Control and Prevention (CDC) in the US, among 1.3 million confirmed COVID-19 cases, 14% of patients were hospitalized, 2% were admitted in the intensive care unit (ICU), and 5% died [274]. In addition, the risk of critical or fatal disease is high among hospitalized COVID-19 patients [275], [276].

The increased need for intensive care units and ventilators due to the unprecedented number of confirmed COVID-19 cases has surpassed the capacity of international healthcare systems. As a result, the World Health Organization (WHO) highlighted the importance of artificial intelligence (AI) as a prominent solution to manage the crisis caused by the virus [277]. AI is a constructive, non-medical intervention approach with a strong potential to overcome the current global health crisis, build next-generation epidemic preparedness, and move towards a resilient recovery [277]. Moreover, AI can shed light into the clinical unmet needs in COVID-19, including the development of robust models for: (i) the prediction of ICU admission, mortality, and the need for mechanical ventilation, (ii) the extraction of prominent risk factors for ICU

admission and mortality, (iii) the early suggestion of targeted interventions/therapeutic treatments, and (iv) the definition of better disease severity indices. Although AI is a promising tool to unveil the underlying mechanisms of COVID-19, the risk of bias and discrimination in its design and deployment must be taken into consideration.

2.3.5. Cardiovascular diseases (CVD)

There are still problems that need to be addressed and resolved despite the employment of sophisticated statistical tools, as was already noted. Big data research may present a distinct perspective, however there may be some discrepancies between the results of small, well-conducted studies and randomized controlled clinical trials. Such differences could be a result of the particular characteristics of the database utilized in the study, as each cardiological database notably differs in terms of the techniques employed to gather and capture data and the population(s) it specifically represents [278]. Data quality may also be impacted by the database's structure (organized vs. unstructured). For example, Hernandez-Boussard et al. [279] mined a dataset containing 10,840 clinical notes and discovered lower recall and precision rates (51.7% and 98.3%, respectively) for structured electronic health records (EHR) compared to unstructured EHRs (95.5% and 95.3%, respectively), which justifies routinely measuring recall for each database/registry before moving forward with data processing and analysis.

2.3.6. Mental disorders (MD)

One of the most significant issues currently facing public health is mental illness [280], [281]. These illnesses affect hundreds of millions of individuals globally and are linked to significant transgenerational transmission [282], [283], to huge economic costs [284], to elevated rates of physical morbidity and mortality [285], and profound personal suffering for patients and their families. However, it is still unclear exactly what these illnesses are. There are no definite cutoff points for when a patient has a disorder and when they do not, nor are there any objective tests or measurements to determine the presence of a mental disorder. The Diagnostic and Statistical Manual of Mental Diseases (DSM) and the International Classification of Diseases (ICD) have been the two main classification and definition systems for mental disorders in recent decades (ICD). The various variants of these systems have been used in the majority of

the existing studies on mental disorders during the past few decades, although they have drawn heavy criticism.

There is evidence, for instance, that the majority of mental diseases should not be viewed as distinct entities but rather as a set of qualities, on which some individuals score well and others poorly [286]. Additionally, significant degrees of comorbidity are more often than not [287]. Some contend that the diagnostic classifications in the DSM and ICD have a narrow range of applicability [288]. Additionally, most treatments are beneficial for several illnesses rather than just one, as is the case with cognitive behavior therapy (CBT) for the majority of mental disorders and pharmacotherapies for mood and anxiety disorders. Therefore, what should be the targets of treatments and how can their effectiveness be measured if we do not yet fully understand what these disorders are and how they should be defined? Treatments' overarching objectives are, of course, to improve patients' conditions or assist them in coping with their issues. But it's not entirely apparent what this entails or when it can be said to have been completed. Not only are the nature and origins of the disorders unknown, but the response also varies on who is asked: the patient, the doctor, the patient's family, the health insurance industry, or society at large.

The prevalence of depression, anxiety, eating disorders, and other mental problems among college students has increased recently. The need for counseling services has also been steadily increasing at the same time. Some people have interpreted these patterns as a mental health emergency that demands immediate examination and the development of potential remedies to meet the requirements of pupils. The prevalence of personal computer technologies, such as social media, and the subsequent growth in symptomatology have been connected in several studies, and it has been hypothesized that time spent using these devices is directly associated to poor mental health. Although the use of personal computing technology has altered how college students interact with one another and may have some negative effects on mental health, these same technologies also have several advantages for improving mental health and treating mental disease. Here, we discuss the difficulties and possibilities that personal computer devices present for the mental health of college students. We emphasize chances for new research in this field as well as chances for people and organizations to use these technologies in more beneficial and health-promoting ways.

Depressive disorders are widespread, expensive, significantly reduce quality of life, and are linked to high rates of morbidity and mortality. Antidepressant drugs and talking therapy are listed as first-line treatments in most guidelines because they are effective treatments for depression. These treatments have improved the lives of countless patients around the world and will do so for many years to come. Although some patients respond well to therapies, there is still much potential for improvement. This Comment presents ten significant data points about the limitations of depression treatment outcomes that, in our opinion, demand more focus. We cannot blame our ignorance on a lack of research into accepted medical practices. Over the past few decades, more than 600 randomized trials have looked at the effectiveness of depression psychotherapies and more than 500 have looked at the effects of antidepressant medicines (although comparatively few are conducted for early-onset depression).

The results are questionable since less than 20% of medication studies and less than 30% of therapy trials have minimal risk of bias. Such trials typically lack the statistical power to determine who a treatment is helpful for, leaving no solid proof of who will benefit from which treatment the most. Additionally, because there are so many distinct outcome measures utilized in treatment research, it is impossible to combine trial results without adding noise. Additionally, most trials do not look at long-term impacts. Despite more than a thousand trials, there are still many fundamental questions that affect both persons who are depressed and those who are trying to help them in real life.

CHAPTER 3. DATA CURATION

3.1. Overview

The technological advances of our era have dramatically increased the amount of generated digital data [2], [4], [83], [119]. The overwhelming need to improve the quality of complex data structures in multiple disciplines varying from the industrial and financial sector to the healthcare sector is more important than ever [66], [68], [104], [125], [136], [137]. This need has led to an emerging demand for the development of automated methods for the quality assessment of big data structures since poor data quality results in data poisoning which makes data useless and hampers further processing yielding poor scientific results. As a result, the design, development, and deployment of automated computational methods for data quality assessment is a great technical challenge.

The data quality assessment process can be considered as a core operation prior to the application of any data analytics workflow. Data curation [66], [70], [71], [104], [106], lies on the core of the data quality assessment process. Its primary focus is to enhance the quality of raw complex data structures by transforming them into high-quality data that fulfill certain data quality indices. According to the literature, a set of qualitative metrics is usually defined to quantify the quality of the data. Examples of such quality metrics include the [52], [66], [106]: (i) consistency, (ii) completeness, (iii) accuracy, (iv) auditability, (v) orderliness, (vi) uniqueness, and (vii) timeliness, among others.

More specifically, consistency refers to the lack of contradictions in the data. Accuracy refers to the percentage of reasonable information in the data. Completeness refers to the extent at which missing values are present in the data whereas auditability refers to whether any changes in the data can be traced or not. Orderliness refers to whether the data conform to a pre-defined format or structure. Uniqueness refers to the extent of the duplicated entries in the data and timeliness refers to the extent at which the data follow a correct timeliness. Data curation can be used as guidance for fixing recording errors which are not easy to be detected, especially when dealing with large scale and highly complex data structures.

In healthcare, the prospects of developing automated data curation workflows are many: (i) they can overcome the complexity of processing medical data, especially big data, where the conventional manual data curation is not feasible, and (ii) they can ensure the reliability and applicability of automation by offering reusable and clinician-friendly data quality reports that can be used for data diagnostics. A data curation workflow consists of a series of steps, including: (i) the development of memory efficient data parsing methods in the case of big data structures, (ii) the detection and elimination of duplicate fields (de-duplication) and fields with highly similar distributions (similarity detection), (ii) the characterization of data according to their context (data annotation), (iii) the identification of data inconsistencies, (iv) the management of missing values (data imputation), and (v) the detection of data anomalies, i.e., values that deviate from the standard data range.

3.2. Beyond the state of the art

According to Table 1, the existing quality metrics, however, are mostly qualitative and not quantitative enough to be considered as part of a computational workflow for data curation. Besides, most of the existing studies from Table 1 focus on providing general guidelines for data quality assessment [136], [289], [290] and methodological steps towards data curation, without, however, focusing on the development and evaluation of a computational framework for data quality assessment on medical data. While the variety of the proposed univariate and multivariate methodologies towards outlier detection [291], [292] lack of an integrated approach that manages to combine both methodologies into a single framework, the presented framework offers an integrated service that includes outlier detection as part of its data quality control strategy.

Meanwhile, most of the clinical studies for data quality assessment [136]–[140] aim to construct a gold standard model (a set of terms which describe the knowledge of a clinical domain) and then use this model to manually or semi-automatically classify the terms of a raw clinical dataset based on their accuracy, relevance, consistency, etc., with the terms of the gold standard model. In addition, most of the proposed methods (Table 6) are not fully automated and are qualitative rather than quantitative to be considered sustainable and viable in terms of their integration into computational workflows for data analytics purposes. Moreover, they provide a series of arbitrary methodological steps which are constrained by predefined semantic representation models for semi-automated quality control and manual data entry, such as, the ExeTera tool [70] and the universal C-Surv model [140]. All in all, the existing frameworks (Table 6): (i) do not use any automated methods towards outlier detection and de-duplication, (ii) focus only on assessing the quality of the terms that are relevant with those from the gold standard model, and (iii) do not provide re-useable data quality assessment reports.

To address these needs, we propose an integrated framework for medical data curation in terms of data quality assessment. The framework consists of a three-layer architecture and serves as a diagnostic tool for managing incomplete terminologies, irrelevant terms, outliers, missing values, data categorization, and duplicated terms. In this work, we extend data standardization as a pre-harmonization process to make data harmonization easier and faster. More specifically, we use lexical matching combined with model-based rules and external sources, i.e., vocabularies, to match and classify terms according to a pre-defined reference model which is a set of parameters which describe the requirements (variables with their types and ranges) of the clinical domain of interest. Through this procedure, we attempt to produce semantic relations between the fields of the raw dataset with those from a reference model and therefore enhance the semantic matching process for data harmonization. The proposed framework accounts also for data standardization since it can produce a set of semantic relations through a rule-driven approach that is developed based on a pre-defined reference model and captures important semantic relations which enable faster data harmonization. In addition, the framework can be easily adjusted with new rules according to a provided reference model that describes the clinical domain of interest.

Table 6. Open issues in the studies from Table 1 and how they are addressed by the proposed framework for medical data curation.

Study	Issues	Proposed framework
[136]	• Lack of quantitative methods for data quality control (e.g., outlier detection, similarity detection). • Conceptual presentation of the methodology for matching terms across patients/records.	• Provides a set of quantitative functionalities to enhance the completeness, relevance, and accuracy of clinical data • Includes functionalities for metadata extraction, outlier detection, de-duplication, and data standardization (based on a set of terms that are lexically matched with those from a standard reference model). • Produces re-usable data quality reports that can be used to fix outliers, duplicates, missing values, and inconsistencies. • Can be iteratively executed until the data quality criteria are met. • Web-based (REST service). • Data standardization in terms of data harmonization.
[137]	• Lack of case studies to prove the superiority of the proposed set of terms against similar ones • Only qualitative measures are defined for quality improvement. • Lack of quantitative methods for data quality control. • Lack of re-usable quality reports.	
[138]	• The quality assessment process is exclusively based on quality criteria that are manually defined for each individual data source. • Lack of quantitative methods for data curation.	
[139]	• Lack of quantitative methods for data quality control • Conceptual presentation of the methodology for matching terms across patients/records	
[70]	• Lack of quantitative methods for data quality control. • Lack of re-usable quality reports. • Focuses on a particular data schema for semantic matching of existing information.	
[140]	• The quality assessment process is exclusively based on quality criteria that are manually defined for each individual data source. • Lack of quantitative methods for data curation.	

3.3. The proposed automated framework for medical data curation

The proposed framework for data curation consists of a three-layer architecture which receives as input a raw dataset and outputs the data evaluation report which provides information related to the data quality and the data standardization outcomes, and the curated dataset (Figure 5). The architecture is comprised of three modules: (i) the data evaluation module, (ii) the data quality control module, and (iii) the data standardization module which serves as a pre-harmonization step.

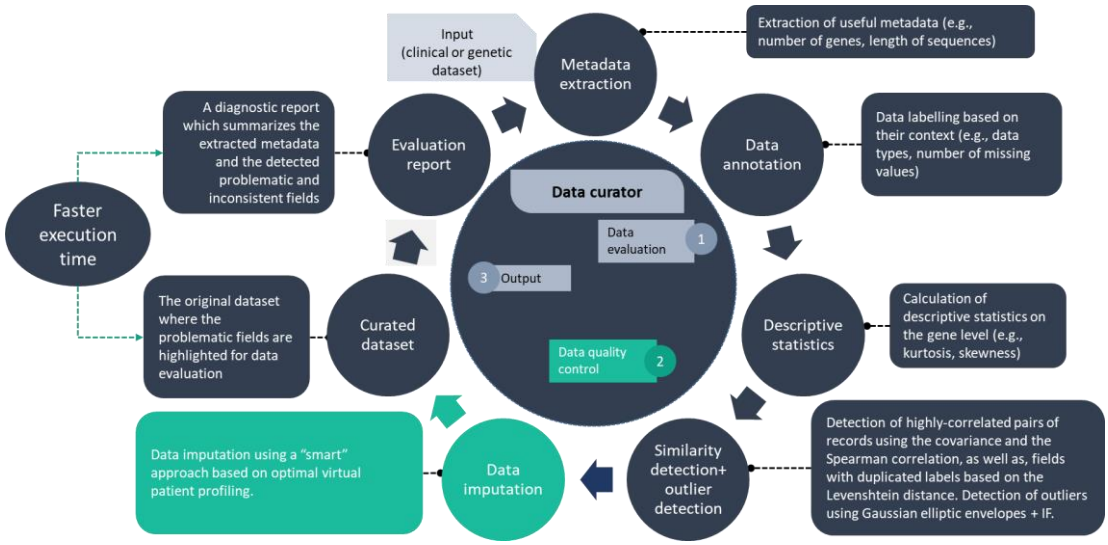


Figure 5. The proposed data curation workflow.

3.3.1. Metadata extraction and data annotation

The data evaluation stage aims to annotate each feature according to its variable and data type, as well as, provide summary metadata and descriptive statistics. During the data annotation, each feature is handled separately in order to determine its variable type (i.e., integer, float, string, date, or unknown) and data type (i.e., numeric or categorical), as well. The variable type of a feature is determined by taking into consideration the range values of that feature. In addition, the value range of each feature varies according to the types of variable types that are present in that feature. For example, if a feature includes only values that are integers, its variable type is set to "int", and if a feature includes values that are integers and floats, its variable type is set to "float". However, if a feature contains values with mixed types, e.g., integer and string data types, as well as, unknown data types, such as, symbols (e.g., "+"), the variable type is set to unknown. The categorical features are defined as those having

two possible values (e.g., the binary features), whereas the numeric features are defined as those having more than two possible values. In case where the feature contains values with unknown variable type, the data type is set to unknown like before. A numeric feature is further characterized as continuous whereas a categorical feature is characterized as discrete. Descriptive statistics can be also computed only on those features that have integer, date, and float variable types, where the mean, kurtosis, skewness, median, maximum, and minimum values are computed on a feature-basis.

3.3.2. Descriptive statistics

Useful descriptive statistics are calculated, on a feature basis, including the mean, median, standard deviation, variance, skewness, and kurtosis. Features with unknown data types are excluded from this process. A table that summarizes these measures is presented in Table 7.

Table 7. Conventional descriptive statistic measures.

Measure	**Abbreviation**	**Mathematical formulation**	**Short description**
samples	$\boldsymbol{x}$	-	-
sample size	N	-	-
mean	μ_x	-	measures the central tendency of a probability distribution
median	m_x	-	the "middle" value of a sorted set of samples that separates the upper half from the lower half
standard deviation	σ_x	$\sqrt{\frac{1}{N-1}(\boldsymbol{x}-\mu^2{}_x)/\sigma}$	measures the statistical dispersion of a set of samples around the mean
variance	$\sigma^2{}_x$	$\frac{1}{N-1}(\boldsymbol{x}-\mu^2{}_x)/\sigma_x$	measures the average of the squared difference of the samples from the mean
skewness	s_x	$E\left[\left(\frac{\boldsymbol{x}-\mu_x}{\sigma_x}\right)^3\right]$	measures the "tailedness" of a probability distribution (third standardized moment)

Measure	Abbreviation	Mathematical formulation	Short description
kurtosis	κ_x	$E\left[\left(\frac{\boldsymbol{x}-\mu_x}{\sigma_x}\right)^4\right]$	measures the asymmetry of a probability distribution (fourth standardized moment)

3.3.3. Outlier detection

Outlier detection, also referred to as anomaly detection, aims at separating a core of regular observations from some polluting ones, known as the outliers, which vary from the majority. According to the literature, a large variety of both univariate and multivariate methods have been proposed so far, some of which are discussed in the sequel. Most of these methods are standard approaches applied by clinical laboratories.

3.3.3.1. Statistical approaches

3.3.3.1.1. z-score and modified z-score

Another widely used statistical univariate measure for outlier detection, is the z-score, which quantifies the distance between a feature's value and its mean value [293]. It is defined as in:

$$z = \frac{\boldsymbol{x} - \hat{x}}{\sigma_x}, \tag{3.1}$$

where $\boldsymbol{x}$ is the feature vector, $\hat{x}$ is its mean value, and σ_x is its standard deviation. In practice, features with z-values larger than 3 or smaller than -3 are considered as outliers [52]. However, the z-score might lead to misidentified outliers since the maximum score is equal to $(n-1)/\sqrt{n}$, yielding small values due to the non-robustness of the standard deviation that is used in the denominator, especially in small size data. For this purpose, a modified version has been proposed [293]:

$$z_{mod} = \frac{\boldsymbol{x} - \tilde{x}}{MAD} = b\frac{\boldsymbol{x} - \tilde{x}}{median(|\boldsymbol{x} - \tilde{x}|)}, \tag{3.2}$$

where MAD stands for the Median Absolute Deviation and $\tilde{x}$ is the median. The constant 0.6745 comes from the fact that MAD is multiplied with the constant 1.483 which is a correction factor that makes the MAD unbiased at the normal distribution (b

= 1/1.483 = 0.6745) [52]. The modified z-score yields more robust results due to the scale and location factors which are introduced by MAD in (3.2).

3.3.3.1.2. Interquartile range (IQR)

The Interquartile Range (IQR) is a widely used approach which measures the statistical dispersion using the 1st and 3rd quartiles of an attribute's range [52], [294]. It is defined as the difference between the upper (Q3) and lower (Q1) quartiles of the data. Q1 is defined as the 25th percentile (lower quartile) whereas Q3 is the 75th percentile (upper quartile). Values lower than the first quartile or larger than the third quartile are outliers [52], [294]. The IQR multiplied by 0.7413 yields the normalized IQR. The term 0.7413 comes from the inverse of the width of the standard normal distribution (1/1.3498).

3.3.3.1.3. Grubb's test

The Grubb's statistical test is a univariate statistical measure which tests for the hypothesis that there are outliers in the data [52], [295]. The test statistics is given as:

$$G = \frac{max\ (|\boldsymbol{x} - \hat{x}|)}{\sigma_x}. \tag{3.3}$$

In fact, the Grubb's test statistics is defined as the largest absolute deviation from the sample mean in units of the sample standard deviation. Here, we are interested in testing whether the minimum value or the maximum value of $\boldsymbol{x}$ is an outlier, i.e., a two-sided test. A value is an outlier if the null hypothesis is rejected at the .05 significance level. Another test statistics is the Hampel's test which is defined as the difference of each sample from its population median value (median deviation). A sample is an outlier if its absolute Hampel value is 4.5 times larger than (or equal to) the median deviation [52], [66], [296], [297].

3.3.3.2. Machine learning approaches

3.3.3.2.1. Local outlier factor (LOF)

The Local Outlier Factor (LOF) [65], [298], [299] is a density-based approach which measures the local density of a given data point with respect to its neighboring points, where the number of nearest neighbors determines the accuracy of the model. The LOF uses the density of a point against its neighbors to determine the degree of the whether

the point is an outlier. For a point x, the local reachability density (lrd) of x, $lrd(x)$, is defined as [65], [298], [299]:

$$rd(x) = \frac{\|N_k(x)\|}{\sum_{x' \in N_k(x)} r(x, x')}, \tag{3.4}$$

where $N_k(x)$ is the set of k-nearest neighbors for $\boldsymbol{x}$, $r(x, x')$ is the reachability distance which is defined as the distance between x and its k-nearest neighbor. The reachability distance is the true distance between two points. The LOF is given by [65], [298], [299]:

$$\begin{aligned} LOF(x) &= \frac{\sum_{x' \in N_k(x)} (lrd(x')/lrd(x))}{\|N_k(x)\|} \\ &= \sum_{x' \in N_k(x)} lrd(x') \sum_{x' \in N_k(x)} r(x, x'), \end{aligned} \tag{3.5}$$

which is equal to the average local reachability density of the neighbors divided by the point's own local reachability density. The lower the local reachability density of x the higher the local reachability density of the kNN of x and thus the higher the LOF. The higher the LOF the more likely the point is an outlier.

3.3.3.2.2. Isolation forests

Isolation forests [52], [298], [300], [301] is a collection of isolation trees which: (i) enable the exploitation of subsampling data to precisely detect outliers, (ii) does not make use of distance or density measures to detect anomalies, (iii) achieves linear time complexity, and (iv) is scalable. The term "isolation" stands for the separation of an instance (a polluting one) from the rest of the instances (the inliers). Isolation trees are binary trees where instances are recursively partitioned and produce noticeable shorter paths for anomalies since: (i) in the regions occupied by anomalies, less anomalies result in a smaller number of partitions – shorter paths in a tree structure, and (ii) instances with distinguishable attribute-values are more likely to be separated early in the partitioning process [52], [298], [300], [301]. Thus, when a forest of random trees collectively produces shorter path lengths for some particular points, they are highly likely to be anomalies [52], [298], [300], [301].

The subsample size controls the training data size and affects the reliability of outlier detection whereas the number of trees controls the size of the ensemble trees [52],

[298], [300], [301]. In practice, M is set to 2^8 and N is set to 100. The anomaly score is finally defined as, as follows:

$$s(x, M) = 2^{-\frac{E(h(x))}{c(M)}}, \tag{3.6}$$

where $c(M)$ the average path length of unsuccessful searches, $h(x)$ is a harmonic number which is defined as $ln(x)$ plus the Euler's constant and $E(h(x))$ is the average of $h(x)$ from a collection of isolation forests [52], [298], [300], [301]. Scores close to 1 indicate anomalies, scores much smaller than 0.5 are inliers and scores close (or equal) to 0.5 are safe instances.

3.3.3.2.3. Gaussian elliptic envelopes

A common distance measure, which is widely used for anomaly detection in properly scaled datasets, is the Euclidean distance. In multivariate datasets however, the Euclidean distance suffers from the covariance that exists between the variables [52], [302]. A distance measure that accounts for such effects, in multivariate datasets, is the Mahalanobis distance which uses the eigenvalues to transform the original space into the eigenspace, so as to neglect the correlation among the variables of the dataset [303] and is defined as:

$$D(\boldsymbol{x}) = \sqrt{(\boldsymbol{x} - \boldsymbol{\mu})\boldsymbol{\Sigma^{-1}}(\boldsymbol{x} - \boldsymbol{\mu})}, \tag{3.7}$$

where $\boldsymbol{x}$ is an n-dimensional feature vector where the observations are stacked in columns, $\boldsymbol{\mu}$ is the mean vector across the observations, and $\boldsymbol{\Sigma^{-1}}$ is the inverse covariance matrix. Note that if the covariance matrix is the identity matrix, (3.7) yields the Euclidean distance, whereas if the covariance matrix is diagonal, (3.7) yields the normalized Euclidean distance. A way to visualize the result of (3.7), is to use an Elliptic envelope. Data within the ellipse surface are inliers, whereas data outside of the ellipse are outliers. The Elliptic envelope (also referred to as elliptical envelope), models the data as high-dimensional Gaussian distributions that consider for the covariance between the observations. The FAST-Minimum Covariance Determinant [304] is widely used to estimate the size and the shape of the ellipsis. The algorithm conducts initial estimations of the mean vector μ and the covariance matrix Σ (3.7) using non-overlapping subsamples of the feature vector. Then the algorithm proceeds with new subsamples until the determinant of the covariance matrix converges.

3.3.4. De-duplication

3.3.4.1. Similarity detection

The data curation service offers additional capabilities for detecting highly correlated and duplicated features. The highly correlated features are identified by first computing the Spearman's correlation coefficient between each possible pair of features. The Spearman correlation coefficient is defined as follows:

$$rho = 1 - \frac{6 \cdot \sum {d^2}_i}{N \cdot (N^2 - 1)} \ , \tag{3.8}$$

where d_i is the difference between the ranks of the values between two features and N is the number of samples per feature. An example of the correlation matrix is depicted in Figure 6.

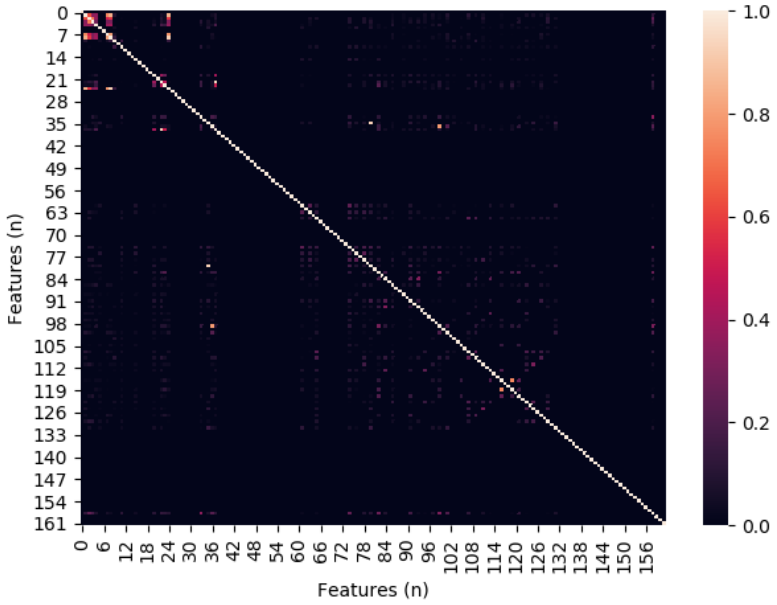


Figure 6. An illustration of the square Spearman correlation matrix for detecting highly correlated pairs of features across the raw input data (those depicted with high intensities).

3.3.4.2. Context based de-duplication

In addition, string similarity metrics, such as, the Jaro distance [52], [305], are also used to detect features with lexical similarities by computing the Jaro distance between each possible pair of features labels. For two strings, x and y, the Jaro string similarity measure, sim_J, is equal to:

$$sim_J = \begin{cases} 0 & , \quad c = 0 \\ \frac{1}{3} \cdot \left(\frac{c}{|x|} + \frac{c}{|y|} + \frac{c-t}{c} \right), & o/w \end{cases} , \tag{3.9}$$

where c is the number of matching (coincident) characters, and t is half the number of transpositions. Finally, the pairs of features having more than 95% correlation are highlighted for further evaluation by the data providers including the features that express lexical similarities (e.g., the ‘’Lymphadenopathy” and the “Lymphadenopathy (fixed)”). An example of the lexical matrix is depicted in Figure 7.

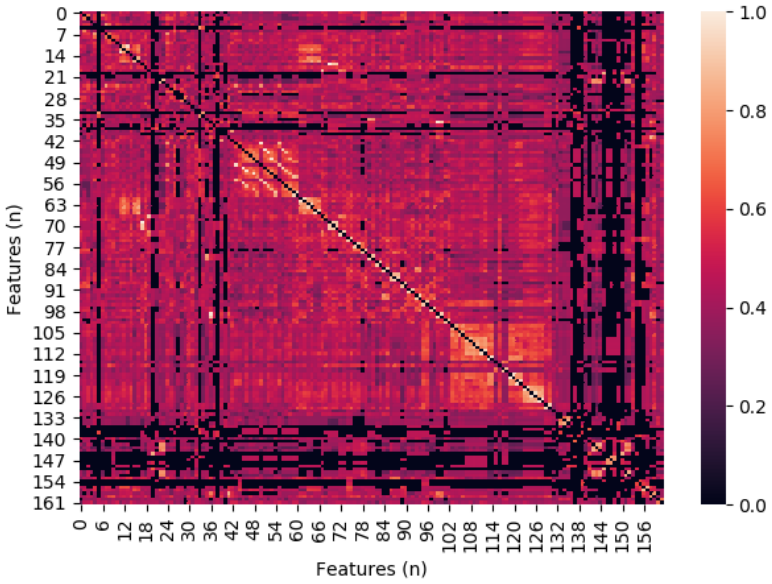


Figure 7. An illustration of the lexical distance matrix for detecting lexically similar (or identical) terms across the raw input data (those depicted with high intensities).

3.3.5. Data imputation

The features are classified according to the number of missing values into three categories, namely the: (i) “bad” features, (ii) “good” features, and (iii) “fair” features. The “bad” features are those having more than 50% of missing values. The “fair” features are those having less than 50% missing values where data imputation can be applied to improve the quality of the dataset. Finally, the “good” features are those without any missing values. The data imputation process is semi-automated and is performed with the support of the data providers. It consists of two options for replacing missing values, namely the average/most frequent (or median) and the mean value for the discrete and continuous features, respectively.

The “bad” features are excluded from the data imputation process since any attempt to replace the missing values would be pointless. In fact, the data imputation process can be performed only for the features that are classified as “fair”, with integer, float, or date data types, and only when the data provider agrees to do so. The “fair” features with unknown data types are excluded from the data imputation process. Features with detected outliers and/or unknown data types are also excluded from the data imputation process since any attempt to impute them would further spread the contaminated values within the dataset.

3.3.5.1. Average/Most frequent

The average/most frequent method is a univariate method according to which the missing values in the continuous features are replaced with the average value of the remaining (non-missing) values in each individual feature whereas the missing values in the discrete features are replaced with the most frequent value of the non-missing values [52], [66]. This approach however might introduce critical biases which often lead to data poisoning and make data useless. For instance, assume that we have two discrete features, say “gender” with values {“male”, “female”, “other”} and “pregnancy_status” with values {“yes”, “no”}. Our goal is to impute the missing values in the discrete feature “pregnancy_status”. For the sake of easiness, let’s assume that the most frequent value in the feature “pregnancy_status” is “yes”. Thus, all missing values in “pregnancy_status” will be set to “yes”. In that case, however, a patient who is “male” will have a “pregnancy_status” set to “yes” which is invalid; thus introducing significant biases in the dataset.

3.3.5.2. Random imputation

As its name implies, random imputation is a univariate approach which draws a random value from the distribution of each individual features to replace the missing ones [52], [66]. Random imputation can also lead to data poisoning since the randomness is highly likely to yield values with no practical reasoning (like in the previous example).

3.3.5.3. k-nearest neighbors (kNN)

The k-nearest neighbors (kNN) method is a multivariate and more straightforward approach where the samples with missing values in a patient are imputed according to

the values of the k-nearest neighboring samples [306]. The imputed values are equal as the weighted average of the neighboring samples, where the weights of all neighboring samples can be either: (i) uniform and thus have equal influence on the missing value of the sample, or (ii) equal to the inverse of the Euclidean distance (or any type of distance measure, e.g., the Mahalanobis distance) of the neighboring samples, where neighboring samples with smaller distance have a greater influence on the missing value of the sample.

As a result, the number of neighbors is proportionate to the computational complexity of the imputation process, where higher number of neighbors (k-valeus) lead to higher computational complexity. In addition, the k-NN approach is mostly useful in time-series data rather than in clinical data where each feature represents a unique information. In that case, large number of neighbors can lead to the definition of incompatible values for each individual feature and thus to data poisoning.

3.3.5.4. "Smart" imputation

3.3.5.4.1. Workflow

The proposed "smart" imputation workflow (Figure 8) consists of four stages, namely the: (i) data pre-processing stage, (ii) generation of virtual patient profiles stage, (iii) "smart" imputation stage, and (iv) nested validation stage. A data pre-processing pipeline is first applied on the raw data to resolve data inconsistencies, such as, outliers and duplicated fields. Four state-of-the art data generators (tree ensembles, Bayesian Networks, Bayesian Gaussian Mixture Models, Artificial Neural Networks) are trained on the curated data to produce virtual distributions with low dispersity.

A search algorithm is then applied to seek for a set of virtual patients having common clinical profiles with the real patients though the definition of a profile matching score (PMS) which quantifies the distance between the non-missing values of a real patient profile against those from the pool of virtual (synthetic) patient profiles. The virtual patient profiles with the smallest PMS are then used for imputation. Nested validation is applied to assess the accuracy of the imputed values by computing the average correlation difference (CD) and scaled squared absolute difference (SSAD) among the original and the imputed data. The validation process is repeated ten times for five different contamination ratios (i.e., 10% to 50%).

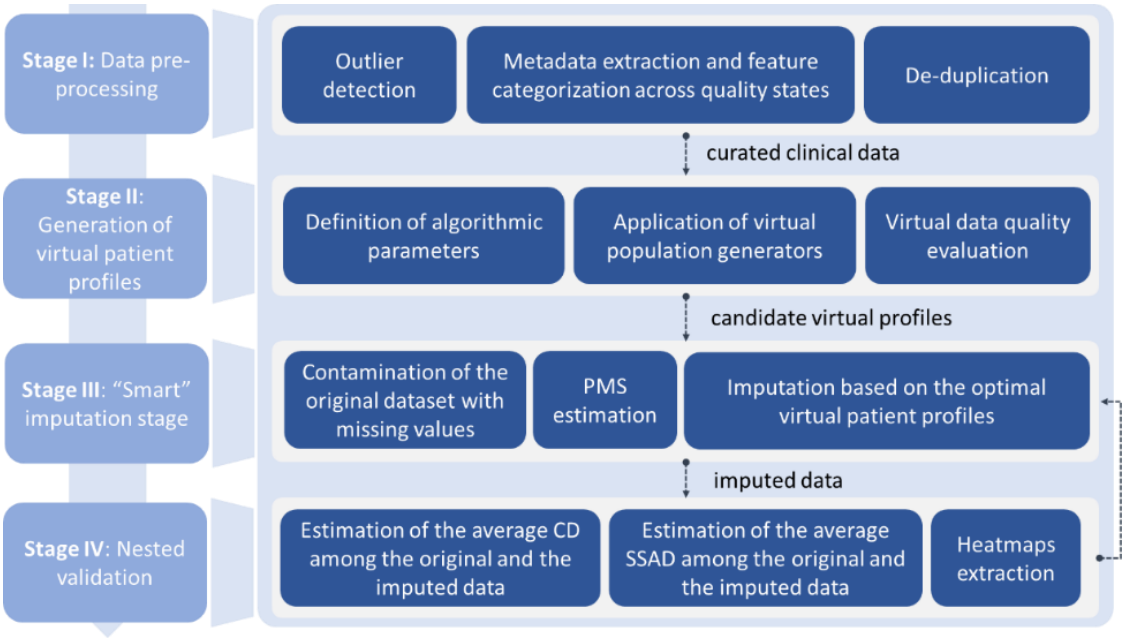


Figure 8. An illustration of the proposed workflow.

3.3.5.4.2. Smart imputation

In this work, the "smart" imputation process is based on the identification of a set of N-candidate virtual patient profiles, say $\{V_1, V_2, \dots, V_N\}$, which will be used to impute the missing values in a set of N-real patient profiles $\{R_1, R_2, \dots, R_N\}$. To do so, a search algorithm was developed to identify the candidate virtual patient profiles as those having the smallest profile matching score (PMS) which is defined as in:

$$PMS_v = e^{-|x_{v,1}-x_{r,1}|^2} + e^{-|x_{v,2}-x_{r,2}|^2} + \cdots + e^{-|x_{v,n}-x_{r,n}|^2}, \tag{3.10}$$

where v is the index of the virtual data generator, $x_{v,i}$ is the i-th virtual feature and $x_{r,i}$ is the corresponding real feature, $i \in [0, n]$ and n is the number of features in the dataset. Since the Euclidean and similar linear distance measures are not able to model the distance between diverse features in the cartesian grid, the exponential function was used to provide a non-linear relationship between them. A high-level pseudocode of the "smart" imputation workflow is presented in Algorithm 1. The input parameters include the curated clinical dataset, the number of features, the virtual population algorithmic specifications and the empty virtual population algorithmic objects. Each virtual data generator is trained on the curated clinical dataset based on the provided specifications. Then, the profile matching score is calculated based on the virtual data. The smallest

profile matching score is then extracted, and the suggested imputed values are returned in the output.

Algorithm 1. A pseudocode of the "smart" imputation process.

1	**Input parameters**
2	D_cur = curated clinical dataset
3	N = number of features
4	specs = virtual population algorithmic specifications
5	vpops = empty virtual population algorithmic objects
6	def **smart_imputation**(D_cur):
7	for vpop in vpops:
8	Vpop_m = vpop(D_cur, specs[vpop])
9	pms = PMS(Vpop_m)
10	S = search(pms)
11	imputed_S = extrac_set(S)
12	return imputed_S

3.3.5.4.3. Synthetic data generation

Gaussian Mixture Models (GMM) with variational Bayesian inference (BGMM) focus on the estimation of a set of hyper-parameter(s) $\boldsymbol{\theta}$ in a search function, say $q(\boldsymbol{x};\boldsymbol{\theta})$, so that the Kullback-Leibler (KL) divergence with the posterior distribution, say $p(\boldsymbol{x})$, is minimized [130]. In this case, the minimization of the logarithm of the evidence yields:

$$argmax_{\theta}\left[\int_{\boldsymbol{x}} q(\boldsymbol{x};\boldsymbol{\theta})\big(R(\boldsymbol{x}) + log\big(\tilde{q}(i|\boldsymbol{x})\big)\big)d\boldsymbol{x} + H(q)\right], \tag{3.11}$$

where $R(\boldsymbol{x})$ refers to the logarithm of the posterior distribution, $H(q)$ is the entropy of $q(\boldsymbol{x};\boldsymbol{\theta})$, and $\tilde{q}(i|\boldsymbol{x})$ is a multivariate normal distribution. Since the number of Gaussian components in the BGMM is arbitrary, we applied k-means clustering to derive robust estimations on the number of Gaussian components, where the Davies-Bouldin index was used to estimate the optimal number of clusters, say K. Then, the number of Gaussian components was set equal to K to initialize the BGMM. The Bayesian networks are modeled as a directed acyclic graph (DAG), where each node $v \in \boldsymbol{V}$ is assigned to a random variable, say x_v, according to [168]:

$$p_v = \prod_{c\in\boldsymbol{C}} p\big(x_c|x_{pa(c)}\big)\prod_{d\in\boldsymbol{D}} p\big(x_d|x_{pa(c)},x_{pa(d)}\big). \tag{3.12}$$

where, $p(x_d|x_{pa(c)}, x_{pa(d)})$ is the probability of x_c given the parents of the discrete and continuous variables, $x_{pa(c)}$, $x_{pa(d)}$, and $p(x_c|x_{pa(c)})$ is the probability of x_c given the parents of the continuous nodes, $x_{pa(c)}$. Tree ensembles were trained on a portion of the real data (e.g., 50%) as described in [131], [173], by estimating a set of density trees. Artificial neural networks (ANNs) were also utilized as described in [131], [172] using Gaussian radial basis functions (RBFs) as activation functions.

3.3.5.4.4. Synthetic data quality evaluation

Synthetic data quality metrics [130], [134] were used to quantify the: (i) similarity (i.e., the average correlation difference; small value indicate reduced dissimilarity), (ii) level of dispersity (i.e., the variance to mean ratio – VMR; values less than 1 denote reduced dispersity) which is also referred to as coefficient of variation, and (iii) convergence (i.e., the Kullback Leibler-divergence or KL-divergence; values close to 0 denote reduced entropy variation) among the real and the virtual patient profiles.

3.3.5.4.5. Validation

A nested validation method was developed to demonstrate the effectiveness of the proposed "smart" imputation process. According to the proposed method, each feature in the real patient dataset is randomly contaminated with different ratios. The process is repeated ten times to avoid biases. In each iteration, the scaled squared absolute difference (SSAD) between the original and the proposed imputed values is computed along with the average correlation difference (CD) between the features in the original and imputed datasets.

3.3.6. The data evaluation report

An instance of the data evaluation report is depicted in Figure 9. As it is already mentioned, the data evaluation report summarizes the contents of the "Info" panel and the "Quality assessment" panel, in a tabular format, thus providing an offline, concise view on the structure and vocabulary of the data. For the sample dataset, the total number of features was equal to 166 and the number of patients was equal to 250. In this example, out of 166 features, 60 were characterized as discrete and 78 as continuous. The number of unknown features was equal to 28 and the total number of missing values was equal to 44.58%. The names of the features are the labels that exist

in the first row of the file “raw_dataset” under the path “raw_data” of the data provider’s private cloud space. The value ranges in each feature include the minimum and the maximum values that exist on each feature’s space. In the case where the feature has unknown or string data type, the complete set of unique values is presented in the value range. For example, the feature “First visit (year)” has a variable type of date in the range [1983, 2018]. On the other hand, the feature “comorbidities” has a variable type string and thus all the unique string values are recorded (e.g., “HEART ARRHYTHMIAS”, “HASHIMOTO”). In the same feature, the outlier detection method is not applicable since it has a string data type. The same occurs for the features “First Symptom”, and “Year of first symptom”. The outlier detection method is also not applicable for the bad features with unknown data types, such as, for: (i) “Rose-Bengal Stain(0-1)”, which includes an unknown symbol “+” that probably denotes positivity instead of the value “1”, (ii) “Positive ocular stain score”, which includes values that are recorded as fractions (e.g., “1/9”), and (iii) “Dry-mouth-Objective (ml of saliva in 15 min)”, which includes an incompatible value “<1.5”.

Metadata	
Number of feature(s)	166
Number of instance(s)	250
Discrete feature(s)	60
Continuous feature(s)	78
Unknown feature(s)	28
Missing values (%)	44.58%

Quality assessment

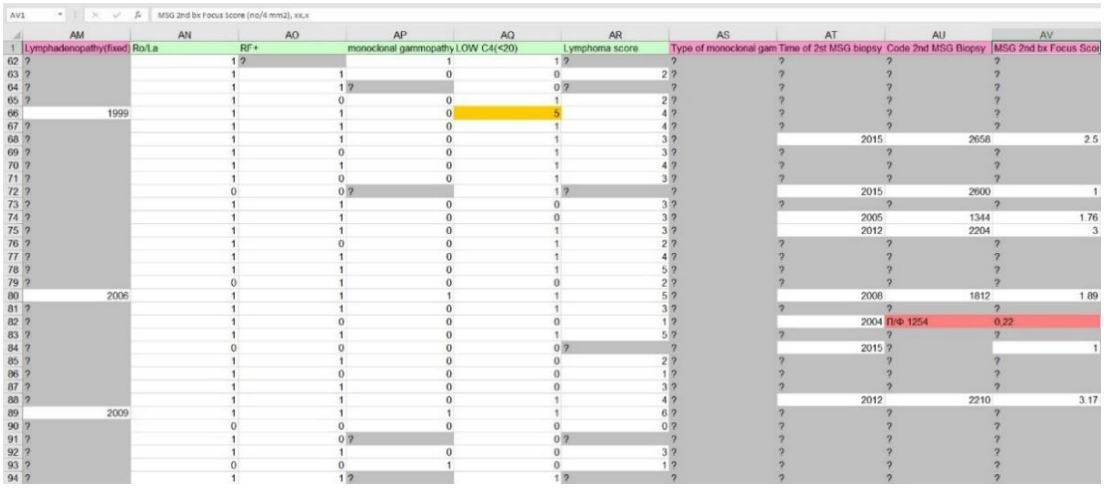
Features	Value range	Type	Variable type	Missing values	State	Outliers	Incompatibilities
comorbidities	ASHIMOTO, HEART ARRHYTHMIAS	categorical	string	0	good	not-applicable	no
First visit (year)	[1983, 2018]	numeric	date	0	good	no	no
Last visit (year)	[1991, 2018]	numeric	date	0	good	no	no
Old (1), new (2), old still in follow up (3)	[1, 2018]	numeric	int	0	good	yes	no
Year of Birth	[1918, 1995]	numeric	date	0	good	yes	no
SEX (female=1)	[0, 1]	categorical	int	2	fair	no	no
First Symptom	arthralgias, dry mouth, dry eyes, dry mou	categorical	string	1	fair	not-applicable	no
Year of first symptom	1992, 1993, 1994, 1995, 1996, 1997,	unknown	unknown	3	fair	not-applicable	yes, unknown type of data
Year of disease diagnosis	[1982, 2018]	numeric	date	1	fair	no	no
Age at SS diagnosis	[14, 81]	numeric	int	1	fair	no	no
Date of blood drawn	[2016, 2017]	numeric	date	247	bad	no	yes, bad feature
Dry mouth-subjective	[0, 1]	categorical	int	4	fair	no	no
Dry mouth, subjctiv Date	[1975, 2017]	numeric	date	22	fair	no	no
Dry mouth-Objective (ml of saliva in 15min)	1.0, 1.2, 1.5, 2.5, 0.15, 0.2, 0.75, 1.6, 1	unknown	unknown	214	bad	not-applicable	yes, unknown type of data
Whole salivary flow Date	[1985, 2018]	numeric	date	217	bad	no	yes, bad feature
Dry eyes subj	[0, 1]	categorical	int	1	fair	no	no
Dry eyes, subjctiv Date	[1970, 2017]	numeric	date	20	fair	yes	no
Rose-Bengal Stain(0-1)	[0.0, 1.0, +]	unknown	unknown	138	bad	not-applicable	yes, unknown type of data
Positive ocular stain score	[1/9, 5/9, 6/9, 9/9]	categorical	string	243	bad	not-applicable	yes, bad feature
Abnormal Shirmer's	[0, 1]	categorical	int	50	fair	no	no

Figure 9. An indicative instance of the data evaluation report.

3.3.7. The curated dataset

An instance of the automatically generated curated dataset, for the same dataset, is depicted in Figure 10. The fact that the features “Ro/La”, “RF+”, “monocloncal gammopathy”, “LOW C4(<20)”, and “Lymphoma score” are filled with light green color denotes that they have less than 50% missing values, whereas the features “Lymphadenopathy (fixed) date(-yr)”, “Type of monoclonal gammopathy”, “Time of

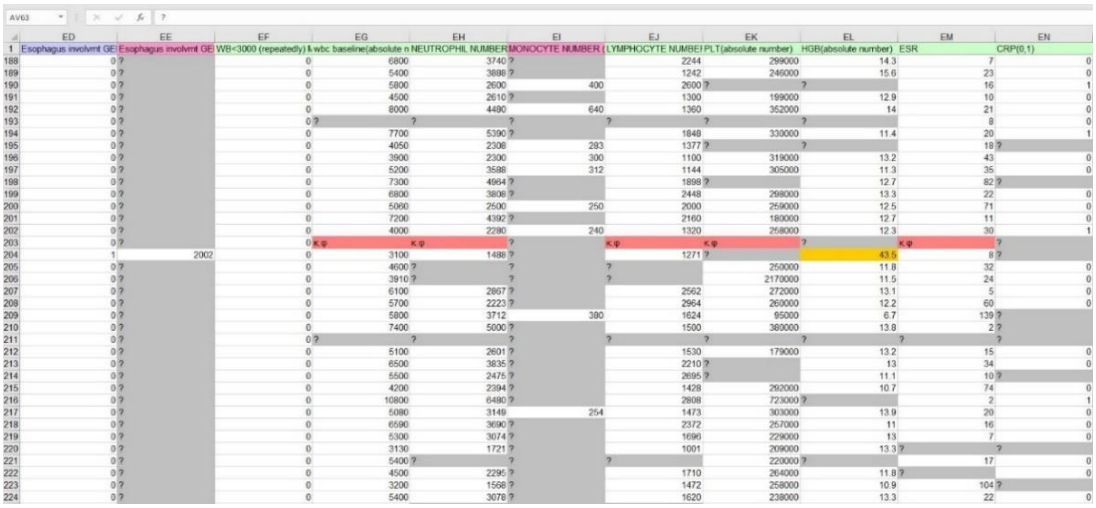
2st MSG biopsy", "Code 2nd MSG Biopsy", and "MSG 2nd bx Focus Score (no/4 mm2), xx, x" have more than 50% missing values and are depicted with light red color. The missing values are recorded with the symbol "?", in a gray background for easier tracking. According to Figure 10, a contaminated value 5 was successfully detected for the feature "LOW C4(<20)" and marked with yellow color since this feature is expected to have binary values (i.e., "1" for C4 values lower than 20 or "0" otherwise). In addition, two incompatible values, namely "Π/Φ 1254" and "0,22" are successfully marked with red color for the features "Code 2nd MSG biopsy", and "MSG 2nd bx Focus Score (no4/mm2), xx, x", respectively. The former value contains a combination of unknown characters (i.e., "Π/Φ") with numbers, whereas the latter value has incompatible format (i.e., "0,22" instead of "0.22").



	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV
1	Lymphadenopathy(fixed)	Ro/La	RF+	monoclonal gammopathy	LOW C4(<20)	Lymphoma score	Type of monoclonal gam	Time of 2st MSG biopsy	Code 2nd MSG Biopsy	MSG 2nd bx Focus Scor
62	?		1	?	1	1	?	?	?	?
63	?		1	1	0	0	2	?	?	?
64	?		1	1	?	0	?	?	?	?
65	?		1	0	0	1	2	?	?	?
66	1999		1	1	0	5	4	?	?	?
67	?		1	1	0	1	4	?	?	?
68	?		1	1	0	1	3	2015	2658	2.5
69	?		1	1	0	1	3	?	?	?
70	?		1	1	0	1	4	?	?	?
71	?		1	0	0	1	3	?	?	?
72	?		0	0	?	1	?	2015	2600	1
73	?		1	1	0	0	3	?	?	?
74	?		1	1	0	0	3	2005	1344	1.76
75	?		1	1	0	1	3	2012	2204	3
76	?		1	0	0	1	2	?	?	?
77	?		1	1	0	1	4	?	?	?
78	?		1	1	0	1	5	?	?	?
79	?		0	1	0	0	2	?	?	?
80	2006		1	1	1	1	5	2008	1812	1.89
81	?		1	1	0	1	3	?	?	?
82	?		1	0	0	0	1	2004	ΠΦ 1254	0,22
83	?		1	1	0	1	5	?	?	?
84	?		0	0	0	0	?	2015	?	1
85	?		1	1	0	0	2	?	?	?
86	?		1	0	0	0	1	?	?	?
87	?		1	1	0	0	3	?	?	?
88	?		1	1	0	1	4	2012	2210	3.17
89	2009		1	1	1	1	6	?	?	?
90	?		0	0	0	0	0	?	?	?
91	?		1	0	?	0	?	?	?	?
92	?		1	1	0	0	3	?	?	?
93	?		0	0	1	0	1	?	?	?
94	?		1	1	?	1	?	?	?	?

Figure 10. An indicative instance of the curated dataset with an incompatible value, an unknown value, and an outlier.

Another instance of the same curated dataset is depicted in Figure 11, where the same incompatible value (i.e., "κ.φ") has been detected by the service for the features "wbc baseline (absolute number)", "NEUTROPHIL NUMBER (absolute number)", "LYMPHOCYTE NUMBER (absolute number)", "PLT (absolute number)" which is the number of platelets, and "ESR" which stands for the erythrocyte sedimentation rate. An abnormal value 43.5 was detected by the service as an outlier for the feature "HGB (absolute number)" which stands for hemoglobin. In this instance, a good feature is also depicted with light blue color, namely the "Esophagus involvmt GER (0-1)". The rest of the features are "fair" and thus depicted with light green color except from the "bad" features "Esophagus involvmt" and "MONOCYTE NUMBER (absolute number)".

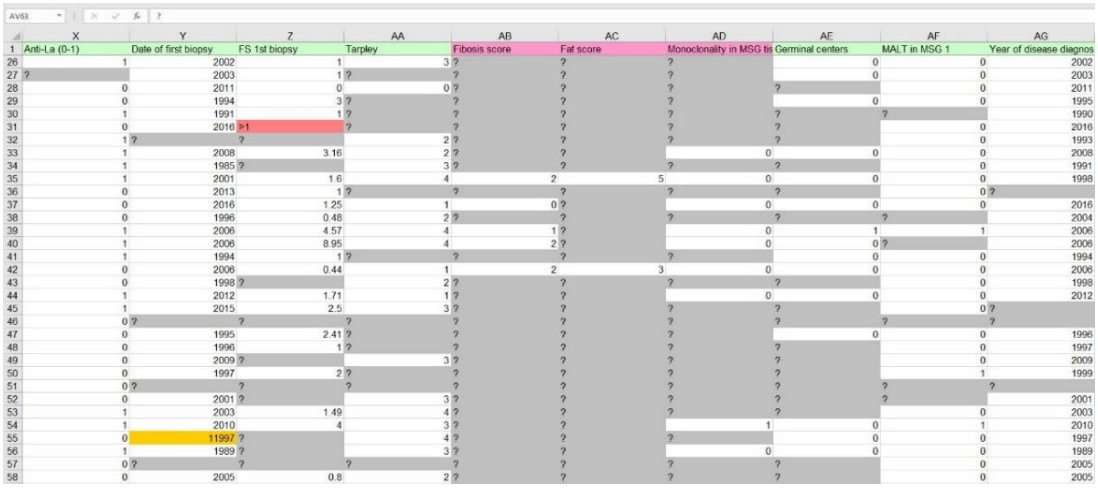


	ED	EE	EF	EG	EH	EI	EJ	EK	EL	EM	EN
1	Esophagus involvmt GE	Esophagus involvmt GE	WB<3000 (repeatedly)	k wbc baseline(absolute n	NEUTROPHIL NUMBER	MONOCYTE NUMBER (	LYMPHOCYTE NUMBE	PLT(absolute number)	HGB(absolute number)	ESR	CRP(0,1)
188	0 ?		0	6800	3740 ?		2244	299000	14.3	7	0
189	0 ?		0	5400	3898 ?		1242	246000	15.6	23	0
190	0 ?		0	5800	2600	400	2600 ?	?		16	1
191	0 ?		0	4500	2610 ?		1300	199000	12.9	10	0
192	0 ?		0	8000	4480	640	1360	352000	14	21	0
193	0 ?		0 ?	?	?	?	?	?	?	8	0
194	0 ?		0	7700	5390 ?		1848	330000	11.4	20	1
195	0 ?		0	4050	2306	283	1377 ?	?		18 ?	
196	0 ?		0	3900	2300	300	1100	319000	13.2	43	0
197	0 ?		0	5200	3588	312	1144	305000	11.3	35	0
198	0 ?		0	7300	4984 ?		1898 ?		12.7	82 ?	
199	0 ?		0	6800	3808 ?		2448	298000	13.3	22	0
200	0 ?		0	5060	2500	250	2000	259000	12.5	71	0
201	0 ?		0	7200	4392 ?		2160	180000	12.7	11	0
202	0 ?		0	4000	2280	240	1320	258000	12.3	30	1
203	0 ?		0 κ.φ	κ.φ	?		κ.φ	κ.φ	?	κ.φ	?
204	1	2002	0	3100	1488 ?		1271 ?		43.5	8 ?	
205	0 ?		0	4600 ?	?		?	250000	11.8	32	0
206	0 ?		0	3910 ?	?		?	2170000	11.5	24	0
207	0 ?		0	6100	2867 ?		2562	272000	13.1	5	0
208	0 ?		0	5700	2223 ?		2964	260000	12.2	60	0
209	0 ?		0	5800	3712	380	1624	95000	6.7	139 ?	
210	0 ?		0	7400	5000 ?		1500	380000	13.8	2 ?	
211	0 ?		0 ?	?	?		?	?	?	?	?
212	0 ?		0	5100	2601 ?		1530	179000	13.2	15	0
213	0 ?		0	6500	3835 ?		2210 ?		13	34	0
214	0 ?		0	5500	2475 ?		2695 ?		11.1	10 ?	
215	0 ?		0	4200	2394 ?		1428	292000	10.7	74	0
216	0 ?		0	10800	6480 ?		2808	723000 ?		2	1
217	0 ?		0	5080	3149	254	1473	303000	13.9	20	0
218	0 ?		0	6590	3690 ?		2372	257000	11	16	0
219	0 ?		0	5300	3074 ?		1696	229000	13	7	0
220	0 ?		0	3130	1721 ?		1001	209000	13.3 ?	?	
221	0 ?		0	5400 ?	?		?	220000 ?		17	0
222	0 ?		0	4500	2295 ?		1710	264000	11.8 ?		0
223	0 ?		0	3200	1568 ?		1472	258000	10.9	104 ?	
224	0 ?		0	5400	3078 ?		1620	238000	13.3	22	0

Figure 11. An instance of the curated dataset with four unknown values and one outlier.

A final instance of the same curated dataset is depicted in Figure 12, where an outlier value 11997 has been detected by the service for the feature “Date of first biopsy” along with an incompatible value “>1” for the feature “FS 1st biopsy”.

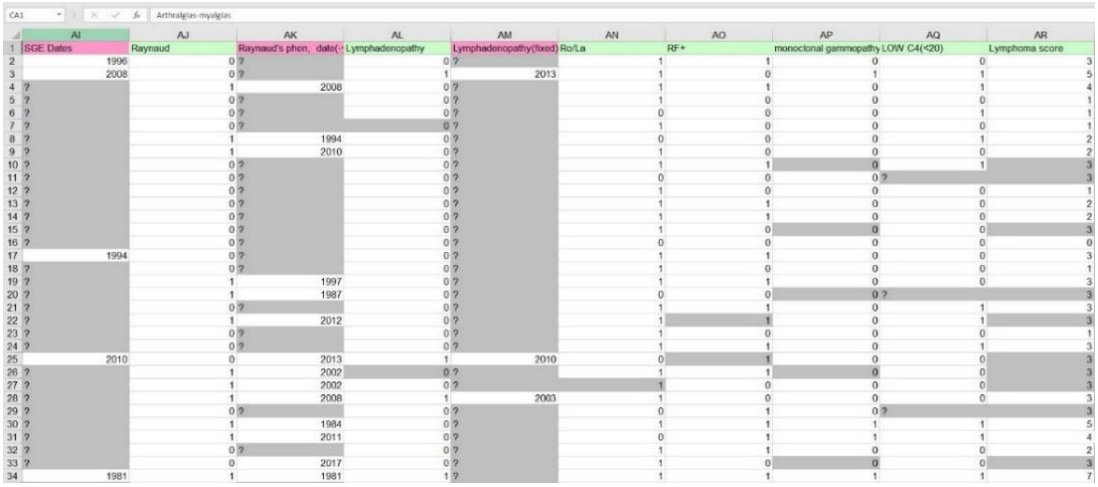
The former value implicates an erroneously parsed year whereas the former value denotes a value which might be larger than 1 but it is not properly recorded.



	X	Y	Z	AA	AB	AC	AD	AE	AF	AG
1	Anti-La (0-1)	Date of first biopsy	FS 1st biopsy	Tarpley	Fibosis score	Fat score	Monoclonality in MSG tis	Germinal centers	MALT in MSG 1	Year of disease diagnos
26	1	2002	1	3 ?	?	?	?	0	0	2002
27	?	2003	1 ?	?	?	?	?	0	0	2003
28	0	2011	0	0 ?	?	?	?	?	0	2011
29	0	1994	3 ?	?	?	?	?	0	0	1995
30	1	1991	1 ?	?	?	?	?	?	?	1990
31	0	2016	>1	?	?	?	?	?	0	2016
32	1 ?	?		2 ?	?	?	?	?	0	1993
33	1	2008	3.16	2 ?	?	?	0	0	0	2008
34	1	1985 ?		3 ?	?	?	?	?	0	1991
35	1	2001	1.6	4	2	5	0	0	0	1998
36	0	2013	1 ?	?	?	?	?	?	0 ?	
37	0	2016	1.25	1	0 ?	?	0	0	0	2016
38	0	1996	0.48	2 ?	?	?	?	?	?	2004
39	1	2006	4.57	4	1 ?	?	0	1	1	2006
40	1	2006	8.95	4	2 ?	?	0	0 ?		2006
41	1	1994	1 ?	?	?	?	?	0	0	1994
42	0	2006	0.44	1	2	3	0	0	0	2006
43	0	1998 ?		2 ?	?	?	?	?	0	1998
44	1	2012	1.71	1 ?	?	?	0	0	0	2012
45	1	2015	2.5	3 ?	?	?	?	?	0 ?	
46	0 ?	?	?	?	?	?	?	?	?	
47	0	1995	2.41 ?	?	?	?	?	0	0	1996
48	0	1996	1 ?	?	?	?	?	?	0	1997
49	0	2009 ?		3 ?	?	?	?	?	0	2009
50	0	1997	2 ?	?	?	?	?	?	1	1999
51	0 ?	?	?	?	?	?	?	?	?	?
52	0	2001 ?		3 ?	?	?	?	?	?	2001
53	1	2003	1.49	4 ?	?	?	?	?	0	2003
54	1	2010	4	3 ?	?	?	1	0	1	2010
55	0	11997 ?		4 ?	?	?	?	0	0	1997
56	1	1989 ?		3 ?	?	?	0	0	0	1989
57	0 ?	?	?	?	?	?	?	?	0	2005
58	0	2005	0.8	2 ?	?	?	?	?	0	2005

Figure 12. An indicative instance of the curated dataset with erroneously parsed values; one incompatible value and one outlier.

To demonstrate the application of the data imputation process, a second experiment was conducted on the same dataset, where the data imputation method has been set to “Average/most frequent” instead of “None”.



	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR
1	SGE Dates	Raynaud	Raynaud's phen, date(-	Lymphadenopathy	Lymphadenopathy(fixed)	Ro/La	RF+	monoclonal gammopathy	LOW C4(<20)	Lymphoma score
2	1996	0	?	0	?	1	1	0	0	3
3	2008	0	?	1	2013	1	0	1	1	5
4	?	1	2008	0	?	1	1	0	1	4
5	?	0	?	0	?	1	0	0	0	1
6	?	0	?	0	?	0	0	0	1	1
7	?	0	?	0	?	1	0	0	0	1
8	?	1	1994	0	?	0	0	0	1	2
9	?	1	2010	0	?	1	0	0	0	2
10	?	0	?	0	?	1	1	0	1	3
11	?	0	?	0	?	0	0	0	?	3
12	?	0	?	0	?	1	0	0	0	1
13	?	0	?	0	?	1	1	0	0	2
14	?	0	?	0	?	1	1	0	0	2
15	?	0	?	0	?	1	0	0	0	3
16	?	0	?	0	?	0	0	0	0	0
17	1994	0	?	0	?	1	1	0	0	3
18	?	0	?	0	?	1	0	0	0	1
19	?	1	1997	0	?	1	1	0	0	3
20	?	1	1987	0	?	0	0	0	?	3
21	?	0	?	0	?	1	1	0	1	3
22	?	1	2012	0	?	1	1	0	1	3
23	?	0	?	0	?	1	0	0	0	1
24	?	0	?	0	?	1	1	0	1	3
25	2010	0	2013	1	2010	0	1	0	0	3
26	?	1	2002	0	?	1	1	0	0	3
27	?	1	2002	0	?	1	0	0	0	3
28	?	1	2008	1	2003	1	0	0	0	3
29	?	0	?	0	?	0	1	0	?	3
30	?	1	1984	0	?	1	1	1	1	5
31	?	1	2011	0	?	0	1	1	1	4
32	?	0	?	0	?	1	1	0	0	2
33	?	0	2017	0	?	1	0	0	0	3
34	1981	1	1981	1	?	1	1	1	1	7

Figure 13. An indicative instance of the curated dataset after data imputation is applied.

An instance of the curated dataset after data imputation is depicted in Figure 13, where the "bad" features "SGE Dates", "Raynaud's phen, date(-yr)", and "Lymphadenopathy (fixed) date(-yr)" have been correctly ignored from the imputation process whereas the missing values on the "fair" features "Raynaud", "Lymphadenopathy", "Ro/La", "RF+", "monoclonal gammopathy", and "Lymphoma score" have been replaced with the median value. Note that the feature "LOW C4(<20)" has not been replaced since it has outliers (Figure 10) and any attempt to replace the missing values would further contaminate that feature. In fact, the content of the "Quality Assessment" panel along with the content of the upper panel with metadata is summarized within the homonymous document (i.e., the data evaluation report).

An instance of the data evaluation report is depicted in Figure 14.

Metadata	
Number of feature(s)	78
Number of instance(s)	50
Discrete feature(s)	60
Continuous feature(s)	17
Unknown feature(s)	1
Missing values (%)	11.97%

Quality assessment

Features	Value range	Type	Variable type	Missing values	State	Outliers	Incompatibilities
SEX (female=1)	[0, 1]	categorical	int	0	good	no	no
Age at SS diagnosis	[22, 78]	numeric	int	0	good	no	no
Dry mouth-subjective	[0, 1]	categorical	int	3	fair	no	no
Dry eyes subj	[0, 1]	categorical	int	0	good	no	no
Rose-Bengal Stain(0-1)	[0, 1]	categorical	int	15	fair	no	no
Abnormal Shirmer's	[0, 1]	categorical	int	3	fair	no	no
ANA+	[0, 1]	categorical	int	1	fair	no	no
RF(<20=0, >20=1)IU/ml	[0, 1]	categorical	int	3	fair	no	no
Anti-Ro (0-1)	[0, 1]	categorical	int	1	fair	no	no
Anti-La (0-1)	[0, 1]	categorical	int	0	good	no	no
FS 1st biopsy	[0.2, 5.26]	numeric	float	13	fair	no	no
SGE	[0, 1]	categorical	int	3	fair	no	no
Raynaud	[0, 1]	categorical	int	1	fair	no	no
Lymphadenopathy	[0, 1]	categorical	int	0	good	no	no
Ro/La	[0, 1]	categorical	int	0	good	no	no
RF+	[0, 1]	categorical	int	3	fair	no	no
monoclonal gammopathy (blood)	[0, 1]	categorical	int	4	fair	no	no
LOW C4(<20)	[0, 1]	categorical	int	2	fair	no	no
Dyspareunia, subjctiv (0-1)	[0, 1]	categorical	int	0	good	no	no
Dry skin, subjctiv (0-1)	[0, 0]	categorical	int	0	good	no	no
Dry upper resp, subjctiv (0-1)	[0, 1]	categorical	int	0	good	no	no

Figure 14. An instance of the automatically generated data evaluation report.

The data evaluation report is automatically generated and is stored inside the private cloud space of a data provider (in the case where the data curation service is executed in a secure cloud computing environment, where each data provider has his/her own private cloud space where their data are stored – the cloud computing environment must fulfill all the necessary GDPR regulations for data sharing and data protection). The same occurs for the curated dataset in the context of the data diagnostics procedure. Thus, the data provider has easy access to them through the private cloud space. Another instance of the data evaluation report is depicted in Figure 15 whereas an instance of the curated dataset is depicted in Figure 17. The curated dataset is automatically generated and is stored inside the private cloud space of the data provider.

According to Figure 17, the features "WB<3000 (repeatedly)" (White Blood Cell - repeatedly), "wbc baseline" (White Blood Cell at baseline) are "good", the features "Neutrophil number", "Lymphocyte number", "PLT (absolute number)" (Number of platelets in the blood), "HGB (absolute number)" (Hemoglobin), "ESR" (Eryhtrocyte Sedimentation Rate), "CRP (0,1)" (C-reactive protein), "γ-globulins" are "fair", and the features "Monocyte number" is "bad".

Arthritis (0-1)	[0, 1]	categorical	int	0	good	no	no
Raynaud's phen.(0-1)	[0, 1]	categorical	int	0	good	no	no
Palpable purpura (0-1)	[0, 1]	categorical	int	0	good	no	no
Vasculitic ulcer(0-1)	[0, 0]	categorical	int	0	good	no	no
Other rash(0-1)	[0, 1]	categorical	int	0	good	no	no
Myositis (0-1)	[0, 0]	categorical	int	0	good	no	no
PNS-entrapment (0-1)	[0, 1]	categorical	int	0	good	no	no
PNS-vasculitic (0-1)	[0, 0]	categorical	int	0	good	no	no
CNS involvmt (0-1)	[0, 0]	categorical	int	0	good	no	no
Psychiatric(0-1)	[0, 0]	categorical	int	0	good	no	no
Lymphadenopathy (0-1) (fixed)	[0, 1]	categorical	int	0	good	no	no
Splenomegaly(0-1)	[0, 0]	categorical	int	0	good	no	no
Liver involvmt (autoimmune cholangitis)/ biopsy	[0, 1]	categorical	int	0	good	no	no
Lung involvmt – interstitial disease Type(0-1)	[0, 0]	categorical	int	0	good	no	no
Lung involvmt – bronchocentric disease (0-1)	[0, 1]	categorical	int	0	good	no	no
Lung involvmt – pleurisy (0-1)	[0, 0]	categorical	int	0	good	no	no
Kidney involvmt -RTA – nephrocalcinosis (0-1)	[0, 0]	categorical	int	0	good	no	no
URINE SPECIFIC GRAVITY AT DIAGNOSIS	[1002, 1030]	numeric	int	6	fair	no	no
interstitial renal disease(0-1)	[0, 1]	categorical	int	2	fair	no	no
URINE PH AT FIRST VISIT	[4.5, 8]	numeric	float	13	fair	no	no
URINE PH AT LAST VISIT	[5, 8]	categorical	float	17	fair	no	no
Kidney involvmt -biopsy –interstitial infiltrates (0-	[0, 1]	categorical	int	0	good	no	no
Kidney involvmt -GN-biopsy(0-1)	[0, 1]	categorical	int	0	good	no	no
Heart involvmt -valvular Type	[0, 0]	categorical	int	0	good	no	no
Heart involvmt -pericardial (0-1)	[0, 0]	categorical	int	0	good	no	no
Heart involvmt -myocardial /CMR (0-1)	[0, 0]	categorical	int	0	good	no	no
Esophagus involvmt GER (0-1)	[0, 1]	categorical	int	0	good	no	no
WB<3000 (repeatedly) MORE THAN THREE (0	[0, 1]	categorical	int	0	good	no	no
wbc baseline(absolute number)	[3130, 10400]	numeric	int	0	good	yes	no
NEUTROPHIL NUMBER(absolute number)	[1551, 8008]	numeric	int	5	fair	yes	no
MONOCYTE NUMBER (absolute number)	[55, 742]	numeric	int	25	bad	no	yes, bad feature

Figure 15. A second instance of the automatically generated data evaluation report.

A final instance of the data quality evaluation report is depicted in Figure 16. This figure includes 3 fields were marked as outliers, and 16 as missing values followed by a question mark in the case where the features are "bad" and/or have incompatibilities and with the imputed value otherwise (e.g., for the features "Lymphocyte number", "CRP", "γ-globulins" the missing values have been imputed since they are "fair" features without incompatibilities).

PLT(absolute number)	[20000, 417000]	numeric	int	1	fair	no	no
HGB(absolute number)	[9, 14.6]	numeric	float	2	fair	yes	no
ESR	[6, 140]	numeric	int	2	fair	yes	no
CRP(0,1)	[0, 1]	categorical	int	6	fair	no	no
γ-globulins(11-18=0,>18=1,<11=2)	[0, 2]	categorical	int	6	fair	no	no
HBsAg(0-1)	[0, 0]	categorical	int	24	fair	no	no
Anti-HCV(0-1)	[0, 1]	categorical	int	24	fair	no	no
Anti-HIV I/II(0-1)	[0, 0]	categorical	int	30	bad	no	yes, bad feature
ANA(titer-1)	250, 1/1280, 1/160, 1/320, 1/640, 1/80,	unknown	unknown	2	fair	not-applicable	yes, unknown type of data
IgG	[795, 2720]	numeric	int	32	bad	no	yes, bad feature
IgM	[82, 797]	numeric	int	29	bad	yes	yes, bad feature
IgA	[140, 519]	numeric	int	29	bad	no	yes, bad feature
LDH	[124, 495]	numeric	int	16	fair	no	no
AMA(titer-1)	[0, 1]	categorical	int	26	bad	no	yes, bad feature
Anti-TPO(0,1)	[0, 1]	categorical	int	27	bad	no	yes, bad feature
Anti-TG (TITER)	[0, 1]	categorical	int	28	bad	no	yes, bad feature
C3(mg/mL)	[32, 214]	numeric	float	2	fair	yes	no
C4 (mg/mL)	[2.4, 65]	numeric	float	2	fair	yes	no
Cryo (mg/mL)	[0, 1]	categorical	int	15	fair	no	no
Lymphoma (0-1)	[0, 1]	categorical	int	0	good	no	no

Figure 16. A final instance of the automatically generated data evaluation report.

The values “10400”, “8008”, and “9” of the features “wbc baseline”, “Nutrophil number”, and “HGB (absolute number)”, respectively, are marked as outliers since they deviate from the standard distribution of each corresponding feature. The Imputation of the missing values takes place only for the “fair” features without any incompatibilities (e.g., outliers, inconsistent data types) to avoid any further data contamination in the curated dataset.

WB<3000 (repeatedly)	wbc baseline(absolute ni	NEUTROPHIL NUMBER(	MONOCYTE NUMBER (	LYMPHOCYTE NUMBEF	PLT(absolute number)	HGB(absolute number)	ESR	CRP(0,1)	γ-globulins(11-18=0,>18
0	4300	1677	387	1849	250000	12.9	11	0	0
0	7700	4081 ?		2695	185000	13	60	1	1
0	6700	3350 ?		2948	250000	12.8	30	0	1
0	5530	3649 ?		1382	273000	12.7	41	0	1
0	3300	1551	165	1485	120000	12.4	13	0	1
0	5640	4455	282	846	201000	10.2	7	0	1
0	3160	1706	252	1169	174000	11.8	43	0	1
0	4480	2640	317	1523	283000	11.7	70	0	1
0	4570	3381	137	776	276000	10.6	10	0	1
0	7600 ?	?		1623	20000	13	58	0	1
0	5800	3016	290	1914	224000	11.6	55	1	0
0	4560	3192 ?		912	136000	14.6	6	0	2
0	5800	4176	348	1218	181000	11.5	10	0	1
0	4040	2585	323	1171	188000	13.2	25	0	1
0	5300	2544	742	2491	178000	12.8	35	1	2
0	7100	3470	500	3124	228000	12.3	72	0	1
0	3300	2079	396	759	220000	11	88	0	1
0	10400	8008 ?		2080	416000	12	38	1	0
0	5500	3685	55	1736	248000	13.7	32	0	1
0	5030	2660	382	1700	414000	9	49	0	1
0	6430	3600	578	2121	213000	12.3	27	0	0
0	5300	4240 ?		795	265000	10.6	50	0	1
0	5800	4176 ?		870	218000	14	25	0	1
0	4300	2967	301	946	417000	12	22	1	1
0	9700	5432	291	3686	387000	12.3	70	0	0
0	4090	2576 ?		1022	169000	13.2	20	0	0
0	5200	3640 ?		1560	276000	12	30	0	0
0	7660	4442	536	2527	208000	12.6	20	0	0
0	3820	2712	382	649	252000	14.5	21	0	1
0	4800 ?	?		1623	217000 ?		23	0	1
0	5500	2750	660	1650	179000	13.2	15	0	1

Figure 17. An instance of the automatically generated curated dataset.

A second instance of the curated dataset is depicted in Figure 18. According to Figure 18, the features “ANA (titer-1)” and “C3 (mg/mL)” are “fair” and the features “Anti-HIV I/II (0-1)”, “IgG”, “IgM”, “IgA”, “AMA(titer-1)”, “Anti-TPO (0,1)”, “Anti-TG (titer)” are “bad”. In addition, 1 field is marked as outlier, 29 as inconsistent and 127 as missing values where 118 are followed by a question mark since they are “bad” features and 9 fields are imputed for the feature “LDH” since it is a “fair” feature without any incompatibilities. The value “797” of the feature “IgM” is marked as outlier since it is a value that deviates from the standard distribution. The values in the feature “ANA (titer-1)” are filled with red color since they are inconsistent (i.e., fractions).

Anti-HIV I/II(0-1)	ANA(titer-1)	IgG	IgM	IgA	LDH	AMA(titer-1)	Anti-TPO(0,1)	Anti-TG (TITER)	C3(mg/mL)
0	1/160	?	?	?	273	0	0	0	117
?	1/80	1940	797	455	406	1	?	?	65
?	1/1280	?	?	?	273	?	?	?	94
?	1/640	2470	108	369	278	?	?	?	104
0	1/640	1690	194	384	213	0	1	1	32
?	1/640	1320	82	249	273	?	?	?	77
?	1/640	?	138	361	169	0	?	?	87
0	1/640	2580	104	519	124	0	0	0	82
?	1/640	?	?	?	273	0	?	?	?
?	1/640	?	?	?	273	?	?	?	80
0	1/160	?	?	?	159	0	0	0	117
?	?	?	?	?	273	?	?	?	84
?	1/320	?	?	?	271	?	1	0	87.7
0	1/640	1830	194	513	169	?	?	?	103
0	1/1250	1630	98	158	324	0	0	1	120
0	1/640	1250	256	245	375	?	?	?	111
0	1/640	2250	112	402	443	0	1	0	78
?	0	?	?	?	381	0	0	0	181
0	1/1250	1720	155	310	306	?	?	?	107
0	1/640	?	?	?	187	0	0	1	97
0	<1/160	795	104	254	174	0	0	0	123
?	1/640	?	?	?	293	?	?	?	191
?	1/320	?	?	?	495	?	?	?	118
0	1/640	1650	636	356	253	0	0	0	104
0	1/1280	?	?	?	167	1	?	?	138
?	1/160	1630	110	140	173	0	1	1	110
?	1/320	?	?	?	273	?	?	?	120
?	1/160	?	?	?	183	0	?	?	103
0	1/640	?	?	?	273	0	0	0	102
?	1/640	?	?	?	273	?	?	?	102
?	1/320	?	?	?	362	?	1	0	118

Figure 18. A second instance of the automatically generated curated dataset.

3.3.8. The "clean" curated dataset

The clean curated dataset is the pure version of the diagnostics dataset from the REST service of the data curator, where the "bad" features are automatically removed from the data.

3.3.9. An instance of the REST API service of the data curator

An instance of the main screen of the REST API service of the data curator which was developed under the aegis of this thesis is depicted in Figure 19. From there, the end-user can select: (i) a method for outlier detection, (ii) a method for similarity detection (de-duplication), and (iii) a method for data imputation.

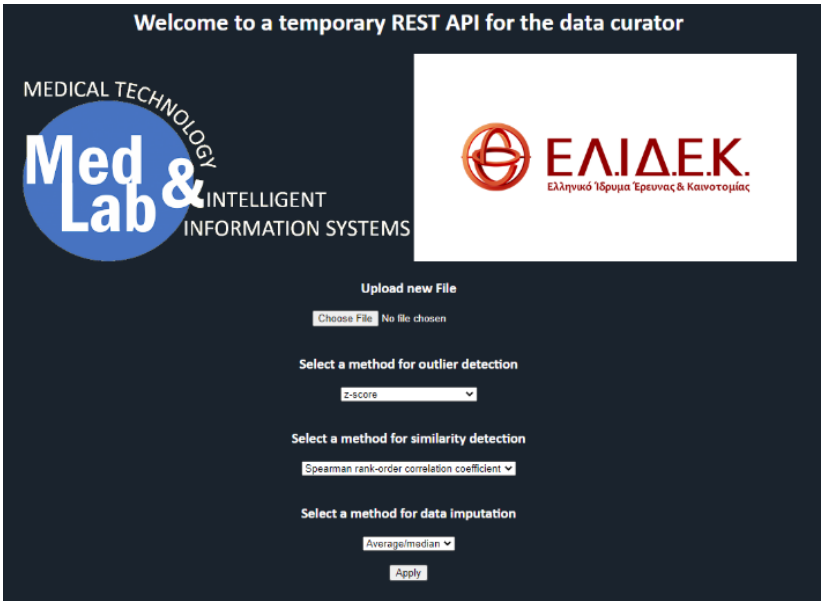


Figure 19. The main screen of the REST API data curation service.

According to Figure 20, the REST API service currently supports both univariate and multivariate methods for outlier detection, including the z-score, the interquartile range (IQR), the Grubb's test, the local outlier factor (LOF), the isolation forests, and the modified version of the isolation forests.

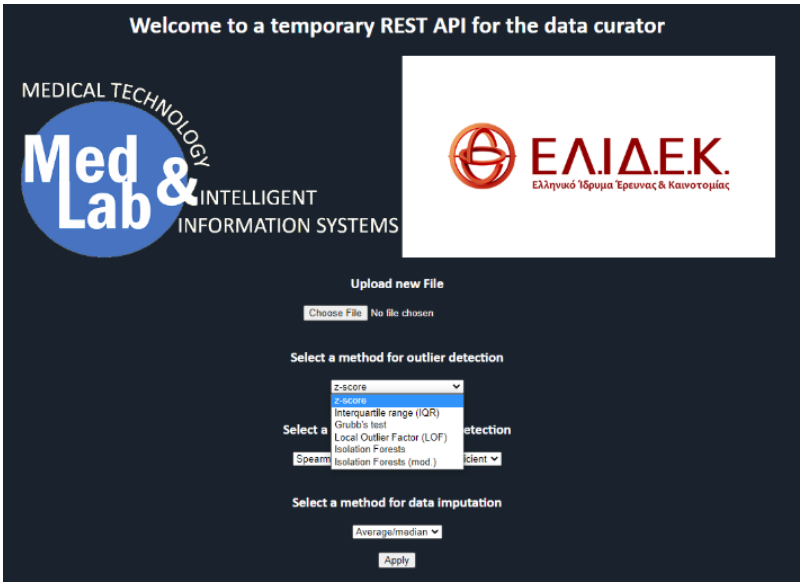


Figure 20. The outlier detection methods of the REST API data curation service.

According to Figure 21, the REST API service currently supports four methods for pairwise similarity detection among the features, including the Spearman rank-order correlation coefficient, the Pearson's correlation coefficient, the Kendall's tau, and the covariance (in the case of a gene expression dataset as the input dataset).

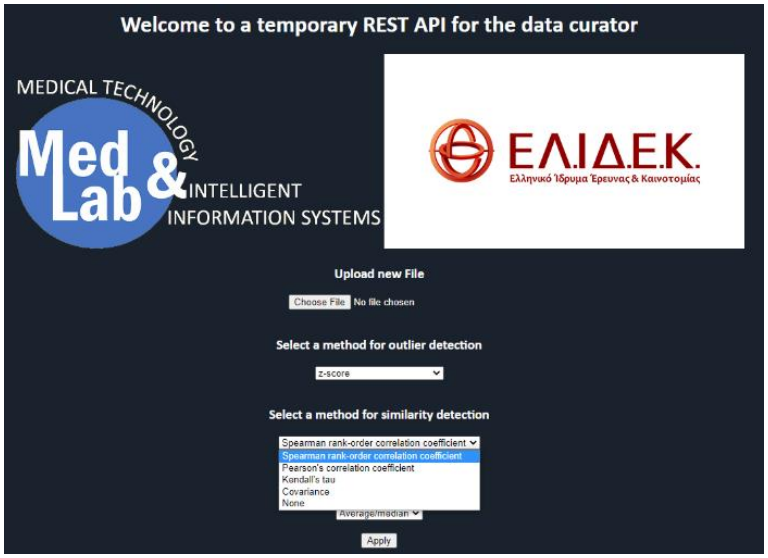


Figure 21. The similarity detection methods of the REST API data curation service.

According to Figure 22, the REST API service currently supports three methods for data imputation, including the average/most frequent, the random and the imputation with zeros which is a special case in the case of gene expression data with missing gene counts.

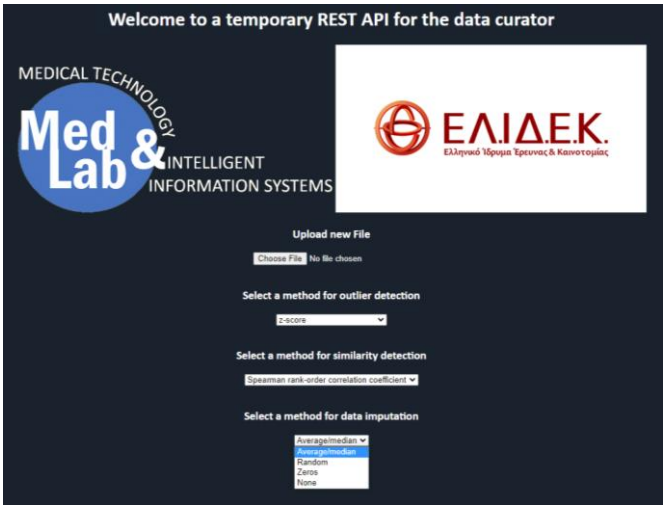


Figure 22. The data imputation methods of the REST API data curation service.

3.3.10. Alternative color coding

An instance of the data assessment report for an indicative SILICOFCM dataset is depicted in Figure 23. For each feature, the value range, data type, variable type, number of missing values, mean/median value, state, presence of outliers and incompatibilities is recorded along with useful metadata at the top row (number of features, number of instances, number of discrete and continuous features, number of unknown features, missing values).

	Number of feature(s)	Number of instance(s)	Discrete feature(s)	Continuous feature(s)	Unknown feature(s)	Missing values (%)
Metadata	157	364	73	84	0	31.73% (18134.0)

Quality assessment

Features	Value range	Type	Variable type	Missing values	Mean/most freq	State	Outliers	Incompatibilities
City	[1, 4]	categorical	int	0	2	good	no	no
Center	[1, 4]	categorical	int	0	2	good	no	no
Wight	[45, 142]	numeric	float	71	81.26	fair	yes	no
Height	[85, 193]	numeric	float	72	171.8	fair	yes	no
BMI	[6.1353938828929, 45.2005259697567	numeric	float	71	27.35	fair	yes	no
Age	[14, 95]	numeric	int	0	55.68	good	no	no
Sex	[0, 1]	categorical	int	0	1	good	no	no
fatigue	[0, 1]	categorical	int	0	1	good	no	no
dyspnoea	[0, 1]	categorical	int	0	1	good	no	no
chest_pain	[0, 1]	categorical	int	0	1	good	no	no
palpitations	[0, 1]	categorical	int	0	1	good	no	no
syncope	[0, 1]	categorical	int	0	1	good	no	no
other	[0, 1]	categorical	int	0	1	good	no	no
NYHA_class	[1, 3]	numeric	float	58	2.12	fair	no	no
systolic	[12, 260]	numeric	int	0	126.24	good	yes	no
diastolic	[50, 140]	numeric	int	0	76.97	good	yes	no
heart_murmur	[0, 1]	categorical	int	0	1	good	no	no
pulmonary_cracles	[0, 1]	categorical	int	0	1	good	no	no
pleural_effusion	[0, 1]	categorical	int	0	1	good	no	no
pretibial_edema	[0, 1]	categorical	int	0	1	good	no	no
venous_congestion	[0, 0]	categorical	int	0	0	good	no	no
hypertrophic_cardiomypathy	[0, 2]	categorical	int	199	1	bad	no	yes, bad feature
dilated_cardiomyopathy	[0, 1]	categorical	int	223	1	bad	no	yes, bad feature
sudden_cardiac_death_age_lt_40	[0, 2]	categorical	int	207	1	bad	yes	yes, bad feature
sudden_cardiac_death_age_40_59	[0, 2]	categorical	int	207	1	bad	yes	yes, bad feature
sudden_cardiac_death_age_geq_60	[0, 1]	categorical	int	209	1	bad	no	yes, bad feature
unexplained_heart_failure	[0, 1]	categorical	int	215	1	bad	no	yes, bad feature
cardiac_transplantation	[0, 1]	categorical	int	213	1	bad	no	yes, bad feature
pacemaker_or_defibrillator_implants	[0, 1]	categorical	int	214	1	bad	no	yes, bad feature
evidence_of_systemic_disease	[0, 1]	categorical	int	254	1	bad	no	yes, bad feature

Figure 23. An instance of the data assessment report.

An instance of the curated dataset is depicted in Figure 24 using updated color coding.

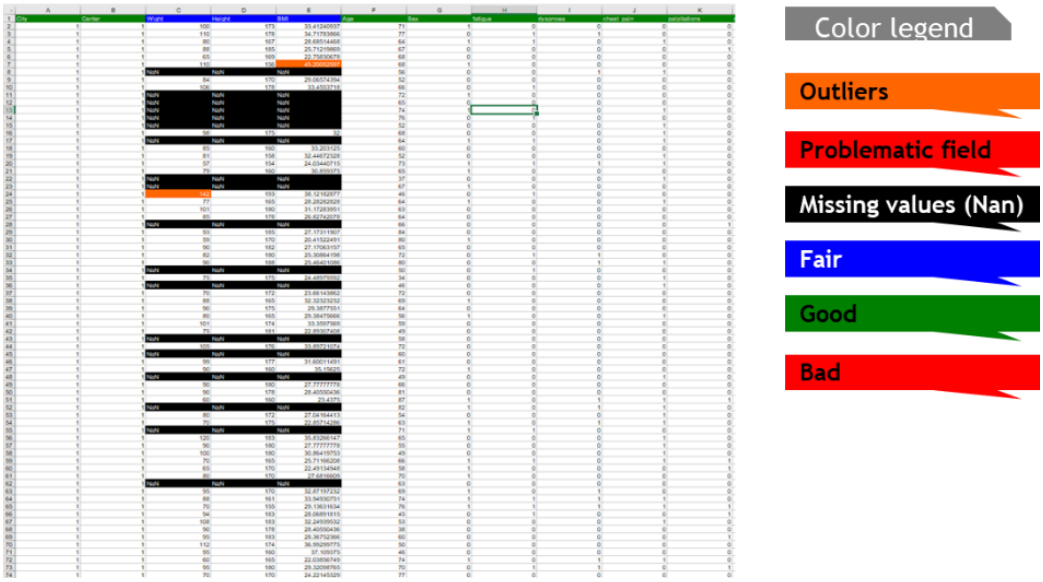
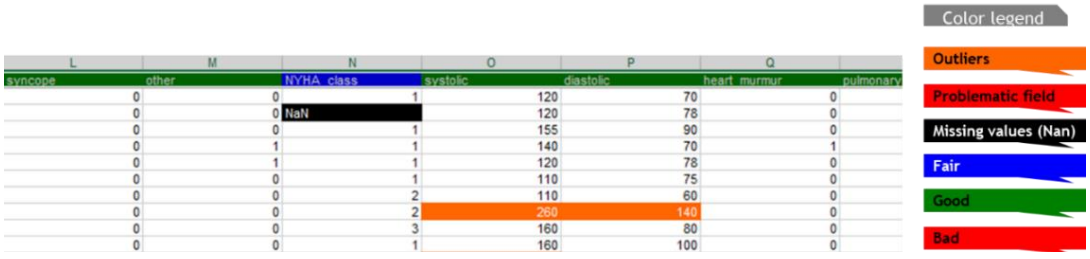


Figure 24. An instance of the curated dataset with appropriate color coding for data quality control.

A second instance of the curated dataset is depicted in Figure 25 presenting the successful identification of two potential outliers in the features “systolic” and “diastolic” pressures. More specifically, a value 260 has been recorded for the systolic pressure and a value 140 for the systolic pressure regarding the same patient. These values deviate from the standard population distribution, in each case, and thus they have been highlighted with orange color for easier inspection by the clinical experts.



L	M	N	O	P	Q	
syncope	other	NYHA_class	systolic	diastolic	heart_murmur	pulmonary
0	0	1	120	70	0	
0	0	NaN	120	78	0	
0	0	1	155	90	0	
0	1	1	140	70	1	
0	1	1	120	78	0	
0	0	1	110	75	0	
0	0	2	110	60	0	
0	0	2	260	140	0	
0	0	3	160	80	0	
0	0	1	160	100	0	

Figure 25. An instance of the curated dataset with the presence of two outliers.

3.4. Summary

Data quality has been recognized as a key factor in all operating processes both in the public and private sectors. It has been characterized as a multidisciplinary process since it reflects the needs for sustainable data of high quality in several domains varying from

business to healthcare. The technological advances of our era combined with the fact that the structure of the current information systems is network-based, have dramatically increased the amount of digital data. A crucial consequence of this evolution is that data management has become more complex and controversial. This need has increased the necessity for automated methods and rules that are able to deal with the quality assessment of big data. Lack of data quality results in bad data manipulation which makes data useless and has numerous negative effects on further processing. Thus, emphasis must be given on the development of new methods for dealing with insufficient data sources.

The most important process of a data management system is the data quality assessment. The data quality assessment process is related to: (i) the evaluation of data protection metrics (e.g., data protection impact assessment), (ii) the organizational structure of the data, and (iii) the overall information management. Several studies [66], [125], [136], [138]–[140] were launched highlighting the leading role of data quality assessment in improving the information quality especially in the medical domain. To assess the quality of the data, one must first define the quality requirements and metrics. Examples of common data quality requirements include the: (i) accuracy, (ii) completeness, (iii) consistency, (iv) interpretability, (v) timeliness, (vi) relevancy, and (vii) ease of manipulation, among many others. In general, multiple metrics can be associated with each quality requirement. For example, completeness, in the field of data science, can be defined as the degree to which a given dataset meets the pre-defined requirements of an optimal dataset. In this case, completeness is a quality requirement that can be quantified from data-driven (quality) metrics, such as, the number of missing values, incomplete terminology, etc.

It has been long proven that the quality of data mining results and related applications highly depend on the quality of the data. For example, a clinician who wishes to apply a simple regression model on a contaminated clinical dataset (in the presence of outliers and/or incompatible values) to identify independent factors for a particular disease or develop a prediction model for the disease progress, will end up with a distorted model with no clinical value. In addition, the structural heterogeneity of the clinical data across different clinical centers introduces biases during the analysis of medical data. The heterogeneity and non-canonical form of medical data resulting from either the bad

quality of the medical data or the lack of a standard clinical vocabulary hampers data mining and other processing tasks.

Data standardization is a promising solution for transforming the data into a common format [52], [73], [75] and interoperability standards and modeling annotations in turn are key factors towards the success of data standardization. Standardization is usually performed according to a gold standard model which serves as a reference model, i.e., a set of parameters which describes the requirements (e.g., variables, types, descriptions) of a disease of interest [52], [73], [109], [129], [141]. Most of the studies make use of gold standard models to assess the quality of different types of clinical data based on various data quality measures, such as, the accuracy, the completeness, etc.

Data standardization has many similarities with data harmonization. The latter is of great importance since it aims to overcome the heterogeneity of medical data worldwide by converting the heterogeneous data into homogeneous ones (e.g., with similar structure and terminologies). Data harmonization [52], [73], [109], [129], [141] involves several mechanisms including data transformation to a common format, data annotation, terminology detection and alignment, most of which are part of the data curation framework and especially of the data standardization process. It is, therefore, important to consider data standardization as a crucial part of a data quality assessment framework and a key part of data harmonization.

According to Table 6, most of the studies on data curation mainly focus on providing general guidelines for data quality assessment, methodological steps towards data curation and standardization, without, however, focusing on the development and evaluation of a computational framework for data quality assessment on medical data. To address this need, we presented the objectives, functionalities, and methodological advances of an integrated framework for medical data curation in terms of data quality assessment. The proposed framework consists of a three-layer architecture and serves as a diagnostic tool for managing incomplete terminologies, irrelevant terms, outliers, missing values, data categorization, and duplicated terms. In the core of this framework lies data standardization.

In this thesis, we extend data standardization as a pre-harmonization process to make data harmonization easier and faster. More specifically, we use lexical matching

combined with model-based rules and external sources, i.e., vocabularies, to match and classify terms according to a pre-defined reference model which is a set of parameters which describe the requirements (variables with their types and ranges) of the clinical domain of interest. Through this procedure, we attempt to produce semantic relations between the fields of the raw dataset with those from a reference model and therefore enhance the semantic matching process for data harmonization. To our knowledge, this is the first fully automated, highly scalable, and efficient data curation framework with a REST API service for medical data quality assessment.

CHAPTER 4. DATA HARMONIZATION

4.1. Overview

The lack of standardized data (data heterogeneity) among biobanks, cohorts, and other sources of medical data is a crucial scientific challenge which obscures the effective analysis of such data, yielding clinical studies with poor statistical power [52], [73], [129], [147]. Data harmonization is an emerging research field which aims to overcome the existing heterogeneities across multiple medical data sources [52], [73], [74], [109], [129], [147], [160], [161]. The heterogeneity of data among biobanks, cohorts, and other sources of medical data is a critical scientific limitation which poses significant obstacles in the effective analysis of such data, yielding clinical studies with poor statistical power and, thus, inaccurate disease outcomes [52], [73], [74], [109], [129], [147], [160], [161]. In computer science, data harmonization is an emerging technique which aims to overcome the structural heterogeneities that are present among the medical data derived from multiple sources by producing homogenized versions of the heterogeneous data that share a common medical domain (context). The overall idea of data harmonization is to transform the heterogeneous data into a common format with the exact same parameters and range values, using data-driven, and other computational approaches, such as, lexical, and semantic matching, to enable the integrative analysis of the heterogeneous data and therefore, enhance the statistical power of the clinical

studies which make use of such data. Based on the above concept, data harmonization can enable the interlinking and subsequent integration of clinical data to deal with the unmet needs in various diseases.

4.1.1. The value of data harmonization in healthcare

The prospects of data harmonization in the medical domain are many. Harmonization can overcome the lack of common (standard) data collection protocols, a fact that introduces biases during the collection of medical data among multiple sources of medical data (e.g., clinical centers), especially during the recording of the measurement units of various laboratory-related attributes. This can be accomplished by either normalizing the measurement units according to a pre-defined range that is already defined in the standard model or based on statistical approaches, through a procedure which is known as data standardization. In contrast to data standardization, data harmonization is a more generalized strategy which aims to first align the structure of the datasets and then apply data standardization as part of the normalization process.

In fact, data harmonization involves: (i) the identification of terminology matches among the heterogeneous data through a procedure which is known as terminology mapping, and (ii) the application of data standardization to normalize the measurement scales across the matched terms through a procedure which is known as terminology alignment. The fact that the data harmonization process can be performed in a semiautomated manner reduces the time effort that is needed by the clinicians to manually homogenize their clinical data, which is rather impossible in the case of large datasets, where the number of parameters is vast, like in omics data [307]–[310]. The clinician's involvement can be further reduced by constructing machine learning algorithms which are able to learn from external sources, such as, medical index vocabularies, through the development of semantic interlinking mechanisms which are able to speed up the data harmonization process.

The fact that data harmonization allows the integration of heterogeneous data can provide great insight on the assessment of the unmet needs in various diseases [52], [73] through: (i) the development of more robust risk stratification models for the early identification of high-risk individuals, (ii) the identification of new prominent

biomarkers for the prediction of disease outcomes, and (iii) the development of new, targeted, therapeutic treatments and health policies.

Why is it necessary to harmonize the data? As it has been already mentioned, data harmonization can enable the integration of clinical data from cohorts, and other similar sources of medical data, that coexist under a specific clinical domain (e.g., that describe a common disease of interest) which can reveal valuable clinical information regarding a disease's onset and progress over time. Indeed, the integration of harmonized data from multiple data sources can significantly enhance the population subgroups that participate in the clinical studies and thus yield more powerful patient stratification models which are able to precisely identify groups of individuals (subjects) that are more prone to the development of a disease outcome [311]–[314]. These individuals are referred to as high-risk individuals who belong to specific subgroups and, thus, the early diagnosis of such subgroups is of great importance for enhancing the quality of the existing healthcare systems. Since data harmonization can transform the clinical data into standardized formats, it can enable the interlinking of electronic health records with electronic health record systems worldwide [311]–[314].

The integration of harmonized data can also lead to the identification of prominent clinical factors having significant contribution in the prediction of one or more disease outcomes, as well as, the confirmation of the existing prominent clinical predictors, especially in the case where the type of the disease under investigation is rare and/or chronic [52], [73]. These clinical predictors are referred to as biomarkers and can reveal the underlying mechanisms of various diseases. The clinical importance of biomarkers is high in the case of genetic data, where phenotypes and genetic variants are present [315]. Moreover, the outcomes of clinical studies that make use of integrated, harmonized data enable the development of more targeted therapeutic treatments for the different population subgroups that can greatly enhance or replace the existing treatments and thus shed light into the progress of a disease over time, as well as, promote the establishment of new healthcare policies by the healthcare stakeholders.

Apart from the integration of harmonized data into centralized databases, another important advantage lies on the interlinking of harmonized data that are stored in distributed clinical databases which can enable the development of data analytics tools for analyzing the data in distributed environments. In general, a centralized data

repository is more vulnerable to security threats and privacy breach and thus it is not always feasible to maintain the data under a common database [80]. A solution to this, is to store the data in distributed databases and then interlink them with multiple authentication levels for increased security. So, the only way to analyze the data that are stored in distributed sites is to develop new machine learning models or extend the existing ones so that they can incrementally process the individual data on each site. For example, assume that a researcher wishes to compute population characteristics (e.g., descriptive statistics) across the distributed data or, in a more demanding case, assume that the researcher wishes to run a risk stratification model for predicting disease outcomes across distributed data. These two scenarios can only be feasible in the case where the individual data, on each site, have a common format, i.e., a common set of parameters and standardized values. It is obvious now that this limitation can be effectively addressed by harmonizing the individual datasets on each site so that the data model can be able to adapt on the same set of parameters.

4.1.2. Types of data harmonization methods

In general, there are two conceptual approaches/strategies to accomplish data harmonization, namely the stringent and the flexible strategy [52], [316], [317]. The former limits the harmonization process only on data that will be (or have been) collected under common measurement procedures (standards) whereas the latter approach extends the harmonization process to include data that have been already collected under different measurement procedures or protocols [52], [316], [317]. Here, emphasis is given on flexible methods that enable the harmonization of retrospective clinical data due to its underlying complexity and its clinical importance.

4.1.2.1. The stringent approach

The stringent approach is an ideal strategy which constrains the harmonization process to clinical data that have been collected under common collection criteria and operating procedures [52], [316], [317], where the common data collection criteria refer to the adoption of identical study specifications (uniform measures) between the clinical studies that participate in the data harmonization process. These specifications, include [52], [316], [317]: (i) common inclusion and exclusion criteria for the definition of the population subgroups, (ii) common follow-up time periods, and (iii) a common set of

qualitative and quantitative measures (e.g., therapies), among others. These specifications together constitute a data collection protocol and are exclusively designed by domain experts who can identify: (i) the domain of the field of interest (type of study), (ii) the set of measures that should be collected for the specified study, and (iii) the standardized measurement units for the recommended set of measures for the particular type of study. The range of diseases that can be covered by the stringent approach can be vast as long as the data follow the same standard operating procedures.

According to the stringent approach, the clinical studies that participate in the data harmonization process must be initially designed to meet these specifications to be harmonized and finally synthesized, otherwise the data harmonization process will fail. It is obvious that these requirements are strict and limited to only a small portion of clinical centers and other similar sources of clinical data that adopt common data collection criteria and standard procedures. Of course, most of the clinical centers do not follow identical procedures for the data generation process and thus stringent harmonization remains a conceptual and ideal strategy for the scientific community. This is highly present in the case of retrospective data harmonization, where the stringent approach is useless since the data have been collected in the past, where a standard data collection protocol is usually absent. The stringent approach would be meaningful in the case of a prospective study of perhaps in a cross-sectional study which focuses on data that have been obtained at a specific time point although it would require a substantial amount of time to be prosperous.

4.1.2.2. The flexible approach

The stringent method is a strict and a rather ideal approach that significantly limits the statistical power of the data harmonization process since it obscures the integrity of the produced harmonized data through the underlying information loss and limits the harmonization to a small portion of data that have been collected under the same standard operating procedure. An alternative approach that aims to deal with the limitations that are posed by the stringent approach is flexible harmonization [52], [316]. As its name implies, the flexible approach allows a certain level of heterogeneity between the data which participate in the harmonization process instead of the stringent case where the complete absence of heterogeneity among the individual data is required. Therefore, the flexible approach can support the harmonization of both

prospective and retrospective data as far as the level of compatibility between them is well-defined. Through this manner, the flexible methodology envisages to enable the harmonization of data that do not necessarily need to be homogeneous or obtained under a common data collection protocol criteria with equal-sized populations.

In flexible harmonization, the level of heterogeneity of the data directly affects the percentage of harmonized variables across them. This implies that the amount of flexibility is constrained to a specific set of requirements that need to be defined. That is, the set of clinically relevant parameters (factors) that will be common among the heterogeneous data. Of course, the clinical domain where the data that participate in flexible harmonization belong to, must be common. To facilitate flexible harmonization, the clinical experts must first define a set parameter (variables) that will serve as the core set for the domain of interest allowing for a specific level of flexibility regarding the data collection protocol and the standard operating procedures [52], [316]. Therefore, flexible harmonization is constrained to specific outcomes that are defined by the clinical experts. In the prospective case, the core set of variables is defined and agreed to by the experts to allow a specific level of flexibility during the recording of the follow-up data. In the retrospective case, the core parameters are combined with pairing rules to identify potential associations and thus quantify the harmonization accuracy.

The flexible strategy is far more realistic and has a much higher clinical value and overall applicability than the stringent approach, although, in both cases, certain compatibility criteria must be carefully defined so that harmonization can be feasible. The compatibility criteria are expressed in the form of a set of standard variables, i.e., a core set of variables, that describes the requirements of the clinical domain of interest. In both cases, however, the standard model is defined by the clinical experts in the field in such a way to: (i) be in line with the majority of the parameters within the data that are collected by different clinical centers, and (ii) explicitly describe the domain knowledge of the disease under investigation. This means that the clinical experts select the variables of the standard model by taking into consideration: (i) the contribution of each variable towards the efficient description of the disease's domain knowledge, and (ii) the extent to which these variables are present in the majority of the data that exist under each clinical center. Additional information regarding the format (and the type)

of the medical data that will be involved in the harmonization process, along with the ethical and legal concerns, the quality of the data, as well as, the precise definition of the objectives, i.e., the reason behind data harmonization, should also be taken into consideration for the realization of the flexible strategy [52], [316].

The level of heterogeneity and diversity between the data is more or less trivial and can be reflected by the structure of the core set of variables. For example, the clinical experts of a specific domain might consider that a set of N-variables is enough to describe the knowledge of the domain under investigation whereas another group of experts, on the same field, might consider the need to add more (or less) variables in the core set. Therefore, the experts must agree on a core set of variables that overlaps with most of the data to increase the harmonization accuracy. This can be extremely difficult especially in the case of retrospective flexible harmonization where the data might have been collected under diverse protocols and might exist under different identifiers, as well. The same stands for the objectives of the data harmonization process. The data providers must clarify the scope of data harmonization, as well as, provide any kind of information regarding the study design and the specifications that were used for the data collection process which are valuable for the definition of a more accurate core set of variables.

According to the literature, a variety of computational methods for medical data harmonization has been proposed so far [52], [73], [111], [141]–[143], [145], [146], [151], [160], [316], [317]. A robust data harmonization method involves the application of lexical and semantic matching algorithms. A lexical matching algorithm [52], [143] uses string similarity techniques to identify common terminologies (i.e., exact sequences or similar block sequences) that are present between the terms of the standard model and those of the original dataset. External vocabularies can also be used to enrich the clinical domain knowledge and thus enhance the accuracy of the overall lexical matching process through the identification of homonyms or synonyms. On the other hand, the semantic matching method [52], [73], [142], [145] uses semantic relationships that exist between the terminologies, apart from the lexical matching process that is already included, to reduce the information loss and enhance the overall data harmonization process. This can be accomplished through the construction of ontologies which represent the clinical domain knowledge of interest in the form of

entities (e.g., classes), and object properties (e.g., “includes”, “has”, “consists of”) [52], [73], [142], [145]. Semantic matching uses a standard (or reference) model which is usually expressed in the form of an ontology, where the classes are considered as categories, e.g., “Clinical tests”, that might consist of further sub-classes, e.g., “Blood Tests”, etc. Each class can include a set of variables which are related to the class they belong to in terms of common meaning or concept [52], [73], [142], [145]. For example, the class “Blood tests” includes the variables “age”, “gender”, “hemoglobin levels”, etc. This can lead to the semantic interoperability of the variables which might not be lexically identical but share a common concept.

Towards this direction, semi-automated data harmonization frameworks have been proposed to co-analyze heterogeneous data in biobanks and other related registries, including the DataSHaPER [141], [147] which was used to harmonize 53 databases on epidemiology with 36% compatibility, the SORTA tool [142] which was used to match 5120 entries in a single biobank with 97% recall, and the BiobankConnect software [143] which was used to harmonize data across 6 biobanks with 74% precision. In [318] an open-source editor was also developed to provide standardized HL7 data formats and in [319] a semi-automated lexical matcher was used to map 78.48% of 1,492 biobank terms. Statistical approaches, such as, multiparameter logistic Item Response Theory (IRT) analysis were deployed in [160]–[162] to examine how a set of items (e.g., psychiatric phenotypes) is affected by other factors for scale homogenization.

Another family of data harmonization methods is statistical harmonization which involves the application of linear and non-linear statistical models to investigate the effect of different latent factors on a set of one or more items [160]–[162]. In statistical theory, the items can be considered as all types of informative variables (e.g., depression) that are observed and the latent factors as variables that are not directly observed but are rather inferred by the items [160]–[162]. The purpose of statistical data harmonization is to homogenize scales that measure the same item and transform them into a common metric of the same scale, where the types of the items might vary from discrete and ordinal to continuous [160]–[162].

For example, a clinical center might record the cholesterol levels using the scale low, medium, high, whereas another clinical center may record the same levels using the scale 0, 1, and 2. Thus, statistical harmonization tries to recode the variables that belong

to the same construct so that they are commensurately scaled at the end [160]–[162]. Of course, the detection of variables that express different (or common) scales and belong to the same construct is challenging since there is no prior knowledge regarding the names of the items like in lexical or semantic matching [160]–[162]. Even if two variables (items) describe the same construct, it is not always proper to match these two variables since there might be differences in the population characteristics between the clinical studies, e.g., differences in the education level, ethnicity, gender, etc. Such differences need to be controlled during the statistical analysis process and thus the complexity of the harmonization process is greatly enhanced.

Moreover, the types of the items directly affect the type of the statistical model to be used. Towards this direction, a variety of statistical methods has been proposed so far, especially for the harmonization of psychometric and cognitive items with different measurement scales across clinical data, including [160]–[162]: (i) simple linear factor analysis (LFA) for continuous items, (ii) 2-parameter and multi-parameter logistic Item Response Theory (IRT) analysis for binary items, (iii) generalized linear factor analysis (GLFA) for mixtures of continuous and discrete items, and (iv) moderated non-linear factor analysis (MNFA) for mixtures of continuous and discrete items with non-linear dependencies, among others.

4.2. Beyond the state of the art

Most of the current computational efforts towards semi-automated data harmonization involve the definition of a global standard (common) procedure for data collection which is ideal in the case of prospective data (i.e., data that will be updated in the future). The difficult and most challenging part though, is the need to harmonize retrospective data (i.e., data that have been already collected in the past with the absence of a pre-defined standard data collection protocol). A fundamental drawback of these methods, however, lies on the fact that they are not easily generalizable to other clinical domains. Moreover, they adopt a semi-automated strategy which is based on the extensive collaboration between the clinical and the technical experts to define pairing rules, i.e., a pre-defined set of rules for lexical matching. The performance of the existing frameworks is low in several cases due to the complexity of the clinical field under investigation, such as, in [141], [143], [147] but also due to the lack of straightforward computational methods to enable automated terminology matching

[141], [142], [145], [147]. To address these needs, we propose an automated data harmonization workflow which adopts a hybrid strategy that combines lexical analysis with semantic models (ontologies) to identify terminologies with lexical and conceptual overlap.

4.3. Lexical matching

4.3.1. The edit distance problem

The most common method for lexical matching involves the computation of the edit distance between two strings, assume x and y which is defined as:

$$d_{x,y}(i,j) = \begin{cases} i & , \quad i = 0 \\ j & , \quad j = 0 \\ d[i-1, i-1] & , \quad i,j > 0 \; and \; x_i = y_j \\ min \begin{Bmatrix} d[i-1, j-1] + 1 \\ d[i-1, j] + 1 \\ d[i, j-1+1] \end{Bmatrix}, & \quad o.w. \end{cases}, \qquad (4.1)$$

where $d_{x,y}(i,j)$ is the distance between the first i-characters of x and the first j-characters of y. In fact, the edit distance aims to transform x into y by performing three possible types of operations, namely: (i) insertion, (ii) deletion, and (iii) substitution. Assume a string $x =$ "abc" with size 3. Insertion involves the addition of a new character, assume d, into x, so that $x =$ "abcd". Deletion involves the removal of an existing character from the string, assume c, so that $x =$ "abd". Substitution involves the replacement of an existing character, assume d, by c, so that $x =$ "abd" becomes $x =$ "abc". The Jaro (and Jaro-Wrinkler) distance [320] and the Levenshtein distance scores [321] are the most common methods for calculating the edit distance although the latter is much closer to the definition of the edit distance against the former method which takes into consideration the number of transpositions between two strings.

The latter is defined as half the number of matching characters between x and y. In fact, the Jaro distance first computes the number of matching characters, between x and y, assume m, as well as, the number of transpositions, assume t. Then, according to (4.1), the Jaro distance measures the edit distance between x and y by computing the average of the percentage of matched characters in each string with the percentage of the

transpositions in the number of matching characters. Thus, the higher the Jaro distance (i.e., the closer to 1) the more similar the two strings are.

The algorithm generates a distance matrix where each cell corresponds to a distance score. An example of the edit distance matrix for the strings x = "lymphadenopathy" and y = "lymphoma" is presented in Table 8. The edit distance value is expected to be 8 and the number of operations is expected to be the following:

1. lymph**a**denopathy -> lymphdenopathy (delete "a")
2. lymph**d**enopathy -> lymphenopathy (delete "d")
3. lymph**e**nopathy -> lymphnopathy (delete "e")
4. lymph**n**opathy -> lymphopathy (delete "n")
5. lympho**p**athy -> lymphomathy (substitute "p" with "m")
6. lymphoma**t**hy -> lymphomahy (delete "t")
7. lymphoma**h**y -> lymphomay (delete "h")
8. lymphoma**y** -> lymphoma (delete "y")

Note that in this example, no insertions are needed. The resulting distance matrix is presented in Table 8. The desired distance is the value in the last cell of the table, i.e., cell (15,8) which is 8 (as expected). This denotes that the total number of operations that is needed to transform x into y is 8. The values which are depicted in bold are equal to the consecutive costs of the operations. More specifically, the zeros in the cells (1,1), (2,2), (3,3), (4,4), and (5,5) denote that the first five characters in both strings, i.e., the characters "l", "y", "m", "p", "h", are equal, and thus the cost is 0 since neither of the three types of operations (insertion, deletion, substitution) is applied. The values 1, 2, 3, and 4 in the cells (6,5), (7,5), (8,5), and (9,5), respectively, denote that the four characters "a", "d", "e", "n" in x shall be deleted with a total cost of 4.

The value 4 in cell (10,5) denotes that the character "o" is equal in both strings and thus it is the same as before since the operation has a 0 cost (4+0). The value 5 in cell (11,6) denotes that "p" shall be replaced by "m" and thus a substitution operation is applied adding an extra one in the cost (4+1). The value 5 in cell (11,7) denotes that "a" is equal and thus the cost remains the same. The values 6, 7, and 8 in the cells (12,8), (13,8), (14,8), and (15,8) denote that the characters "t", "h", and "y" should be removed and thus 3 deletions are applied yielding a final cost 8 (5+3).

Table 8. The edit distance table for the terminologies “lymphadenopathy” and “Lymphoma”.

	-	L	y	m	p	h	o	m	a
-	0	1	2	3	4	5	6	7	8
l	1	**0**	1	2	3	4	5	6	7
y	2	1	**0**	1	2	3	4	5	6
m	3	2	1	**0**	1	2	3	4	5
p	4	3	2	1	**0**	1	2	3	4
h	5	4	3	2	1	**0**	1	2	3
a	6	5	4	3	2	**1**	2	3	2
d	7	6	5	4	3	**2**	3	4	3
e	8	7	6	5	4	**3**	3	4	4
n	9	8	7	6	5	**4**	4	4	5
o	10	9	8	7	6	5	**4**	5	5
p	11	10	9	8	7	6	5	**5**	6
a	12	11	10	9	8	7	6	6	**5**
t	13	12	11	10	9	8	7	7	**6**
h	14	13	12	11	10	9	8	8	**7**
y	15	14	13	12	11	10	9	9	**8**

4.3.2. Levenshtein distance

Another popular metric for sequence matching is the Levenhstein distance [321] which measures the similarity between two strings, assume a and b, in terms of the number of deletions, insertions, or substitutions that are required to transform a into b:

$$lev_{a,b}(i,j) = \begin{cases} max(i,j) & , \quad min(i,j) = 0 \\ min \begin{cases} lev_{a,b}(i-1,j) + 1 \\ lev_{a,b}(i,,j-1) + 1 \\ lev_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & , \quad o.w. \end{cases} , \qquad (4.2)$$

where a Levenhstein distance of zero denotes identical strings.

4.3.3. Jaro distance

The Jaro distance [305], [320] is another widely-used string similarity measure which quantifies the similarity between two strings. For two given strings, a and b, the Jaro string similarity measure, sim_J, is defined as:

$$sim_J = \begin{cases} 0 & , \quad x = 0 \\ \frac{1}{3} \cdot \left(\frac{x}{|a|} + \frac{x}{|b|} + \frac{x - y}{x} \right), & o.w. \end{cases} , \tag{4.3}$$

where x is the number of coincident characters and y is half the number of transpositions. The Jaro distance is useful for quantifying the similarity between two strings in the interval [0, 1] against the Levenshtein distance, which measures the total number of different characters and thus is preferred in the design of pairing rules for lexical matching. For example, in the case where x = "lymphadenopathy" and y = "lymphoma", the Jaro distance is equal to 0.7329 whereas the Levenshtein distance is equal to 8 and thus is less informative.

4.3.4. Jaro-Winkler distance

The Jaro-Winkler distance measure is a modification of the Jaro distance measure that uses an additional prefix scale c to give more weight to strings with common prefix of a specific length. It is defined as follows:

$$sim_{JW} = sim_J + (lc(1 - sim_J)) , \tag{4.4}$$

where l is the length of common prefix at the start of the string up to a maximum of four characters. The prefix weight is the inverse of the l that is needed to consider both strings as identical. For example, the Jaro Winkler distance between the terms "lymphocyte number" and "lymphoma score", is equal to 0.89 whereas the Jaro distance is equal to 0.73. In the same example, the Levenhstein distance is equal to 9, which denotes the number of the operations that are needed to match the two strings. Lexical matching does not consider for semantic relations but instead focuses more on matching variables with identical patterns whereas semantic matching further seeks for semantic relations.

The Jaro-Winkler distance [320] is a weighted version of the Jaro distance which uses a prefix scale p to give more weight to strings that match from the beginning for a length l. This property, however, is statistically weak since, in most cases, it yields falsified pairs. For example, in the previous example, the Jaro-Winkler distance would be 0.8664 since the subsequence "lymph" is common from the beginning and thus is given more weight due to its large length. The Jaro-Winkler distance would be useful

in the case where two strings differ in the spelling near the end, e.g., when $x =$ "Raynaud" and $y =$ "Raynaud's", where, in that case, the Jaro-Winkler distance is 0.977 whereas the Jaro distance is equal to 0.926, thus giving more emphasis to the length of the common prefix.

4.4. Semantic matching

The lexical matching method can sometimes ignore terminologies that share a common meaning or relationship. For instance, the terms "Blood tests" and "Hematological tests" are lexically heterogeneous but share a common basis since they describe the exact same type of laboratory test. The accuracy of the lexical matching process, however, can be reduced by using external vocabularies which include medical dictionaries that can be used to identify and match synonymous or homonymous terminologies. However, in the case where a dataset, assume A, includes the blood test-related variables hemoglobin, white blood cell count, and number of platelets, and another dataset, assume B, includes the blood test-related variables erythrocyte sedimentation rate, and cholesterol levels, there is no lexical matching algorithm that is capable of capturing the lexical similarity between these terms although they might express the same concept (i.e., they are both related to blood tests). As a result, the absence of knowledge regarding the relationships between the variables reduces the percentage of the matched terms and thus the harmonization performance.

4.4.1. Relational modeling

What if we could somehow use this knowledge to distil the relationships between the variables that share same context? This can be accomplished by constructing a hierarchical data presentation model or a semantic presentation of the data, where the relationships (i.e., the object properties) between the variables will be well-defined and then use these relations to match the semantic presentations of the data instead of matching the variables of the data themselves.

4.4.2. Ontologies

One way to construct a semantic presentation of the data this is to construct an ontology [322]–[325]. In an ontology [322]–[325], the data are described in the form of entities and object properties, where the entities are classes and sub-classes, and the object

properties are the relationships between them. An example of the format of an ontology is presented in Figure 26. The main class in the ontology is the “Patient”. The “Patient” is connected to the sub-class “Laboratory tests” through the object property “has”, i.e., the patient has laboratory tests. A sub-class can also consist of further sub-classes. For example, the sub-class “Laboratory tests” consists of the sub-classes “Blood tests”, “Oral tests”, “Urine tests”, and “Ocular tests”, where the object property “consist of” is used to denote this relationship.

A sub-class can also include variables where the relationship between the sub-classes and the variables are denoted by the object property “include”. In this example, the sub-class “Blood tests” includes the variables “hemoglobin” and “white blood cell”. In a similar manner, the sub-class “Urine tests” includes the variables “urine pH” and “urine gravity flow”, and the sub-class “Ocular tests” includes the variables “Rose Bengal score” and the “ocular staining score (OSS)”. To demonstrate the structural complexity that an ontology might have, the sub-class “Schirmer’s test” has been added under the sub-class “Oral tests” which is part of the sub-class “Laboratory tests”. The sub-class “Schirmer’s test” includes two variables, namely the “date” when the test was conducted along with the test’s “score”. In general, the levels of an ontology can be larger, especially in disease-oriented ontologies where the domain knowledge is vast.

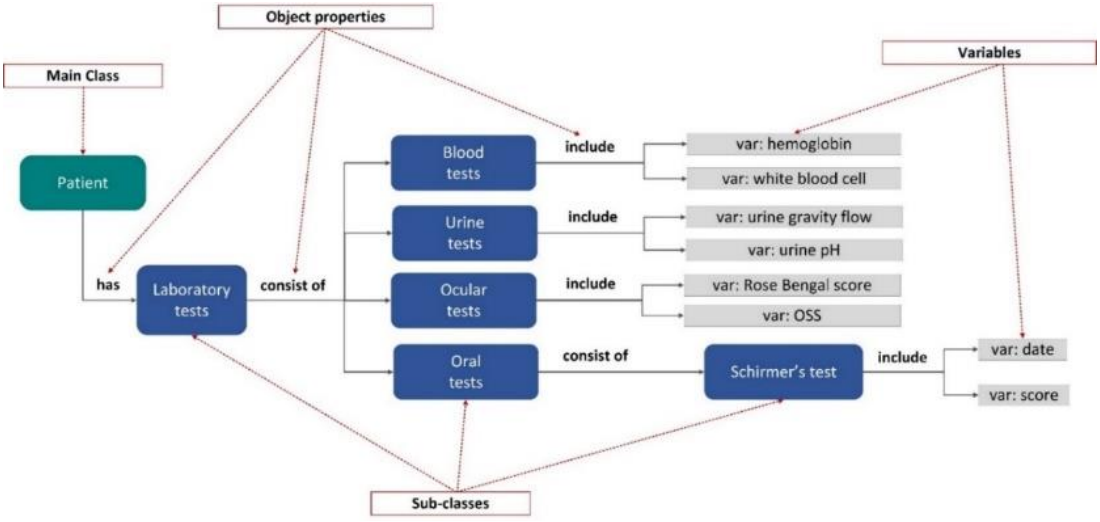


Figure 26. The fundamental components of an ontology *[52]*.

Semantic web technologies [326], [327] provide a rigorous solution to automatically integrate disparate information sources and database schemas. The World Wide Web Consortium (W3C) [328] provides solutions to the expression of both data and rules for

reasoning. The semantic models can be expressed in different formats including: (i) the Resource Description Framework (RDF) [329], (ii) the Web Ontology Language (OWL) [330], and (iii) the Extensible Markup Language (XML) [331], among others.

4.4.3. Reference ontologies as data quality standards

In order to enable the harmonization of individual cohort data it is necessary first to develop a standard model which will describe the minimum requirements that are needed to exclusively describe the domain knowledge of the disease of interest. In the case where only associations need to be computed between the individual cohort data, one cohort dataset can be used as a standard model that can be aligned (harmonized) with each one of the remaining cohort datasets. These minimum requirements consist of a set of medical parameters, value ranges and descriptions that can effectively describe the related domain. This set of parameters is used as a reference template according to which the data will be harmonized. Therefore, the third requirement is the existence of a reference model for the disease under investigation. Data analytics workflows can then be applied on the harmonized data to extract valuable information regarding the disease's onset and progress. Finally, the transparency of the workflows and any related operations that are performed during the data harmonization and the data analytics procedures needs to be reassured.

The reference model is usually expressed in a semantic form through an ontology which provides a hierarchical representation of a specific clinical domain based on a set of entities (i.e., classes and subclasses) and object properties (i.e., parameters) that explicitly describe the knowledge of a particular clinical domain [52], [73], [109], [141], [147]. Most of the methods for retrospective data harmonization make use of a pre-defined, standardized model which describes the requirements of a particular clinical domain and serves as a common template (i.e., a gold standard) for harmonization [52], [73], [109], [141], [147]. This standard template includes a set of clinical variables (parameters) that can describe the domain knowledge of a disease of interest. Of course, the existence of a standard template is not always necessary especially in the case when the scope of data harmonization is to seek for variable associations between two (or more) heterogeneous datasets which are clinically relevant. In that case, one dataset can be considered as the reference model.

4.4.4. HL7-standards

No matter how healthcare data is kept in the various computer systems, it may be transferred between them according to the HL7® FHIR® (Fast Healthcare Interoperability Resources) [332] standard. It makes it possible for clinical and administrative data related to healthcare to be securely accessible to those who require it and to those who have the right to do so on behalf of a patient receiving care. Collaboration is used by the standards-setting body HL7® (Health Level Seven®3) to create and improve FHIR. The creation of FHIR started in 2012 in response to consumer demands for quicker, simpler, and more effective ways to share the vastly increasing volume of health data. The requirement for physicians and consumers to be able to share data in a lightweight, real-time manner using contemporary internet technology and standards was brought on by the growth in the availability of new health data and the developing "app" economy.

FHIR is built on internet standards that are often used in sectors other than healthcare. These in especially include the REST technique, which explains how discrete packets of data (referred to as Resources) can be easily transferred. FHIR considerably lowers the entry barriers for new software developers to address healthcare needs by adopting current standards and technologies that are already familiar to software developers. The FHIR standard also offers software developers the following benefits: A heavy emphasis on quick and simple implementation; developers have noted that they were able to implement straightforward interfaces in just one day.

Free to use without any limitations. Assistance from well-known companies including Apple, Microsoft, Google, Epic, Cerner, and the majority of EHR manufacturers. There are lots of free downloads, online resources, and implementation libraries, as well as reference servers. Numerous publicly accessible examples are available to jump-start the creation of new apps. Out-of-the-box interoperability: base resources can be utilized as is or modified to meet local needs (the process of Profiling). An evolutionary development path from Version 2 and Clinical Document Architecture (CDA®) [333], prior HL7 healthcare standards, allowing them to coexist and benefit from one another. A solid base of XML, JSON, HTTP, and OAuth web standards. Clear and understandable online specifications. A serialization format that is easy for humans to read for developers. A worldwide network to support implementers.

4.4.5. Types of medical index terminologies

4.4.5.1. ICD-10 and ICD-11

The ICD-11's structure [142] is more complex than the ICD-10's. The ICD-11 provides a high degree of detail in the coding of these conditions, with about 55,000 codes that can be used to describe diseases, syndromes, injuries, and reasons of death. On June 18, 2018, a draft version of the new ICD was made available. Beginning on January 1, 2022, member nations will adopt it as the official reporting method. It was formally introduced at the World Health Assembly in May 2019. The ICD-11 provides translations into 43 different languages as well as instructions for using it with various cultural contexts. By offering a universal coding language that can be utilized by researchers and healthcare practitioners everywhere, the redesigned method facilitates utilization and international comparisons.

4.4.5.2. SNOMED-CT

SNOMED CT [155] is among a group of standards that have been designated for use in U.S. Federal Government systems for the electronic exchange of clinical health information. It is also a standard that is required by the interoperability requirements of the U.S. Healthcare Information Technology Standards Panel. The SNOMED International owns and maintains the medical jargon. The National Library of Medicine (NLM) offers SNOMED CT information and resources to NLM UMLS Metathesaurus licenses in its capacity as the United States National Release Center for SNOMED CT. An EHR can employ the structured clinical vocabulary known as SNOMED CT which is the world's most complete and accurate clinical health terminology package. To ensure that data is captured consistently and accurately across the NHS, SNOMED CT is used as a vocabulary for patient clinical information. This makes it easier to transfer clinical data between systems. Without requiring a care worker to manually enter the data again, clinical information from a discharge summary, for instance, can be instantly added to a patient record, saving time, and reducing human error.

4.4.5.3. ATC

The active substances are categorized using the Anatomical Therapeutic Chemical (ATC) classification system [334] based on the organ or system they affect as well as

their therapeutic, pharmacological, and chemical characteristics. There are five main levels of classification for drugs. When pharmacological subgroups are thought to be more appropriate than therapeutic or chemical subgroups, the second, third, and fourth levels are frequently utilized to identify them. The International Nonproprietary Name (INN) is preferred for the chemical compound. Usually, USAN (United States Adopted Name) or BAN (British Approved Name) names are selected if INN names are not assigned. For epidemiological research to yield accurate results, coding is crucial. According to the study's objectives, comparisons can be made at various levels thanks to the five different levels.

- **ATC 1st level:** The system has 14 main anatomical or pharmacological groups.
- **ATC 2nd level:** Pharmacological or Therapeutic subgroup.
- **ATC 3rd and 4th levels**: Chemical, Pharmacological or Therapeutic subgroup.
- **ATC 5th level**: Chemical substance.

4.4.5.4. LOINC

The common denominator used globally to identify health measures, observations, and records [335]. To identify health measurements, observations, and documents, LOINC is a common language (IDs, names, and codes). If you consider an observation to be a "question" and the value of the observation as the "answer," Today, most clinical and laboratory systems use the HL7 version 2 messaging standard to transmit data. You can see how a SNOMED CT code reflects the response and a LOINC code identifies the query by looking at an example of where the test results appear in an HL7 message.

4.4.5.5. OHDSI Athena

For all instances of an OMOP CDM [336], there is a web application for distributing and reading the Standardized Vocabularies. Different observational databases can be systematically analyzed using the OMOP Common Data Model. The idea behind this method is to convert the data present in those databases into a standard format (data model) as well as a standard representation (terminologies, vocabularies, coding schemes), and then perform systematic analyses using a library of common analytic routines that have been created based on the standard format. A unique solution is provided by the Observational Medical Outcomes Partnership (OMOP) CDM, which is now at version 6.0. According to OMOP, diverse coding systems can be harmonized

to a common vocabulary with little information loss. Utilizing standardized analytics tools, evidence can be produced after a database has been transformed to the OMOP CDM. There are other sources of such tools, some of them commercial, but we at OHDSI are currently developing Open-Source tools for data quality and characterization, medical product safety surveillance, comparative effectiveness, quality of care, and patient-level predictive modeling.

4.5. A hybrid method for data harmonization

4.5.1. Overview

The data harmonization workflow is depicted in Figure 27 and consists of 4 stages, including the: (i) metadata extraction and relational modeling, (ii) construction of reference ontologies for the CVD and mental disorders, (iii) development of medical dictionaries by interlinking the word embeddings from the ontologies with external knowledge repositories, such as, the OHDSI (Observational Health Data Sciences and Informatics) [337], and (iv) lexical and semantic analysis. The latter are built on top of the dictionaries to identify terminologies with lexical and conceptual basis. The output is a data harmonization report which includes the matching scores for each identified terminology along with useful metadata.

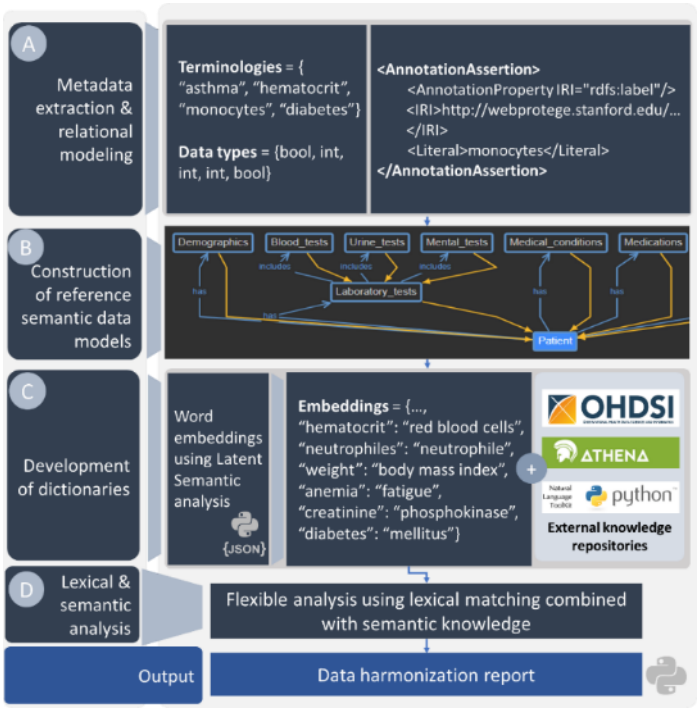


Figure 27. An illustration of the proposed hybrid data harmonization workflow.

4.5.2. Medical corpus definition and interlinking with external medical index repositories

Metadata collection is a crucial step prior to the definition of the reference semantic data models and involves the construction of a reference data template including: (i) the terminologies in the input data, (ii) the value ranges, and (iii) any relational information. The template is transformed into an ontology which consists of classes, subclasses and object properties that hierarchically represent the relationships among the terminologies. Object properties, such as, "include", and "has" are used to represent the relationships. Each ontology is expressed into .RDF (Resource Description Framework) format for compatibility with international knowledge bases, such as, the ICD-11 [1], and SNOMED-CT [155].

The terminologies of each ontology are first extracted and then are enriched with medical terminologies from external knowledge bases to define a dictionary, assume J, in the form:

$$J = \{t, c, r\}, \tag{4.4}$$

where t is the list of the extracted terminologies from the reference ontology, c is the list of the classes where the terminologies in t belong to, and r is the list with the value ranges of these terminologies. Each domain's dictionary is enriched with ICD-10 and SNOMED CT related terminologies from the OHDSI [337] using the Natural Language Processing Toolkit (NLTK) [338]. Latent Semantic Analysis (LSA) [339], [340] was also used to enrich each dictionary through the extraction of word embeddings, i.e., vector space model representations for each terminology. LSA learns latent topics by performing a matrix decomposition on a matrix with terminologies using Singular Value Decomposition (SVD) [341].

4.5.3. Latent Semantic Analysis (LSA)

Latent Semantic Analysis (LSA) [339], [340] was used to extract vector space model representations of terminologies. LSA learns latent topics by performing a matrix decomposition on the document-term matrix using Singular value decomposition (SVD). Given a matrix, say $X \in R^{mxn}$, where the i-th row, t_i, is a vector that represents a term and the i-th column, d_i, represents a document, the SVD is defined as:

$$X = U\Sigma V^T, \tag{4.5}$$

where U is a left singular matrix, S is a diagonal matrix with the eigenvalues, and V is a right singular matrix. SVD can be used to quantify the relation between two documents, say d_i and d_j, and/or two terminologies, say t_i and t_j by comparing Σd_i and Σd_j and/or Σt_i and Σt_j based on a distance measure, such as, the cosine or the Euclidean distance.

4.5.4. Lexical and semantic matching

The idea behind lexical analysis lies on the fact the higher the similarity score between two terminologies, the higher their overlap in terms of lexical relevance. Towards this direction, string similarity metrics, including the Jaro distance [52], [305], [320], and the Levenshtein distance [52], [342], were used to extract common terminologies among the input data and the dictionaries. The most common method for lexical matching is to calculate the edit distance between two terminologies, assume x and y, as in (4.1). In fact, the edit distance aims to transform x into y by performing three types of operations, namely: (i) insertion, (ii) deletion, and (iii) substitution. Here, the Levenshtein distance is used to measure the similarity among x and y, in terms of the number of operations needed to transform x into y, as in (4.2). The Jaro distance, $J_{x,y}(a, b)$, is also used to provide a more explainable similarity score as in (4.3). The Jaro-Winkler distance is used as a strict modification of (2) based on a scale factor which assigns higher weights to terminologies with common prefix [320].

A fundamental issue in lexical analysis is the fact that the lexical matches of an input terminology might be more than one. To deal with this, we utilize semantic analysis to extract the most prominent match by taking into consideration the semantic overlap of the related object properties. For example, if a laboratory-related terminology, assume x, is lexically similar with terminologies y_1 and y_2, where y_1 is blood test-related and y_2 is demographic-related, then y_1 will be the best match. A pseudocode of the harmonization process is presented in Algorithm 2. The object properties are first extracted from the reference ontologies along with the classes and subclasses. Word embeddings are computed for the terminologies and entities of the ontologies and fused into a large dictionary which is interlinked with terms from the OHDSI [337]. Lexical

and semantic analysis is recursively applied to identify the most prominent matches for the given terminologies.

Algorithm 2. A pseudocode of the hybrid data harmonizer.

```
def semantic_matching(ontology, terms):
    obj_prop = ontology.object_properties;
    var = [obj_prop[i] for i, x in obj_prop if x == “include”];
    class = [obj_prop[i] for i, x in obj_prop if x == “has”];
    we = embedding_analysis(terms,var,class);
    terms = dictionary_fusion(OHDSI, we);
    for (i,j) in zip(terms, var):
        if (match(i, var[i], emb) == 1)):
            matched_vars[k] = list([var_A[i],]);
    return matched_vars;
end
```

4.6. Summary

The lack of standardized data (data heterogeneity) among biobanks, cohorts, and other sources of medical data is a crucial scientific challenge which obscures the effective analysis of such data, yielding clinical studies with poor statistical power [52], [73], [74], [109], [129], [147], [160], [161]. Data harmonization is an emerging research field which aims to overcome the existing heterogeneities across multiple medical data sources. The rationale of data harmonization is to transform the heterogeneous data into a standardized format with common parameters and range values to enable the integrative analysis of such data and thus enhance the statistical power of the outcomes.

Hence, data harmonization can deal with the clinical unmet needs in various diseases. Although notable data harmonization approaches have been proposed, they mainly focus on prospective data based on the definition of standardized data collection procedures. The most challenging field in data harmonization though is retrospective data harmonization, where standardized protocols are absent. Two main types of harmonization exist: the “stringent” and the “flexible”, where the former one is based on the harmonization of data which have been already collected in standardized formats

whereas the latter one expands the harmonization concept to analyze heterogeneous data which have been collected without common data collection protocols.

Medical data harmonization overcomes the structural heterogeneities through the identification of lexically or conceptually similar terminologies between two or more heterogeneous clinical datasets. This is most commonly achieved by lexically and/or semantically matching the terms of the heterogeneous datasets using a reference model which serves as a gold standard. The reference model is defined as a set of terminologies which describe the domain knowledge of a disease of interest and is usually expressed in the form of an ontology using classes, sub-classes and object properties describing the relationship between the terms. This set of terms is usually defined by the clinical experts in the field and includes various clinical parameters which are related to laboratory tests, biopsies, treatments, etc. The terminology matching process, however, is not always enough since the values of the matched terminologies need to be transformed according to the pre-defined range values in the reference model.

Towards this direction, semi-automated data harmonization frameworks have been proposed to co-analyze heterogeneous data in biobanks and other related registries, including the DataSHaPER [141], [147] which was used to harmonize 53 databases on epidemiology with 36% compatibility, the SORTA tool [142] which was used to match 5120 entries in a single biobank with 97% recall, and the BiobankConnect software [143] which was used to harmonize data across 6 biobanks with 74% precision. In [318] an open-source editor was also developed to provide standardized HL7 data formats and in [319] a semi-automated lexical matcher was used to map 78.48% of 1,492 biobank terms. Statistical approaches, such as, multiparameter logistic Item Response Theory (IRT) analysis were deployed in [160]–[162] to examine how a set of items (e.g., psychiatric phenotypes) is affected by other factors for scale homogenization.

A fundamental drawback of these methods, however, lies on the fact that they are not easily generalizable to other clinical domains. Moreover, they adopt a semi-automated strategy which is based on the extensive collaboration between the clinical and the technical experts to define pairing rules, i.e., a pre-defined set of rules for lexical matching. The performance of the existing frameworks is low in several cases due to the complexity of the clinical field under investigation, such as, in [141]–[143], [145], [147] but also due to the lack of computational methods in [141], [142], [145], [147],

[318], [319] to enable automated terminology matching. To address these needs, we proposed an automated data harmonization workflow which adopts a hybrid strategy that combines lexical analysis with ontologies to identify terminologies with lexical and conceptual overlap.

Data harmonization differs a lot from simply putting the data together. During manual data integration, several variables must be removed due to the differences and incompatibilities in the measurement units, the different type of data collection protocols, etc. As a result, there is a limited exploitation due to the small subset of original data which limits the potential of new scientific discoveries. In addition, the fact that the data integration process is manual along with the nomenclature impose a high risk for mistakes and obscure the definition of the data sharing principles since data integration requires direct data access. On the other hand, data harmonization involves processes that transform the variables into compatible ones to make them compatible. The legal and ethical issues are well defined in advance through data governance mechanisms. Data harmonization limits the direct access to the data since only the data schema is required to transform the variables. Moreover, the fact that the data harmonization procedure is semiautomated in most of the existing tools and frameworks reduces the risk for manual mistakes. Besides, the fact that it uses interoperable data schemas can overcome the nomenclature factor during the analysis.

As we have already mentioned, data harmonization can address the unmet needs in chronic and rare diseases by enabling the interlinking and subsequent co-analysis of heterogeneous cohort data. An example of a promising initiative that deals with the harmonization of longitudinal cohorts of patients with chronic and rare diseases is the HarmonicSS project [343]. HarmonicSS envisages to harmonize and coanalyze more than 20 longitudinal cohorts of patients that have been diagnosed with a rare autoimmune disease known as primary Sjögren's Syndrome (pSS). In short, pSS is a chronic autoimmune disease causing severe salivary gland dysfunction yielding clinical manifestations which vary from dry eyes and dry mouth to severe rheumatoid disorders and lymphoma development [241]. We adopted a semi-automated semantic matching approach to align heterogeneous pSS-related terms with the terms of a pSS reference model which was developed in co-operation with the clinical experts of the project.

The pSS reference model consists of a set of parameters that efficiently describe the pSS domain knowledge including seven classes (i.e., laboratory tests, medical conditions, demographics, lifestyle, SS disease activity indices, and interventions), where each class includes additional subclasses, e.g., the class laboratory tests includes blood tests, oral tests, ocular tests, urine tests, biopsies, and even further subclasses, e.g., the subclass blood tests consists of the lipid tests, hematological tests, serum protein tests, complement tests, etc [127]. The semantic matching process is applied through a user-friendly interface, where the clinical and the technical experts can align terms that share similar concepts by defining mapping scenarios (e.g., a Lab-test outcome yes/no scenario) along with the related value mappings (e.g., set "0" to "no" and "1" to "yes") and evaluate suggested terminology mappings and finally extract the mapping rules. The terminologies in the pSS reference ontology are FHIR compliant using ICD-10 and SNOMED-CT terminologies.

With the majority of the existing data harmonization tools and frameworks being semiautomated, emphasis must be given on the development of new strategies to eliminate the "semi-" term. A promising solution would be to create a repository with a collection of biomedical ontologies that lie under a specific medical domain. This would greatly increase the interoperability of the harmonization process since the available core set of terms would cover a much larger portion of the domain. Another idea that is more straightforward would be to enrich this repository with information regarding the mapping of heterogeneous data schemas with the ontologies to introduce the "smart" repositories. A fundamental objective of the "smart" repository is the fact that it could be used to train a proper machine learning algorithm that could learn from the existing knowledge to automatically align the data schema of an upcoming heterogeneous dataset. This would greatly reduce the time effort needed for semiautomated harmonization and enhance the applicability of such an approach across different domains by including the mapping information of the related ontologies.

There is no doubt that data harmonization has a leading role in the co-analysis of heterogeneous medical data. Harmonization is the "key" factor that can enable the integrative analysis of data from heterogeneous data sources and thus envisages to make the sharing of data meaningful by distilling the power of data sharing into the construction of a set of interoperable and homogeneous data schemas that can be used

to deal with the unmet needs in rare and chronic diseases, as well as, in biobanks, omics registries, and cohorts. The success of data harmonization towards this direction has been proven by the existence of several initiatives which have demonstrated promising results towards the harmonization of biobanks, electronic health records, and cohorts, in various medical domains worldwide. Practices and actions need to be taken from healthcare stakeholders to invoke the inclusion of biomedical ontologies from rare diseases, as well as, the update of the existing ones in the international repositories to enable the development of data harmonization tools that will be able to make the interlinking of clinical centers worthy for the public.

CHAPTER 5. SYNTHETIC DATA GENERATION AND DATA AUGMENTATION

5.1. Overview

Virtual population generation has gained a lot of attention in the healthcare sector due to the overwhelming need to overcome the significant lack of sufficient population size, particularly for in silico clinical trials (ISCTs), where the financial burden of expensive drugs leverages the orchestration of viable Phase II/III CTs by pharmaceutical companies worldwide [115], [344], [345]. Furthermore, the lack of medical databases with increased statistical power (e.g., in rare diseases) obscures the deployment of machine learning pipelines that can identify risk factors for disease progression and treatment due to the reduced amount of available training data. As a matter of fact, all these factors have a significant negative impact in the capacity of the existing healthcare systems, where the costs and delays for treatment and re-admission are already high. Virtual population generation envisages to address these needs through the development of virtual data generators which are trained on the real data to produce virtual (or synthetic) distributions which can "mimic" the real ones in terms of reduced divergence and dispersion with the real data. Since the virtual data quality is directly

affected by the quality of the real data, it is first necessary to enhance the quality of the real data including the data completeness and conformity.

The current advances in data science have led to the development of an emerging branch of applications which focuses on the augmentation of medical data. Its aim is to shed light into the underlying structure of clinical problems towards the development of robust machine learning models for predicting disease outcomes and their risk levels. Virtual population generation [115], [130], [131], [165], [344]–[346] refers to the development of computational methods that can be used to generate artificial (or synthetic) patient data by producing virtual distributions like those in the real world. Its desired usage is to enhance the statistical power of clinical research databases with significant lack of population size. Data augmentation [134], [347] refers to the aggregation of the real with the virtual patient data to yield AI (artificial intelligence) models with increased performance for classification tasks. For example, in medical imaging, data augmentation refers to the application of data mirroring and data cropping methods to enhance the performance of the existing deep learning models for image segmentation by increasing the size of the input training data with virtual training data. Clinical data augmentation refers to the aggregation of the real with the high-quality virtual clinical data to address various clinical unmet needs including the development of robust risk stratification and disease classification models, as well as, the detection of biomarkers, among others.

As a result, the clinical value of data augmentation lies on the quality of the virtually generated data. Indeed, the aggregation of the real data with poor quality virtual data (i.e., virtual data with increased divergence and reduced similarity with the real data) is expected to have a negative impact on the performance of the AI models. As a matter of fact, particular emphasis must be given on the development of robust virtual population generators. In addition, prior the development of the virtual population generators it is crucial to apply data curation methods towards the detection and removal of data recording errors, inconsistent data types and problematic fields that are present in the input clinical data. This step is important since the application of the virtual population generators on contaminated data might produce virtual data with poor performance and reduced statistical power of downstream applications. Thus, data

curation must be applied to meet standard data quality criteria in terms of data completeness and conformity [52], [66]–[69], [106], among others.

Several studies have been launched towards the design of efficient synthetic medical data generators based on both probabilistic and machine learning approaches as already described in Section 2.2.3. In [164], [167] the multivariate normal distribution (MVND) was applied to generate virtual data given the mean vector and the covariance matrix of the real data. Bayesian networks (BN) have been also used in [166], [168]–[171], [348] for the generation of virtual distributions based on the modeling of conditional probabilities across diverse network topologies. The BN and the MVND, however, suffer from mathematical assumptions; the MVND algorithm assumes that the real data are normally distributed whereas in the BNs the conditional probabilities are modeled using assumptions on the prior distribution of the features, where the network topology is not pre-defined. Towards this direction, machine learning based generators have been applied in several studies [131], [172], [173], such as, the artificial neural networks (ANNs) with radial basis functions, the supervised tree ensembles (STE), the unsupervised tree ensembles (UTE), and yielding favorable performance against the probabilistic approaches. However, they are not computationally efficient since they require increased training and testing time. In addition, the STE, and the ANN require a target feature which influences the associations of the virtual features and introduces biases in the generated data. Moreover, in the case of Bayesian networks, the number of all possible permutations of edges in one topology is infinite.

5.2. Beyond the state of the art

The emerging need for the development of computationally efficient and unbiased synthetic data generators remains a technical challenge, particularly in the case of large-scale clinical trials, where the computational complexity is important. A computationally efficient data generator has been introduced in [174], [176], where Gaussian Mixture Models (GMMs) were used to generate virtual data based on Dirichlet processes. Since GMM maximizes only the likelihood based on the expectation maximization (EM) approach, it might yield specific structures that might or might not apply to the data. A solution to this is to use variational inference (VI) as in [130], [177], [178] which maximizes a lower bound on the model evidence instead of the data likelihood like in the EM to reduce the computational complexity compared

against the MVND, BN, UTE, STE, and ANN. However, none of these studies has investigated the optimal selection of the number of Gaussian components for the model training process, where the number of Gaussian components is arbitrary. In addition, the number of Gaussian components has a direct effect on the estimation of the weight concentration (or gamma) parameter which is the most important hyperparameter of the BGMM since it affects the log-likelihood of the model. In this thesis, we focused on the optimal estimation of the Gaussian components in the BGMM algorithm to yield concrete estimations of the VI at reduced computational complexity for large-scale synthetic data generation (we refer to this approach as BGMM with Optimal Components Estimation: BGMM-OCE) [133]. To do so, we first apply spectral clustering based on the Locally Optimal Block Preconditioned Conjugate Gradient (LOBPCG) method to identify the best clustering solution as the one with the highest Davies Bouldin score (DBS) at small complexity. Then, we set the optimal number of clusters as the number of Gaussian components, and we define an exponentially decaying gamma value. The BGMM-OCE's performance was compared against state-of-the-art synthetic data generators (BN, UTE, STE, ANNs) in the context of in silico clinical trials for HCM.

In addition, none of the above virtual population/synthetic data generation studies has investigated the effectiveness of clinical data augmentation in terms of not only enhancing the size of the real patient data but also aggregating the virtually generated patient data with the real data to enhance the performance of disease classification and risk stratification models. In this thesis, we proposed a fully automated, highly scalable, data augmentation pipeline to enrich patient databases with insufficient population size and statistical power [134]. Data augmentation yields hybrid and robust ML models for risk stratification and disease prediction. To this end, we deployed five state-of-the art virtual data generation methods to produce high-quality virtual patient data for 1,000 patients with an increased level of similarity to the real patients across two clinical domains; the primary Sjögren's Syndrome (pSS) and the hypertrophic cardiomyopathy (HCM). The number of virtually generated patients is relatively large for both clinical domains especially in pSS considering that it is a rare systemic autoimmune disease. The novelty of the proposed computational pipeline lies on the fact that it: (i) enhances the quality of the input clinical data through the precise detection and elimination of outliers and data inconsistencies using data curation workflows, (ii) augments the

curated clinical data with high-quality virtual data that enhance the population size of two rare clinical research databases through the development of high-performance virtual data generators, including both supervised and unsupervised tree ensembles, as well as, artificial neural networks (ANNs) with Gaussian kernels, which are extended to resolve overfitting effects during the generation stage, and (ii) builds supervised machine learning models on the aggregated real and virtual data for the robust classification of lymphoma pSS patients and for the risk stratification in HCM.

5.3. Methods for synthetic data generation

5.3.1. Statistical methods

5.3.1.1. Multivariate normal distribution (MVND)

Given a univariate feature, $\boldsymbol{X} \in R^{pxn}$, the multivariate normal distribution (MVND) can be defined as an extension of the normal distribution as in:

$$f(\boldsymbol{X}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-(\boldsymbol{X}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})/2}, \quad (5.1)$$

where p is the dimension, $\boldsymbol{\mu}$ is the mean vector of $\boldsymbol{X}$, $\boldsymbol{\Sigma}$ is the covariance matrix of $\boldsymbol{X}$, and $\boldsymbol{\Sigma}^{-1}$ is the pseudoinverse of $\boldsymbol{\Sigma}$. A multi-dimensional normal distribution is constructed from the mean vector and the covariance matrix of the input data.

5.3.1.2. Multivariate log-normal distribution (MVND)

To ease the assumption of normality within the data, the log-normal distribution is defined, where the logarithm of the exponential term in (4) fulfills the condition:

$$ln\left(\boldsymbol{e}^{\boldsymbol{f}(\boldsymbol{x})}\right) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \quad (5.2)$$

5.3.2. Machine learning methods

5.3.2.1. Supervised tree ensembles (STE)

A more advanced approach to virtual population generation is to train a tree ensemble [172], [173] for a given set of training features and a target feature. During the training phase of the generator, we build an ensemble similar to random forests with some additional data needed for data generation phase [172], [173]. In each interior tree node,

we store the generator for the splitting feature based on its univariate empirical cumulative distribution function (ECDF). In each leaf node, we store ECDF-based generators for all variables not encountered on a path from the root to that leaf. To avoid overfitting effects which are introduced in the training process during the construction of the tree ensemble we introduce a new approach according to which one of the trees from the ensemble is randomly chosen when producing a new instance which is passed down the tree starting in the root node. The ensemble, as the generator, approximates the probability density function of the regions where features are assumed to be independent (the dependencies are likely to be resolved on the path from the root to the leaves). As the ensemble contains a sufficient number of different trees, the probability density function of the original feature space is reasonable well approximated with the generated instances. During the training process, the Gini impurity index [52] is used to measure the probability of a variable, I, being classified in the wrong class:

$$I = 1 - \sum_{i=1}^{n} {p_i}^2, \tag{5.3}$$

where p_i is the probability of a sample falling in class $i \in \{1,2,\dots,k\}$, and k is the number of classes.

5.3.2.2. Unsupervised tree ensembles (UTE)

The unsupervised tree ensemble generator is built in a similar way as the supervised tree ensemble, but instead of random forests ensemble, this generator builds a density forest ensemble [131], [173]. Here, the ensemble members are density trees built with a similar top-down manner as decision trees but using the variance of the features as the criterion for selection of the splitting feature. To avoid overfitting effects which are introduced during the construction of the density forest ensemble, each density tree in the ensemble is randomly selected as the one with the smallest convergence rate. Any information regarding the target feature is not necessary. Other components are identical to the supervised tree ensemble both in the learning and generation phase.

Density forest ensembles were used as high-quality virtual data generators [131], [173] instead of the conventional probabilistic methods which are restricted to oversampling with biased assumptions. Density trees are built in a top-down way, where the splitting

process is based on the variance of each feature. A density forest is as a mixture of Gaussian densities [131], [173]:

$$p(v) = \frac{1}{M}\sum_{k=1}^{M} p_k(v) = \frac{1}{M}\sum_{k,q} g_q(v) N\left(v; \mu_q(v), \Sigma_q(v)\right), \tag{5.4}$$

where $v \in V$ is a tree node, $N(v; \mu_q(v), \Sigma_q(v))$ is a multivariate Gaussian distribution with mean $\mu_q(v)$ equal to the mean of all points reaching the leaf $q \in Q$, $\Sigma_q(v)$ is the covariance and $g_q(v)$ is the proportion of all points reaching q.

5.3.2.3. RBF-based ANNs

Robnik-Šikonja [1] has proposed an approach for virtual population generation with artificial neural networks (ANNs), that uses radial base functions (RBFs) as activation functions. The RBF-based ANN's output is defined as in:

$$y(\boldsymbol{q}) = \sum_{i=1}^{N} w_i \, exp\left(-\beta \left|\left|\boldsymbol{q} - \boldsymbol{q_i}\right|\right|^2\right), \tag{5.5}$$

where $y(\boldsymbol{q})$ is the output of the ANN, w_i is the weight of the i-th neuron, $\boldsymbol{q_i}$ is the center vector of the i-th neuron, $||\boldsymbol{q} - \boldsymbol{q_i}||$ is the distance of each sample in $\boldsymbol{q}$ from the center vector $\boldsymbol{q_i}$ in the i-th neuron, and β is a standard Gaussian parameter. The RBF generator is created with a standard training algorithm which estimates the Gaussian parameters in the neurons. In the generation phase, the RBF generator uses Gaussian kernels as multivariate generators to deal with overfitting effects and produce new instances from each one in proportion to their presence in the training set.

5.3.2.4. Bayesian networks (BN)

Bayesian networks [133], [166], [168], [169] are based on the definition of a directed acyclic graph (DAG), say $\boldsymbol{D} = (\boldsymbol{V}, \boldsymbol{E})$, where $\boldsymbol{V}$ is a set of nodes and $\boldsymbol{E}$ is a set of directed edges between the nodes in $\boldsymbol{V}$. Each node $v \in \boldsymbol{V}$ is assigned to a random variable, say x_v, with parents, say $x_{pa(v)}$, with a probability distribution:

$$p_v = p\left(x_v \middle| x_{pa(v)}\right). \tag{5.6}$$

Assuming conditional independencies among the random variables, (1) can be re-written as:

$$p_v = \prod_{c \in \boldsymbol{C}} p\left(x_c | x_{pa(c)}\right) \prod_{d \in \boldsymbol{D}} p\left(x_d | x_{pa(c)}, x_{pa(d)}\right). \quad (5.7)$$

where, $p\left(x_d | x_{pa(c)}, x_{pa(d)}\right)$ is the conditional probability of x_c given the parents of both the discrete ($x_{pa(c)}$, set $\boldsymbol{C}$) and the continuous ($x_{pa(d)}$, set $\boldsymbol{D}$) variables, and $p\left(x_c | x_{pa(c)}\right)$ is the conditional probability of x_c given $x_{pa(c)}$. The DAG structure is used to generate new instances consistent with causal dependencies between the features. If the node is discrete, the probability distribution in (1) is uniform, otherwise a mean and a variance is attached per discrete parent configuration.

5.3.3. Probabilistic methods

5.3.3.1. Gaussian Mixture Models (GMM)

A Gaussian mixture model (GMM) is a probabilistic model which assumes that the samples are generated from a mixture of a finite number of Gaussian distributions with unknown parameters [130], [133], [174]. A GMM approximation is defined as:

$$q(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{i=1}^{N} q(i; \boldsymbol{\theta}) q(\boldsymbol{x} | i; \boldsymbol{\theta}), \quad (5.8)$$

where i is the mixture component, $\boldsymbol{\theta}$ is the set of hyper-parameters, $q(i; \boldsymbol{\theta})$ are the mixture weights, and $q(\boldsymbol{x} | i; \boldsymbol{\theta})$ is a multivariate normal distribution (MVND) with mean $\boldsymbol{\mu_o}$ and covariance matrix $\boldsymbol{\Sigma_o}$, $N(\boldsymbol{x} | \boldsymbol{\mu_o}, \boldsymbol{\Sigma_o})$. A common approach for estimating $\boldsymbol{\theta}$ is based on the expectation-maximization algorithm which maximizes the data likelihood.

5.3.3.2. Bayesian Gaussian Mixture Models (BGMM)

The EM, however, might yield GMMs with topologies that might not fit well to the underlying data structures. A solution to this is provided by variational inference (VI), which seeks for a lower bound on the model evidence instead of the likelihood. The goal of the GMM with variational Bayesian inference (BGMM) is to estimate the hyper-

parameter(s) $\boldsymbol{\theta}$ in $q(\boldsymbol{x};\boldsymbol{\theta})$, so that the Kullback-Leibler (KL) divergence with the posterior distribution $p(\boldsymbol{x})$ is minimized.

The KL-divergence [349] is defined as:

$$KL\big(q(\boldsymbol{x};\boldsymbol{\theta})||p(\boldsymbol{x})\big) = \int_{\boldsymbol{x}} q(\boldsymbol{x};\boldsymbol{\theta})\, log\left(\frac{q(\boldsymbol{x};\boldsymbol{\theta})}{p(\boldsymbol{x})}\right) d\boldsymbol{x}, \tag{5.9}$$

where the quotient of the search model over the posterior is the logarithm of the evidence, $L(\boldsymbol{\theta})$. Minimizing (5.9) is the same as maximizing a lower bound on $L(\boldsymbol{\theta})$:

$$argmax_{\theta}\left[\int_{\boldsymbol{x}} q(\boldsymbol{x};\boldsymbol{\theta})\big(log\big(p(\boldsymbol{x})\big) - log(q(\boldsymbol{x};\boldsymbol{\theta}))\big)d\boldsymbol{x}\right], \tag{5.10}$$

which refers to as the Evidence Lower Bound Objective (ELBO) [350]. In the case of GMM, where the search model is a multivariate normal distribution, (5.10) becomes:

$$argmax_{\theta}\left[\int_{\boldsymbol{x}} q(\boldsymbol{x};\boldsymbol{\theta})\big(R(\boldsymbol{x}) + log\big(\tilde{q}(i|\boldsymbol{x})\big)\big)d\boldsymbol{x} + H(q)\right], \tag{5.11}$$

where $R(\boldsymbol{x})$ is equal to $log(p(x))$ and $H(q) = H\big(q(x|i)\big) = -\int_{\boldsymbol{x}} q(\boldsymbol{x};\boldsymbol{\theta})\, log\big(q(\boldsymbol{x};\boldsymbol{\theta})\big)d\boldsymbol{x}$ is the entropy of $q(\boldsymbol{x};\boldsymbol{\theta})$.

5.3.3.3. BGMM with optimal component estimation (BGMM-OCE)

5.4. Robust initialization of Gaussian components

5.4.1. Fast estimation of the eigenvalues and the eigenvectors though the LOBPCG approach

A scaling approach robust to "hidden" outliers was applied to standardize the input data, where the scaling and centering process was conducted independently for each feature according to the median and the interquartile range. In this work, the eigensolver is based on the Locally Optimal Block Preconditioned Conjugate Gradient Method (LOBPCG) which is ideal for large symmetric positive definite (SPD) generalized eigenproblems [351], [352]. Given the set of n-input features, assume $\boldsymbol{x}$, spectral clustering was applied to project the original data into a different dimensional space, where the separation is easier. This is done by first computing the affinity matrix of the

original data, say $\boldsymbol{Q} \in R^{nxn}$, where the element q_{ij} is the similarity between the input features i and j, and then transforming the affinity matrix into its Laplacian form. Once the affinity matrix is constructed, its Laplacian matrix is defined [353]:

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{Q}, \tag{5.12}$$

where $\boldsymbol{D}$ is an nxn diagonal matrix. The next step is to apply eigenvalue decomposition on $\boldsymbol{L}$. The Laplacian matrix is expressed in the form of a symmetric-definite pencil $(\boldsymbol{A}, \boldsymbol{B})$, where $\boldsymbol{A}$ is an Hermitian matrix and $\boldsymbol{B}$ is Hermitian positive definite with eigenvectors, say $U = \{\boldsymbol{u_1}, \boldsymbol{u_2}, \dots, \boldsymbol{u_k}\}$ and eigenvalues, say $\boldsymbol{P} = diag\{p_1, p_2, \dots, p_k\}$. The eigenvectors $\boldsymbol{U}$ can be estimated by solving the trace-minimization objective [133], [351], [352]:

$$min_{\boldsymbol{U^T BU=I}}\, trace(\boldsymbol{U^T AU}). \tag{5.13}$$

According to the LOBPCG algorithm the eigen vector at step i, $\boldsymbol{U}^{(i)}$, is defined as a block $[W^{(i)}, Q^{(i)}]$, where $W^{(i)}, Q^{(i)}$ are the preconditioned gradient of the Lagrangian and the aggregated update direction of the previous searches [133], [351], [352]:

$$\boldsymbol{W}^{(i)} = \boldsymbol{K}^{-1}\left(\boldsymbol{AU}^{(i)} - \boldsymbol{BU}^{(i)}\boldsymbol{U}^{(i)^T}\boldsymbol{AU}^{(i)}\right), \tag{5.14}$$

and

$$\boldsymbol{Q}^{(i+1)} = \boldsymbol{U}_{\perp}{}^{(i)}\boldsymbol{C}^{(i+1)}, \tag{5.15}$$

with $\boldsymbol{C}$ denoting the coefficient matrix. Once the matrix $\boldsymbol{U}$ is estimated, the k-means algorithm is applied on the k-largest eigenvectors, say $\boldsymbol{u_1}, \boldsymbol{u_2}, \dots, \boldsymbol{u_k}$ of $\boldsymbol{U}$, to assign each row say $\boldsymbol{u_1}^{\boldsymbol{k}}, \boldsymbol{u_2}^{\boldsymbol{k}}, \dots, \boldsymbol{u_m}^{\boldsymbol{k}}$ into a cluster C_j, $j = 2, \dots, k$.

5.4.2. Clustering evaluation based on the DB score (DBS)

For a given clustering set C, the Davies-Bouldin score (DBS) [354], is defined as:

$$DBS(C) = \frac{1}{k}\sum_{i=2}^{k} max\left(R_{ij}\right) = \frac{1}{k}\sum_{i=2}^{k} max\left(\frac{d_i + d_j}{d_{ij}}\right), \tag{5.16}$$

where k is the number of clusters, R_{ij} is the similarity score between cluster C_i and cluster C_j, where $j \neq i$, d_i is the average distance between the observations in cluster

C_i from its centroid, d_j is the average distance between the observations in cluster C_j from its centroid, and d_{ij} is the distance between the centroids in clusters C_i and C_j. DB values close to 1 indicate a good clustering performance since.

5.4.3. Extraction of the optimal number of clusters

The DBS was evaluated on a pre-defined number of clusters, say $2, \dots, k$, and the clustering number that achieves the lowest DB score is the one that yields well-separated clusters in terms of large distance between the clustering centroids.

5.5. BGMM training

5.5.1. Gaussian Mixture Models with Variational Inference

A Gaussian mixture model (GMM) is a probabilistic model which assumes that the samples are generated from a mixture of a finite number of Gaussian distributions with unknown parameters [130], [133], [178]. A GMM approximation is defined as:

$$q(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{i=1}^{N} q(i;\boldsymbol{\theta}) q(\boldsymbol{x}|i;\boldsymbol{\theta}) \tag{5.17}$$

where i is the mixture component, $\boldsymbol{\theta}$ is the set of hyper-parameters, $q(i;\boldsymbol{\theta})$ are the mixture weights, and $q(\boldsymbol{x}|i;\boldsymbol{\theta})$ is a multivariate normal distribution (MVND) with mean $\boldsymbol{\mu_o}$ and covariance matrix $\boldsymbol{\Sigma_o}$, $N(\boldsymbol{x}|\boldsymbol{\mu_o}, \boldsymbol{\Sigma_o})$. A common approach for estimating $\boldsymbol{\theta}$ is based on the expectation-maximization algorithm which maximizes the data likelihood. However, EM might yield GMMs with structural topologies that might not fit to the underlying data structures. A solution to this is provided by variational inference (VI), which seeks for a lower bound on the model evidence instead of the likelihood. In variational inference, the goal of the variational Bayesian Gaussian Mixture Models (BGMM) is to estimate the hyper-parameter(s) $\boldsymbol{\theta}$ in $q(\boldsymbol{x};\boldsymbol{\theta})$, so that its Kullback-Leibler (KL) divergence with the posterior distribution $p(\boldsymbol{x})$ is minimized. In the GMM, the search model is a multivariate normal distribution, and the search function becomes:

$$argmax_{\theta}\left[\int_{\boldsymbol{x}} q(\boldsymbol{x};\boldsymbol{\theta})\big(R(\boldsymbol{x}) + log\big(\tilde{q}(i|\boldsymbol{x})\big)\big)d\boldsymbol{x} + H(q)\right], \tag{5.18}$$

where $R(\boldsymbol{x})$ is equal to $log(p(x))$ and $H(q) = H\big(q(x|i)\big) = -\int_{\boldsymbol{x}} q(\boldsymbol{x};\boldsymbol{\theta})\, log\big(q(\boldsymbol{x};\boldsymbol{\theta})\big)d\boldsymbol{x}$ is the entropy of $q(\boldsymbol{x};\boldsymbol{\theta})$. For the BGMM training process, the prior distribution was set to the Dirichlet processes. The Dirichlet distribution is a multivariate generalization of the beta distribution. More specifically, given a set of multinomial distributions with probability outcomes, say $\boldsymbol{\theta} = \{\theta_1, \theta_2, \dots, \theta_m\}$, and a set of equal-sized parameters, say $\boldsymbol{a} = \{a_1, a_2, \dots, a_m\}$, the Dirichlet distribution is defined as [355]:

$$Dir(\boldsymbol{\theta}|\boldsymbol{\alpha}) \sim \frac{\Gamma(\sum_k \alpha_k)}{\Pi_k \Gamma(\alpha_k)} \prod_{k=1}^{m} {\theta_k}^{a_k - 1}, \tag{5.19}$$

which is considered as a distribution of distributions. The Dirichlet process (DP) is simply a generalization of the Dirichlet distribution, where a process, say H, is said to be a DP with a base distribution G over a probability space $\boldsymbol{\theta}$, a concentration parameter α, say $H \sim DP(\boldsymbol{a}, G)$, given the following condition:

$$\big(H(\theta_1), H(\theta_2), \dots, H(\theta_m)\big) = Dir\big(\boldsymbol{a}G(\theta_1), \boldsymbol{a}G(\theta_2), \dots, \boldsymbol{a}G(\theta_m)\big). \tag{5.20}$$

5.5.2. Weight concentration parameter estimation

The precise definition of the weight concentration parameter is challenging since a low value makes the model put most of the weight on few components whereas large values might lead to poor performance. In practice, the weight concentration parameter is defined as the inverse of the number of components. However, this approach introduces biases since it assumes a linear relationship between them. To deal with this, we use an exponential function during the evaluation on multiple number of components to capture non-linear effects among them, as exp $(-opt)$. Through this way, the BGMM yields better local minima in the variational inference and thus more stable number of components across virtual populations.

5.5.3. Model implementation, training, and random sampling

A pseudocode of the BGMM-OCE algorithm is described in Algorithm 3. The input includes the curated dataset, the number of virtual patients and the initial parameters of the BGMM model. The BGMM-OCE approach applies a sequential spectral clustering process, for a set of k clusters under evaluation, based on the LOBPCG method and

extracts the best clustering solution, i.e., the one having the highest DBS, say k_{DBS}. Then, the BGMM training process is initialized, where the number of Gaussian components is set equal to k_{DBS} and the weight concentration parameter is set equal to $exp\ (-k_{DBS})$. Random sampling is then applied on the trained model based on Dirichlet distributions to yield the virtual samples. Quality evaluation metrics are then applied to compare the real with the virtual distributions.

Algorithm 3. A pseudocode of the BGMM-OCE algorithm.

1	**Input**
2	**X:** curated and transformed dataset;
3	**N:** number of virtual patients;
4	**params:** initial model parameters;
5	def **get_clustering_score**(X, k, params):
6	labels = LOBPCG(k, params).fit(X).labels;
7	DBS = DB(X, labels);
8	return DBS;
10	def **BGMM-OCE**(X, N, k, params):
11	DBS = get_clustering_score(X, 2:k, params[0]);
12	model = train_BGMM(max(DBS), exp(-DBS), params[1]);
13	VP = model.sample(N);
14	eval(VP);
15	return VP;

5.6. Synthetic data quality metrics

5.6.1. Kolmogorov-Smirnov Goodness of fit (gof)

More specifically, the gof test statistics, say D, is given by:

$$D = max\ (|F_o(x) - F_v(x)|), \tag{5.21}$$

where $F_o(x)$ and $F_v(x)$ are the empirical distribution functions (EDFs) of the original and virtual populations, respectively. In fact, the gof measures whether $F_o(x)$ and $F_v(x)$ are similar by calculating the largest distance, D, between the two EDFs. If D is larger than a critical value then the null hypothesis is rejected at the given confidence level. As a matter of fact, a large gof value between $F_o(x)$ and $F_v(x)$ denotes distributions

with large vertical distance and thus the null hypothesis is rejected, whereas small gof values denote similar distributions.

5.6.2. Inter-correlation difference

To calculate the inter-correlation difference, we first estimate the correlation matrix of the real and the virtual data, say $X \in R^{MxN}$ and $V \in R^{M'xN}$, respectively, where M corresponds to the number of real patients, M' to the number of virtual patients, and N to the number of features. To do so, we estimate the Pearson correlation coefficient between each pair of features in matrices X and V resulting to the 3D correlation matrices $C_X \in R^{NxN}$ and $C_V \in R^{NxNxQ}$, respectively, where Q refers to the index number of the virtual patients (e.g., index Q=1 corresponds to the C_V for 1000 virtual patients). Finally, the average correlation matrices are extracted $E[C_X] \in R^{NxN}$ and $E[C_V] \in R^{NxN}$, and averaged over the pairs of features yielding a scalar score representing the inter-correlation difference.

5.6.3. Intra-correlation difference

As far as the intra-correlation difference is concerned, we follow a more complex, patient wise procedure this time, according to which we first estimate the Pearson correlation coefficient between each pair of features in matrices X and V for every individual patient resulting to a 3D real correlation matrix $C'_X \in R^{NxNxM}$ and a 4D virtual correlation matrix $C'_V \in R^{QxNxNxM'}$, where Q refers to the index of the number of virtual patients, M corresponds to the number of real patients, and M' to the number of virtual patients for the corresponding Q index. The average correlation matrices, say $E[C'_X] \in R^{NxN}$ and $E[E[C'_V]] \in R^{NxN}$, are extracted as the average correlation matrix over the possible pairs of features and patients yielding a scalar score representing the intra-correlation difference.

5.6.4. Kullback-Leibler (KL) divergence

For two features, $\boldsymbol{x_r}$ and $\boldsymbol{x_v}$, from the real and virtual data, respectively, with probability distributions, $\boldsymbol{p_{x_r}}$ and $\boldsymbol{p_{x_v}}$ defined on the same probability space, $\boldsymbol{K}$, the Kullback-Leibler (KL) divergence [349] quantifies the divergence between the two distributions, in an asymmetric manner, as in:

$$KL(\boldsymbol{p}_{x_r}||\boldsymbol{p}_{x_v}) = \sum_{\boldsymbol{k}\in K} \boldsymbol{p}_{x_r}(k_i)\, log\left(\frac{\boldsymbol{p}_{x_r}(k_i)}{\boldsymbol{p}_{x_v}(k_i)}\right), \tag{5.22}$$

where KL values close to 0 denote that the probability distributions $\boldsymbol{p}_{x_r}$ and $\boldsymbol{p}_{x_v}$ are almost identical in terms of highly reduced divergence or highly increased convergence.

5.6.5. Coefficient of variation (or variance to mean ratio)

The coefficient of variation (cV) [356] is a quantitative metric which is defined as the ratio of the standard deviation over the mean of a distribution. In fact, the cV score quantifies the variability of a given population with respect to the population mean and thus can be used to quantify the level of dispersity in the virtual distributions. Here, the cV is calculated for each real and virtual distribution per feature.

5.7. Data augmentation

The proposed pipeline for data augmentation is depicted in Figure 28, which consists of three modules, namely the: (i) data quality control module for assessing the quality of the data, (ii) virtual population generation module for producing high-quality virtual data, and (iii) the "hybrid" machine learning module for the development of disease classification and risk stratification models on the aggregated real and virtual patient data. The outcomes of the proposed pipeline include curated clinical data, high-quality virtual data and enhanced disease classification and risk stratification models.

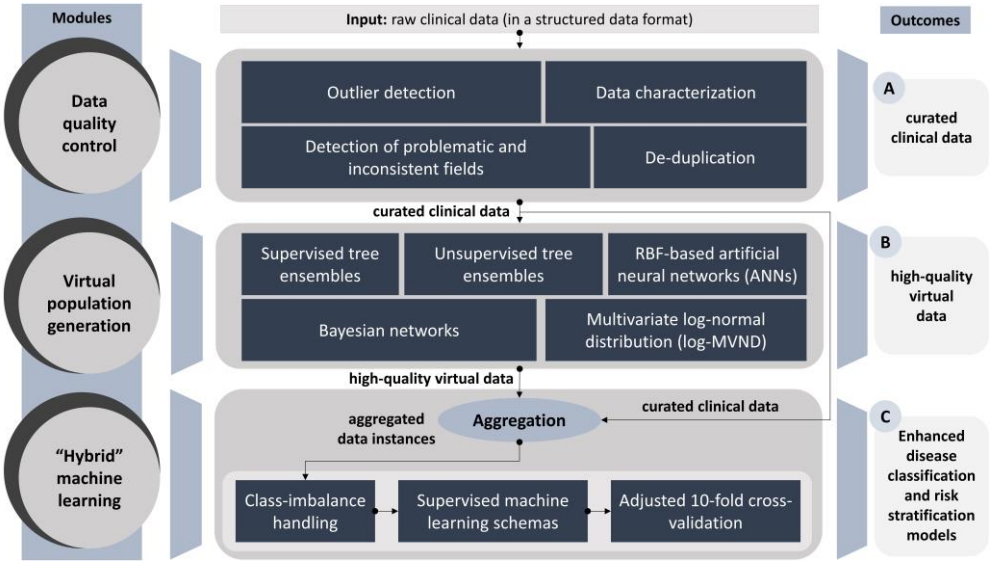


Figure 28. An illustration of the proposed computational pipeline.

A data quality control pipeline presented in a previous study [66] was utilized to automatically resolve problematic fields within the input clinical data, including outliers, data inconsistencies, and missing values. The curated clinical data are introduced into the virtual population generation module to yield virtual distributions that "mimic" the real ones. Towards this direction, state-of-the art machine-learning and statistical methods were developed to ensure the high-quality of the virtually generated data, including: (i) the supervised tree ensembles, where in each tree node, the generator for each feature is captured during the node splitting process based on its univariate empirical cumulative distribution function (ECDF), (ii) the unsupervised tree ensembles, where density forest ensembles are built in a top-down manner using the variance of the features as the criterion for the node splitting process, (iii) the artificial neural networks (ANNs), where radial basis functions (RBFs), such as, the Gaussian kernels are used as multivariate generators of virtual data instances, (iv) the Bayesian networks, where diverse network topologies are evaluated based on the causal relationships between the features (i.e., nodes in the network), and (v) the Log-MVND (multivariate log-normal distribution), where multivariate normal distributions are applied on the log transformed data. For each virtual population generation method, similarity scores, such as, the Kolmogorov-Smirnoff goodness of fit (GOF), the Kullback-Leibler (KL) divergence and the correlation coefficient are used to evaluate the level of agreement among the real and virtual distributions.

In the "hybrid" machine learning module, the virtual data from each generator are aggregated with the real data to assess whether the performance of the machine learning algorithms that are trained on the aggregated data is better than in the case where the algorithms are trained on the real data. Two case studies were conducted towards the development of robust lymphoma classification models in pSS and risk stratification models in HCM. Class imbalance handling was utilized to deal with the population imbalance among the control and target groups through the application of random downsampling with replacement on the control group. The XGBoost was deployed as a robust tree ensemble algorithm [357], [358] which was trained on aggregated data instances along with the Adaptive Boosting (AdaBoost) [359] and the Random Forests [360] which were also deployed to evaluate the overall impact of data augmentation. An adjusted stratified 10-fold cross validation procedure was utilized to train the algorithms on aggregated data and evaluate them on testing subsets of real patients.

To better understand the adjusted cross-validation process, lets denote the real dataset as $\boldsymbol{T}$, and the virtual datasets that were generated by the methods from 5.3 as $\boldsymbol{A}$ for the unsupervised tree ensembles, $\boldsymbol{B}$ for the supervised tree ensembles, $\boldsymbol{C}$ for the supervised RBF-based neural networks, $\boldsymbol{D}$ for the Bayesian networks, and $\boldsymbol{E}$ for the multivariate log-normal distribution. A 10-fold cross validation process is applied to $\boldsymbol{T}$, yielding a training subset $\boldsymbol{T_{train}}$ and a testing subset $\boldsymbol{T_{test}}$, on each iteration. On each round, each virtual dataset is aggregated with $\boldsymbol{T_{train}}$, yielding new instances, say, $\boldsymbol{A', B', C', D', E'}$, where $\boldsymbol{A' = A \cup T_{train}, B' = B \cup T_{train}, C' = C \cup T_{train}, D' = D \cup T_{train}, E' = E \cup T_{train}}$. Each supervised machine learning algorithm from 5.3 is trained on the training instances $\boldsymbol{A', B', C', D', E'}$ and tested on the corresponding testing instance $\boldsymbol{T_{test}}$ where the accuracy, specificity, sensitivity, and area under the curve (AUC) scores are computed and averaged across the folds. In this way, the training instances of $\boldsymbol{T}$ are augmented with the virtual data.

5.8. Summary

Virtual population generation has gained a lot of attention in the healthcare sector due to the overwhelming need to overcome the significant lack of sufficient population size, particularly for *in silico* clinical trials (ISCTs), where the financial burden of expensive drugs leverages the orchestration of viable Phase II/III CTs by pharmaceutical companies worldwide. Furthermore, the lack of medical databases with increased statistical power (e.g., in rare diseases) obscures the deployment of machine learning pipelines that can identify risk factors for disease progression and treatment due to the reduced amount of available training data. As a matter of fact, all these factors have a significant negative impact in the capacity of the existing healthcare systems, where the costs and delays for treatment and re-admission are already high. Virtual population generation envisages to address these needs through the development of virtual data generators which are trained on the real data to produce virtual (or synthetic) distributions which can “mimic” the real ones in terms of reduced divergence and dispersion with the real data. Since the virtual data quality is directly affected by the quality of the real data, it is first necessary to enhance the quality of the real data including the data completeness and conformity. So far, virtual population generation has multiple applications in ISCTs specifically in drug testing and development, as well as, in pharmacokinetics.

The emerging need for the development of computationally efficient virtual data generators yielding virtual data with reduced inter- and intra- correlation with the real data remains a technical challenge. The state-of-the-art virtual generators yield high-quality virtual data with reduced gof values, like the UTE. The gof, however, assumes that the distributions belong to a particular set of distributions which introduces biases in the outcomes. In addition, the STE, and the ANN require a target feature which affects the associations of the features in the virtual data. Furthermore, in the case of Bayesian networks, the number of all possible permutations of the edges within the network is infinite. Moreover, most of these methods are computationally demanding due to the increased training time.

Towards this direction, Gaussian Mixture Models (GMMs) with variational Bayesian inference (BGMM) were developed to generate large-scale virtual populations. The first method utilizes Dirichlet process mixtures as the BGMM's prior structure, where the concentration of each component on the weight distribution is an exponential function of the number of Gaussian components. We then extended the BGMM to reduce the biases which were introduced during the arbitrary selection of the number of Gaussian components by seeking the optimal number of components (BGMMOCE). To do so, we applied spectral clustering to get a first view on the number of clusters in the data. Then we extracted the optimal number of clusters based on the Davies Bouldin index and utilized it as the number of Gaussian components. In addition, we defined the weight concentration (gamma) parameter as the inverse of the number of Gaussian components.

The BGMMOCE algorithm is an extension of the conventional BGMM which aims to address open issues regarding hyperparameter estimation in BGMM which is a crucial technical challenge. BGMMOCE introduces a highly efficient spectral clustering stage based on the LOBPCG method to cluster patients with similar profiles within the input data towards the estimation of the optimal number of Gaussian components. In addition, the BGMMOCE is highly sustainable since it can be applied in any clinical domain. The need for large-scale synthetic data generation is more important than ever not only due to the COVID-19 crisis a few years ago but also because they are expensive and lengthy.

CHAPTER 6. FEDERATED/DISTRIBUTED LEARNING AND DATA ANALYTICS

6.1. Overview

So, how can we enable the co-analysis of disparate sources to deal with the unmet needs for various medical diseases and conditions? The answer to this question has been the core of discussion from the first chapter of this book. Everything begins from data sharing. The sources and types of big medical data are many and thus data sharing is a primary step to interlink disparate sources of medical data, such as, cohorts and omics registries to increase the population size and enhance the scientific impact of the clinical studies that make use of such data to provide outcomes with significant statistical power. The interlinked data might share a common medical domain but often exhibit structural heterogeneities due to the different data collection protocols and data recording schemas that are adopted by the data providers. So, once the interlinking mechanisms for data sharing are established, the next step is to overcome the structural heterogeneities that are present in the shared data. Towards this direction, data harmonization using lexical, semantic, and statistical matching based on reference schemas has been proven to be a powerful strategy that can enable the homogenization of the heterogeneously structured, interlinked medical data. Once the medical data are harmonized at the highest level of available information, the next step is to co-analyze the harmonized datasets using machine learning and data analytics.

Prior to the application of any data analytics strategy for distilling knowledge from medical data it is first necessary to apply a pre-processing pipeline on the medical data. Conventional computational approaches for data pre-processing include: (i) data curation [66] for enhancing the quality of the medical data, such as, the methods that have been already presented in 2.2.1, and (ii) data discretization [361] for dealing with recording errors during the data collection process, where the continuous data are discretized into an equal number of bins or bins with equal frequency or into a specific number of bins that minimizes the information entropy or maximizes the overall information gain, such as, the Entropy-MDL approach, among others. Feature selection and feature ranking approaches [52] are also useful for reducing the dimensionality of the input features by highlighting a subset of prominent features based on a target one, which is usually a disease outcome.

Once the quality of the data is ensured, a data analytics pipeline can be applied to mine knowledge from the medical data. Regarding the application of machine learning in healthcare, a notable progress has been made over the past years towards the extensive analysis of large amounts of clinical, laboratory, histological, and omics data to develop machine learning models for: (i) the prediction of disease outcomes, (ii) the detection of biomarkers, (iii) the effective treatment monitoring, and (iv) the development of patient stratification models, among others [129]. The capabilities of machine learning in medical data analytics are tremendous [4], [72] with increased impact towards the clinical decision-making process.

6.1.1. Machine learning in healthcare

Machine learning can shed light into complex data structures to reveal hidden patterns and associations between the variables that can lead to the detection of prominent variables with high contribution towards the prediction of a specific disease outcome [362], [363]. Towards this direction, machine learning algorithms have been extensively applied in omics data and clinical data for the development of supervised learning models that are able to predict disease outcomes given a specific subset of annotated data for training.

Popular supervised learning algorithms, such as, regression, support vector machines, the decision trees, the Naïve Bayes and the artificial neural networks have been utilized

on large subsets of clinical data to develop robust patient stratification models towards the identification of groups of individuals with high risk for the development of a disease outcome [52]. In addition, machine learning has been also applied on clinical and genetics data to develop unsupervised learning models for grouping (clustering) complex data structures with similar patterns [364]. Unsupervised learning algorithms, such as, the hierarchical clustering and the k-means [365], along with more complex ones, such as, the spectral clustering [366] and the hierarchical clustering [367] have been extensively employed for the categorization of highly associated features.

The fundamental difference between a supervised and an unsupervised learning algorithm lies on the fact that the former is trained on a set of input features (variables), where one feature is set as the target feature, whereas the latter is directly applied on the set of features without any prior knowledge regarding the existence of a target feature. As far as the development of prediction models is concerned, a supervised learning algorithm is usually trained on a set of training features where the feature that represents the disease outcome is specified as the target feature. The performance of the supervised learning model is then evaluated using conventional methods, such as, the k-fold cross-validation and train/test split ratio.

6.1.2. Problems with centralized analysis

It is obvious now that the adoption of a proper machine learning algorithm depends on the type of application study and the definition of the scientific problem that needs to be addressed by the study. Apart from the variety of the existing machine learning algorithms for data analytics in healthcare, emphasis must be given on the data storage environment. The most common way for storing medical data is through the adoption of a centralized database where all the medical data are stored into a common physical environment. Mining knowledge from large amounts of medical data requires the application of deep learning algorithms [368], such as, multi-layer neural networks with error propagation (e.g., the Long short-term memory neural network [369]) and convolutional neural networks (CNNs) [370]which are capable of detecting hidden motifs within the complex big data structures and dealing with the development of supervised learning models for predicting disease outcomes with multiple applications in medical imaging segmentation [55], [371] and bio-signal analysis [372]. Several methods have been proposed towards the effective analysis of big medical data in

centralized databases, such as, the batch processing method according to which the data are divided into smaller subsets, i.e., batches, and the ML algorithm is sequentially applied on the batches until all the batches are parsed.

Keeping the data in a common, centralized database, however, poses significant security threats in the case of a data breach, as well as, obscures the efficient analysis of big data, especially of omics data, where the amount of generated data is so huge that it significantly hampers the application of any machine learning algorithm due to the lack of sufficient memory units and the demand of high computational power [80], [373]. A prominent solution to this is to use the distributed database schema. ˝

Distributed healthcare environments [202], [204], [374] have gained a lot of attention these days due to the need to process massive amounts of accumulated medical data. In distributed databases, the medical data are stored in multiple sites (or locations). An overwhelming scientific challenge, however, in distributed databases, is the need to develop prediction models across the data that are stored in multiple databases, without the data to leave these databases at all. Another challenge lies on the fact that the application of the existing machine learning algorithms is not always feasible due to the non-convex optimization problem that the majority of these algorithms try to solve [375].

6.2. Types of learning

6.2.1. Online learning

Towards this direction, batch processing methods have been proposed to provide an adequate solution for the development of machine learning models in distributed environments [170]. Online learning [181] is such an approach which updates an existing machine learning model according to a global cost function that is sequentially adapted on upcoming data streams. In fact, online learning uses stochastic gradient descent optimization methods to update the existing machine learning model on upcoming training samples by minimizing a global cost function, where the model is continuously updated on new data points or on a series of accumulated data points, over time. Existing machine learning implementations that support online learning include, linear SVM [376], hybrid online learning attempts using non-linear kernels [377], as well as, gradient descent approaches for convex optimization [378].

6.2.2. Incremental learning

A similar strategy that shares a common basis with the former one, is the incremental learning [135], [184], [185], [187]. In contrast to online learning, incremental learning tries to adapt an existing machine learning model on upcoming data streams without the burden of being applied only on upcoming data streams, i.e., in an "online" manner but also on existing data streams or batches. More specifically, incremental learning uses a batch processing method to train a machine learning model on an initial batch and then adjust the model on a series of upcoming batches, by solving additive optimization tasks [135], [184], [185], [187]. This makes incremental learning ideal in the case of out-of-core learning, where the large-scale data do not even fit into the memory and thus need to be processed sequentially, as well as, when the batches are treated as harmonized data which are stored in multiple locations. Existing methods that support incremental learning include methods for convex optimization [183], [379], [380], gradient boosting trees [381], [382], and Naïve Bayes [188]. Meanwhile, stacked generalization techniques [383], [384] have been also proposed for combining individual classification outcomes. Such parallelized methods, however, suffer from biases introduced by the assembly stage [385].

6.2.3. Meta-learning

Meta learning [182] is a rigorous category of machine learning strategies where individual classification outcomes (metadata) are collected from the training of multiple classifiers on the same data and are finally combined to reduce the computational complexity of the incremental learning or online learning process offering some kind of parallel execution. Such methods, however, suffer from biases that are introduced during the assembly stage where the classification outcomes from different classifiers are combined.

6.2.4. Instance based learning

Perhaps the simpler of the two approaches is instance-based learning [386]. It is based on the solutions of previous instances (problems) to provide outcomes for new inputs. This is achieved by producing predictions based on the similarity (distance) of a new input to its nearest neighbors in the training dataset. This in turn implies that all known instances are stored in memory for use. In this approach the generalization is explicit,

no abstract models are involved in the process. This can lead to more adaptive generalizations, given that the implementation can simply store a previously unknown instance for future use. On the other hand, instance-based learning may lead to increased computational complexity due to the need of storing in memory large datasets of instances. A learner of this category may prove susceptible to data noise and overfitting.

6.2.5. Model based learning

This approach aims to create internal knowledge representations (abstractions) based on raw inputs [387]. A model-based learner attempts to construct and consequently refine its model of the environment it operates into to deduce a set of underlying properties which in turn are to be used for producing predictions when new/unknown data are inserted. In this scenario, direct interaction with the environment (fundamentally represented by the set of inserted raw inputs) is minimized in comparison to instance-based learning. This may lead to faster learning sessions in some cases and may also produce more robust learning paradigms. Robustness is demonstrated particularly when the Machine Learning implementation operates under lack of prior knowledge. Data noise issues may also be overcome by a valid knowledge representation. On the other hand, an invalid model will inevitably produce invalid predictions in all cases.

6.3. Beyond the state of the art - The proposed federated AI framework

6.3.1. Overview

Federated learning lies on the additive adjustment of a single estimator across multiple data structures [65], [81], [112], [135]. Given a set of M-distributed nodes (or databases in a federated environment), say $\boldsymbol{DN_1, DN_2, \dots, DN_M}$, we train a machine learning algorithm on the dataset $\boldsymbol{X_1}$ in node $\boldsymbol{DN_1}$, yielding an ML model, say $\boldsymbol{ML_{DN_1}}$, and then update the model through the following function:

$$F(x) = F(x-1) + \beta h(x), \tag{6.1}$$

where $F(x)$ corresponds to the estimated mapper that is trained on the dataset $\boldsymbol{X_i}$, in database $\boldsymbol{DN_i}$, $F(x-1)$ corresponds to the estimated mapper that was trained on the

dataset in the database $\boldsymbol{DN_{i-1}}$, where $i \leq M$, and $\beta h(x)$ is the learner function on $\boldsymbol{DN_i}$. To achieve this, we update the weights of the estimator through the stochastic gradient descent (SGD) method which seeks for a loss function, $h(F(x_i), y_i)$, that minimizes [129], [388]:

$$L(\boldsymbol{w}) = argmin\left(\frac{1}{N}\sum_{i=1}^{N} h(F(x_i), y_i) + ar(\boldsymbol{w})\right) \tag{6.2}$$

where, x_i is the i-th instance, y_i is the target, $\boldsymbol{w}$ is a weight vector, $h(.)$ is a loss function, a is a hyperparameter, $r(\boldsymbol{w})$ is a regularizer, $L(.)$ is the objective, and $F(x_i)$ is a linear score function. Solving (6.2) yields the weight update formula:

$$w_i = w_{i-1} - \eta_t\left(\nabla_w h(F(x_i), y_i) + a\nabla_w r(w)\right) \tag{6.3}$$

where, i is the stage, w_{i-1} is the weight estimation at stage $i-1$, η_t is a non-negative learning rate parameter, and $\nabla_w h(F(\boldsymbol{x_i}), \boldsymbol{y_i})$ is the gradient of the loss function $h(.)$. A pseudocode that summarizes the backbone of federated learning is presented in Algorithm 4. An ML algorithm is trained on the first dataset yielding the initial weights which are additively updated across the rest of the datasets through (6.3) using the weights of the previous ones.

Algorithm 4. A pseudocode for federated learning.

1	**def distributed_learning(F, $\boldsymbol{T} = \{\boldsymbol{T_0}, \boldsymbol{T_1}, \boldsymbol{T_2} \ldots, \boldsymbol{T_M}\}, \boldsymbol{w_0}$):**
2	fit an estimator $\boldsymbol{F_o}$ on the dataset $\boldsymbol{T_o}$ yielding $\boldsymbol{w_0}$
3	for $i = 0: M$ do:
4	retrieve weight vector w_{i-1} from the previous execution
5	solve $w_i = w_{i-1} - \eta_t\left(\nabla_w h(F(x_i), y_i) + a\nabla_w r(w)\right)$
6	update the weights based on (6.3)
7	return $[w_i, F_i]$

6.3.2. Federated learning algorithms

6.3.2.1. Federated Stochastic Gradient Descent (FSGD) based algorithms

The incremental strategy which is adopted by the federated AI modeling process (Algorithm 4) offers a unique scalability which allows us to extend conventional supervised machine learning classifiers for federated learning tasks. More specifically, the loss function, $L(f(\boldsymbol{d_i}), \boldsymbol{y_i})$, in (5) can be adjusted to build supervised machine

learning classifiers for federated training and testing. To develop the federated logistic regression (FLR) classifier we can replace the regularization term in (6.2) with the logistic loss function:

$$L(f(\boldsymbol{d_i}), \boldsymbol{y_i}) = ln\big(1 + exp\big(-\boldsymbol{y_i} f(\boldsymbol{d_i})\big)\big). \quad (6.4)$$

6.3.2.2. Federated Support Vector Machines (FSVM)

In a similar manner, we can develop the federated SVM (FSVM) algorithm using the hinge loss function:

$$L(f(\boldsymbol{d_i}), \boldsymbol{y_i}) = max\ (0, 1 - \boldsymbol{y_i} f(\boldsymbol{d_i})). \quad (6.5)$$

6.3.2.3. Federated Multinomial Naïve Bayes (FMNB)

In the case of discrete features, the multinomial Naïve Bayes (MNB) is preferred. Given an N-dimensional input vector, assume $d = (d_1, d_2, \dots, d_N)$, where d_i is the frequency of an event e_i, and the class, say c_k, with the highest probability or the maximum a-posterior (MAP) class, can be solved as a linear function [389] using the logarithm expression as follows:

$$c_{MAP} = argmax_{c_k}\left[log\big(P(c_k)\big) + \sum_{i=1}^{N} log\big(P(e_i|c_k)\big)\right]. \quad (6.6)$$

where $P(e_i|c_k)$ is the conditional probability of the event e_i given the class c_k, and k is the class index.

6.3.2.4. Federated Multilayer Perceptron (FMLP)

If we replace the loss function with the Perceptron loss:

$$L(f(\boldsymbol{d_i}), \boldsymbol{y_i}) = max\ (0, -\boldsymbol{y_i} f(\boldsymbol{d_i})), \quad (6.7)$$

we can develop the federated Perceptron classifier, as well as, the federated Multi-layer Perceptron (FMLP).

6.3.2.5. Federated Gradient Boosting Trees (FGBT)

In the case of the gradient boosting trees (GBTs) schema, regression trees ensembles are used as weak learners to minimize the expected value of the loss function. In the

case of the GBTs, we incrementally seek for the mapper $F(\boldsymbol{x})$ at a stage m, $F_m(x)$, as in [129], [357], [381]:

$$F_m(\boldsymbol{d_i}) = F_{m-1}(\boldsymbol{d_i}) + p_m \cdot h(\boldsymbol{d_i}; \boldsymbol{a_m}), \tag{6.8}$$

where p_m is the line search function, and $h(\boldsymbol{x}; \boldsymbol{a_m})$ is a regression tree learner with parameters $\boldsymbol{a_m}$.

6.3.2.6. Federated Gradient Boosting Trees with dropout (FDART)

A crucial problem in GBTs though is the fact that the trees added early in the ensemble tend to become more significant in the decision-making process than those added later. A solution to this issue is to use dropout rates [193], where the dropped trees and the newly added tree are scaled by a factor which ensures that the combination of the dropped trees and the new trees have the same effect on the outcome. To do so, the DART is trained on random subsets to prevent the definition of trivial trees. For a model, say Q, where $Q(d)$ is the prediction for sample d, and $L\big(Q(d)\big)$ is the loss function DART creates the random subset [193]:

$$\left\{\left(d, -\nabla_t L\big(Q(d)\big)\right)\right\}, \tag{6.9}$$

where, a new label with values $-\nabla_d L\big(Q(d)\big)$ is assigned for each sample d in the training dataset.

6.3.2.7. Federated Hybrid Boosted Forests (FHBF)

6.3.2.7.1. Issues with FGBT and FDART implementations

A common problem with FGBT, however, is the fact that trees added early in the ensemble, at a particular stage, tend to have a higher impact during the decision-making process those added later [193]. Dropouts have been recently adopted by the deep learning community to deal with this issue by scaling the most prominent trees in the ensemble with a specific rate of rejected trees. On the other hand, a main problem in FGBT with dropout rates is to account for overfitting effects in the selection of the dropout rate which is arbitrary. Besides, the data consistency in each database combined with increased class imbalance can leverage the weight update process yielding zero or infinite weights. On the other hand, even though federated implementations of

conventional supervised learning algorithms, like the support vector machines and logistic regression are easy to be implemented and deployed in federated environments they are often prone to overfitting effects since they suffer by linearity assumptions and thus fail to capture complex data structures.

On the other hand, Naïve Bayes approaches, such as, the multinomial Naïve Bayes are partially affected by overfitting they are often less flexible since they assume feature independence. Besides, although the GBT (with and without dropouts) algorithm has been widely used in the literature as a state-of-the art classifier with advanced implementations both in centralized and federated environments [65], [250], [389] none of these studies have investigated the loss during the training and testing across multiple and highly imbalanced data structures within federated environments.

6.3.2.7.2. Architecture

The FHBF architecture (Figure 29) is comprised of three individual layers: (i) the weight update layer, (ii) the separation layer, and (iii) the decision-making layer. In the weight update layer, the FHBF algorithm is utilized and recursively applied across the federated databases. The weight update process is repeated K-times by applying random downsampling with replacement with respect to a set of pre-defined confound factors among the control group and the target group in each federated database.

The weight update process is orchestrated by the central node (CN) which communicates with the federated AI model handler and the federated AI model collector. The former is responsible for the transmission and storage of the individual model weights from one database to another whereas the federated AI model collector is responsible for gathering the individual hybrid FDART (HFDART) models to formulate a set of clusters with the HFDARTs from each round. This set is referred to as a forest of HFDARTs.

In the separation layer, the “weak” HFDARTs models in the forest are identified by a log loss score and eliminated. The remaining HFDARTs are used for the final decision making based on majority voting. The output stage includes the final predictions along with explainable AI scores.

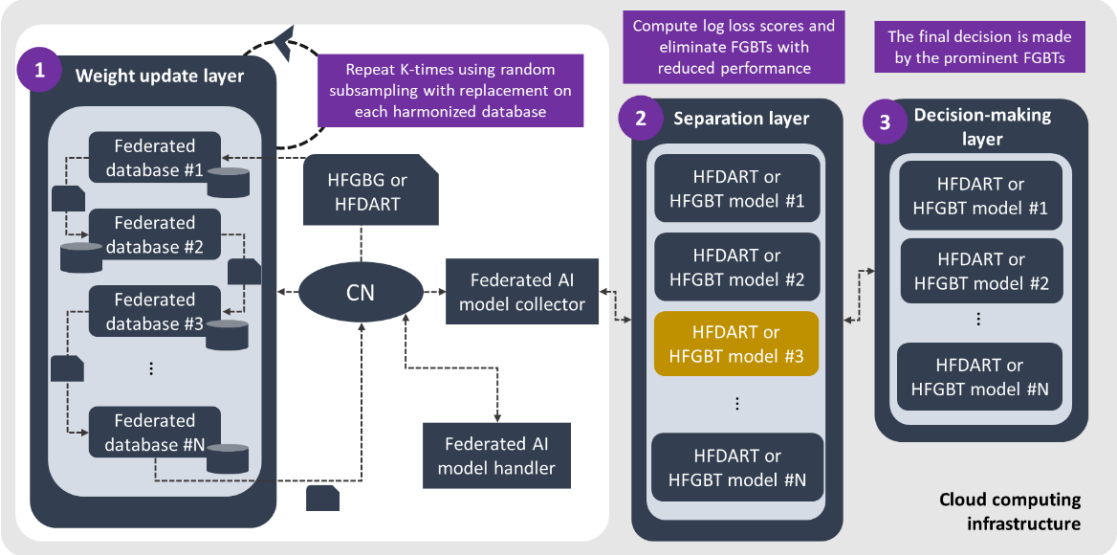


Figure 29. An illustration of the FHBF architecture.

6.3.2.7.3. Information flow

The information flow which reflects the core operations in the FHBF is depicted in Figure 30. According to Figure 30, the FHBF core parameters, including the number of rounds (K), the number of training databases (N), and the number of testing databases (M), are first defined. In each round, say $j \in [1, K]$, the algorithm gets sequential access to each training database, say $i \in [1, N]$, in the federated system. Random downsampling with replacement is applied on the training database to match the target group with the control group according to a predefined downsampling ratio (usually 1:1). The matching process is applied with respect to confound factors, such as, the age, gender, and the disease duration, to avoid biases during population matching. The hybrid loss function is then defined by the scale parameter with respect to the dropout rate. The first and the second order gradient of the loss function are then computed and utilized for the weight update process.

The updated weights are entered in a boosting process consisting of k rounds. In each boosting round, the weights of the model are updated to minimize the prediction loss. When the boosting process ends, the central node (CN) is invoked, and the weights of the model are stored. These weights are used for the training process in the next training database until all training databases participate in the analysis (i.e., until $j = N$). Once the training process is terminated, a cluster of federated models is created (i.e., a federated hybrid boosting forest - FHBF). Each cluster is evaluated in each testing

database, say $m \in [1, M]$ and the binary cross entropy loss (or log loss) is then estimated for each cluster.

As soon as the testing process is terminated, the "weak" clusters (i.e., those with log loss score less than the average log loss of the forest) are eliminated from the final decision-making process. Shapley additive explanation analysis is finally applied on the strong clusters to derive explainable scores for each input feature that participated in the workflow with respect to the target outcome.

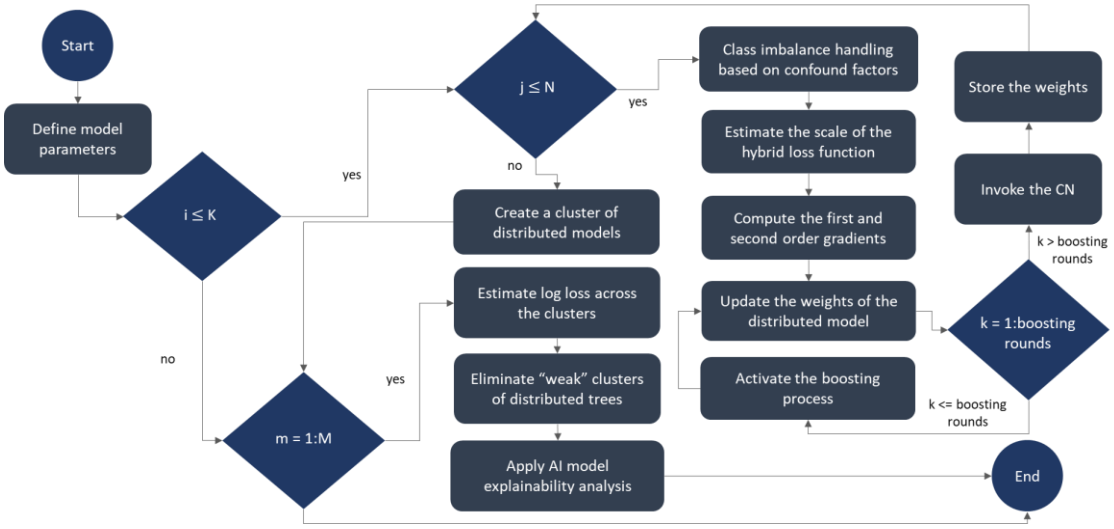


Figure 30. An illustration of the FHBF information flow.

6.3.2.7.4. Development of a hybrid loss function

A main problem in FDART is to account for overfitting effects in the selection of the dropout rate, say r. The fact that the algorithm combines many regression trees with a small learning rate and thus trees that are added early in the ensemble are more significant than trees added late. To solve this we propose a hybrid loss function which combines the *logcosh* loss [390], [391], say f, with the Huber loss [390], [391], say g, where the topology of the loss function is controlled by a parameter δ value. The *logcosh* loss f is defined as:

$$f(y, \hat{y}) = log\ (cosh(y - \hat{y})), \tag{6.10}$$

where y is the target vector and $\hat{y}$ is the vector with the estimations. On the other hand, the modified Huber loss, say $g(y, \hat{y}, \delta)$ is defined as:

$$g(y,\hat{y},\delta)=\begin{cases}\frac{1}{2}|y-\hat{y}|^2, & |y-\hat{y}|\le\delta\\ \delta\left(|y-\hat{y}|-\frac{1}{2}\delta\right), & |y-\hat{y}|>\delta\end{cases}, \tag{6.11}$$

where δ is a scaling parameter that corresponds to the point where the Huber loss transitions from a quadratic to linear shape. Then, f and g are combined into a hybrid loss function, say $h=f*g$ that is calculated based on the product rule yielding the first-order gradient:

$$\nabla h=log\ (cosh(y-\hat{y}))\big((y-\hat{y})/\sqrt{s}\big)+g(y,\hat{y},\delta)\ tanh(y-\hat{y}), \tag{6.12}$$

and the second-order gradient:

$$\nabla^2 h=\left(\frac{1}{cosh^2(y-\hat{y})}\right)g(y,\hat{y},\delta)+2\left(tanh(y-\hat{y})\left(\frac{y-\hat{y}}{\sqrt{s}}\right)\right)+tanh(y-\hat{y})\left(\frac{1}{s\sqrt{s}}\right), \tag{6.13}$$

where s is an approximation factor defined as $1+((y-\hat{y})/\delta)^2$ [392], [393]. The dropout rate rd was finally set equal to the scaling parameter δ so that the shape of the loss function would be steeper around 0 to avoid weight overfitting for large r.

6.3.2.7.5. Weight update function

The rationale of the gradient boosting process lies on the transformation of a set of weak learners into much a stronger one by additively updating the weights of the model until the prediction error is minimized. The error minimization process is usually based on the stochastic gradient approach (SGD). Given a set of N-observations $\{(\boldsymbol{x_1},y_1), (\boldsymbol{x_2},y_2), \ldots, (\boldsymbol{x_N},y_N)\}$, where $\boldsymbol{x_i}\in\boldsymbol{R^N}$, the objective is to obtain an estimated function, $\tilde{F}(\boldsymbol{x})$, mapping $\boldsymbol{x}$ to y, that minimizes the expected value of a loss function, assume $L(y,F(\boldsymbol{x}))$. Recall from Section 6.3.2.5 that the gradient boosting process incrementally seeks for estimations of a mapper at a stage $m\in M$, assume $F_m(x)$ as in:

$$F_i(x)=F_{i-1}(x)+\gamma_i f_i(x)F_i(x)=F_{i-1}(x)-\gamma_i\sum_{j=1}^{n}\nabla_{F_{i-1}}L(y_j,F_{i-1}(x_j)), \tag{6.14}$$

where the regularization objective can be approximated according to Taylor's theorem [357] as follows:

$$E(t) \approx \sum_{i=1}^{N} \left[L\left(y_i, \tilde{y}_{i,t-1}\right) + g_i f_t(x_i) + \frac{1}{2} h_i {f_t}^2(x_i) \right] + r, \tag{6.15}$$

where $l(.)$ is the loss function at step t, $\tilde{y}_{i,t-1}$ is the estimated target at step $t-1$, y_i is the real target and r is a regularization function:

$$r = \gamma L + \frac{1}{2} \lambda \sum_{j=1}^{L} {w^2}_j, \tag{6.16}$$

where w is the weight on the leaves, γ is a constant value, and L is the total number of leaves in each tree. Here, the first and second order gradients are used in (6.15), yielding the FHBF regularization objective:

$$\begin{aligned} E(t) \approx \sum_{i=1}^{N} \Bigg[& l\left(y_i, \tilde{y}_{i,t-1}\right) + (log\ (cosh(y - \hat{y})) \left(\frac{y - \hat{y}}{\sqrt{s}}\right) \\ & + g(y, \hat{y}, \delta)\ tanh(y - \hat{y})) f_t(x_i) \\ & + \frac{1}{2} \Bigg(\left(\frac{1}{cosh^2(y - \hat{y})}\right) g(y, \hat{y}, \delta) \\ & + 2 \left(tanh(y - \hat{y}) \left(\frac{y - \hat{y}}{\sqrt{s}}\right) \right) \\ & + tanh(y - \hat{y}) \left(\frac{1}{s\sqrt{s}}\right) \Bigg) {f_t}^2(x_i) \Bigg] + r. \end{aligned} \tag{6.17}$$

6.3.2.7.6. Confound based class imbalance handling

A crucial challenge that is introduced during the training across federated databases is the increased class imbalance among the control and the target groups in each database. To solve this, random downsampling with replacement was applied to extract a balanced set of training instances, in each database. The downsampling process is based on a pre-defined ratio, say dr, which determines the population size of the control group with respect to the target group. The downsampling process was applied on each database separately and was finally repeated K-times to obtain an unbiased estimation of the model performance. In each iteration, confound factors were taken into consideration to ensure subgroup matching without statistically significant differences.

More specifically, given a set of N-profound factors (features), say $\{\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_N}\}$, we seek for a random control subgroup where the patients' clinical profiles do not statistically deviate from those in the target group. To do so, the non-parametric Wilcoxon rank sum test (or Student's t-test in the case of normality upon a Shapiro-Wilk test) was applied on the continuous confound factors and the chi-square test/Fisher's exact test in the case of the discrete factors to evaluate whether the target subgroup and a randomly selected control subgroup, in each database, does not significantly deviate at a 95% confidence level.

6.3.2.7.7. Assembly stage and scoring procedure

In the assembly stage, the K individual HFGBT models are collected to formulate a set of HFGBTs in the form of a forest, say C, as in:

$$C = \{HFGBT_1, HFGBT_2, \dots, HFGBT_K\}, \tag{6.18}$$

where $HFGBT_i$ corresponds to the HFGBT from the i-th federated training round. In the case where the HFDART is used as a booster, then (16) is updated accordingly. The binary cross entropy (log loss) score is estimated for each model in the forest C as in:

$$H(C_j) = -\frac{1}{N}\sum_{i=1}^{N} y_i log\big(p(y_i)\big) + (1 - y_i) log\ (1 - p(y_i)), \tag{6.19}$$

where C_j corresponds to the model $HFGBT_j$, $j = 1, \dots, K$, y_i is the target of the i −th instance in the testing database, and $p(y_i)$ is the probability of the target class.

6.3.2.7.8. Collecting the final survivors

Clusters of HFGBTs (or HFDARTs in the case where the HFDART is used as a booster in the FHBF) with log loss score below the average log loss score in the forest C are marked as "weak" candidates and are discarded from the final decision-making process.

6.3.2.7.9. Decision making

As a final step, majority voting is applied on the "survivors" to derive the final predictions from a testing database. To this end, the decision-making process is formulated as follows:

$$\tilde{y}_i = \begin{cases} 1, & if \sum_{j=1}^{K} \tilde{y}_{i,j} > K/2 \\ 0, & o.w. \end{cases}, \quad (6.20)$$

where $\tilde{y}_i$ is the predicted value for the $i-$th instance. Weighted voting is also supported.

6.3.2.7.10. The FHBF pseudocode

A pseudocode of the FHBF is presented in Algorithm 5. The input parameters of the FHBF include the: (i) **'K'** which refers to the number of HFGBTs or HFDARTs in the forest, (ii) **'N'** which refers to the number of training databases, (iii) **'M'** which refers to the number of testing database(s), (iv) **'train_databases'** which refer to the locations of the training databases for the WebDAV API function, (v) **'test_databases'** which refer to the locations of the testing databases for the WebDAV API function, (vi) **'matching'** which refers to whether the user wants to apply population matching or not during the downsampling phase in each database, (vii) **'booster'** which refers to the type of booster (either HFGBT or HFDART), (viii) **'rate_drop'** which refers to the dropout rate, (ix) **'delta'** which refers to the scale of the hybrid loss (by default equal to 'rate_drop'), (x) **'loss_score'** which refers to the scoring function, and (xi) **'voting'** which refers to the voting approach (majority or weighted). Then, the FHBF estimates the weight of the model in the first federated database and stores them in the CN. The scaling parameter is determined along with the first and second order gradients based on (6.12), (6.13). The weights of the model are sequentially updated according to (6.17) and stored in the CN. The final model, say M_N is then retrieved and evaluated on a set of one or more testing databases. The weights of M_N and the predictions are stored in cluster C_j. The process is repeated K times and the collected clusters are aggregated into the forest C. Majority voting is applied to derive the final predictions y which are returned in the output.

Algorithm 5. A pseudocode of the FHBF algorithm.

Input parameters
K: number of HFGBTs or HFDARTs in the forest
N: number of training databases
M: number of testing databases

train_databases: locations of the training databases for access
test_databases: locations of the testing database(s) for access
matching: whether to apply triple confound based downsampling or not
booster: hybrid FGBT (HFGBT) or hybrid FDART (HFDART)
rate_drop: the dropout rate
delta: scaling hyperparameter of the hybrid loss function
loss score: the scoring function that determines the percentage of dumped trees in the forest
voting: whether to apply majority voting or weighted voting on the remaining trees
def FHBF (K, M, N, train_databases, test_databases, matching, delta, loss score, voting**)**:
for j in range(0, $\boldsymbol{K}$) do:
for i in range(0, $\boldsymbol{N}$) do:
determine the scaling parameter δ based on the dropout rate rd
compute first and second order gradients of the hybrid loss function according to (6.12), (6.13)
update the weights of M_i on federated database $i+1$ according to (6.17)
store the weights of the federated model M_{i+1} in the CN
retrieve the final federated model M_N from the training stage
evaluate the performance of M_N on the testing databases as in **test_databases**
store the weights of the M_N in cluster C_j along with the predictions
estimate the **log loss** to drop the "weak" clusters in C having the highest loss
apply **voting** to derive the final predictions, say y, from the survivors in C
return y;

6.4. AI model explainability

The SHapley Additive explanation analysis (SHAP) is a novel method from coalition game theory which can shed light into an AI model's decision-making process [124]. To do so, SHAP utilizes explanation models that yield interpretable and explainable classification outcomes. Given a subset of input features, say $P \subseteq \{d_1, d_2, \dots, d_Z\}$, from a larger set of K-features $\{d_1, d_2, \dots, d_K\}$, where $Z \leq K$, the SHAP value of a feature $d_j \in D$, say S_j, is defined as the overall contribution of this feature to the outcome [124]:

$$S_j = \sum \frac{|D|!\,(P - |D| - 1)!}{P!} \left(f_d(D \cup \{d\}) - f_d(D)\right), \tag{6.21}$$

where, K is the set of all input features, $|D|$ is the number of features in D, and $f_d(D)$ is the expected value of the function conditioned on P. To deal with the computational burden introduced in (6.21), we adopt an estimation process [394] which reduces the complexity from $O(TL2^Z)$ to $O(SLD^2)$, where T is the number of trees, L is the total number of leaves, Z is the number of features, and D is the tree depth. The cover metric was also used to measure the number of observations which are related to a particular feature. For each feature, the relative number of observations is calculated as the number of splits that this feature participated across each ensemble and averaged across the training instances on each distributed database.

6.5. Summary

Big data in healthcare can provide broader and more comprehensive insight on the optimization of the existing healthcare services to leverage the financial burden of unnecessary patient readmission, enable cost effective treatment, and improve the patient's quality of life (QoL). There is no doubt that the sharing of diverse clinical data from multiple data sources can enhance the statistical power of the studies that make use of such data. Towards this direction, the conventional strategy for knowledge mining across complex big data structures from multiple data sources is based on the co-analysis of the shared data which is usually referred to as centralized analysis. This type of analysis, however, is not always feasible either viable due to GDPR (General Data Protection Regulation) violations and increased risk for data breach, as well as, due to heavy computational burdens which are introduced during the training of demanding machine learning (ML) workloads across complex data structures. A solution to this critical issue is to deploy federated environments, where the diverse data from multiple sources are shared and stored under federated databases which are orchestrated by a federated data management system.

The technical advancements towards the application of data analytics in healthcare has made a significant progress over the past years. Data analytics is more useful and more powerful than even before yielding high performance machine learning and deep learning models for mining knowledge across massive amount of medical data. In general, a data analytics pipeline consists of three fundamental steps, namely: (i) data curation for enhancing the quality of the medical data by removing outliers, incompatibilities and inconsistencies, (ii) data mining using machine learning methods

for mining useful knowledge across the medical data through the development of patient stratification models, and the detection of biomarkers, among others and (iii) evaluation of the performance of such models using various performance indicators.

Then, we have gone one step further in order to examine how can we apply supervised and unsupervised learning algorithms across clinical data that are distributed across multiple sites. A solution to this can be given by the incremental learning and stacked generalization strategies. Incremental learning focuses on updating an existing machine learning model on upcoming data streams or batches, similar to online learning, although the former one can also support the offline processing of medical data. This means that incremental learning can be used to train machine learning models on massive amounts of medical data by diving them into subsets (batches), training the model on the first batch and sequentially updating the initial model on the remaining batches.

Indeed, if we replace the batches with harmonized datasets which are stored in multiple sets and use a central engine that will coordinate the communication between the sites then we can use incremental learning to sequentially update a machine learning model on these sites. Apart from incremental learning, someone would adopt the stacked generalization strategy and instead apply an individual machine learning model on each harmonized dataset and simply combine the classification outcomes using a meta-learner or a majority voting rule (e.g., weighted average) to yield the final ones. This approach, however, is prone to biases that are introduced during the assembly stage and limits the "horizon" of the training process since the individual models are trained on individual subsets.

Once the data are harmonized they need to be co-analyzed. The conventional approach is to integrate the harmonized data under a common database and apply machine learning to deal with the unmet needs in various medical domains, such as, the development of disease prediction models. A centralized database, however, is prone to privacy breach and computationally inefficient in the case of big data, where the memory and processing requirements are demanding. Towards this direction, batch processing methods have been proposed to deal with the analysis of big data by sequentially fetching the data into smaller subsets, where the machine learning algorithms are applied on an initial batch and then updated on the upcoming batches

until all the batches are being processed. Examples of batch processing methods include online learning and incremental learning. The main difference between them is the fact that the latter does not expect the data to arrive “online” in the form of data streams. Conventional implementations include the SVM with linear kernel stochastic gradient boosting based on ensemble classifiers, Naïve Bayes, etc.

So, we have come to a final question: Can machine learning (and artificial intelligence in general) be used to predict the future? Undoubtedly, machine learning can shed light into hard and complex scientific problems varying from the prediction of rare disease outcomes and the detection of biomarkers and therapy treatment to the prediction of environmental disasters and economic breakthroughs. The existing technology offers the basis for distilling knowledge across huge amounts of generated data including built-in hyperparameter optimization methods, parallelized computing units and high throughput technologies. Although the benefits of artificial intelligence are vast, so are the dangers of misusing it.

CHAPTER 7. CASE STUDIES

7.1. Autoimmune diseases

This case study involves the application of the beyond the state-of-the-art methods that were developed for data curation (CHAPTER 3), data harmonization (CHAPTER 4), synthetic data generation and augmentation (CHAPTER 5) and federated learning (CHAPTER 6) to address open issues and clinical unmet needs (Section 1.4) in the domain of the autoimmune diseases (Section 2.3.1) and particularly in patients who have been diagnosed with primary Sjögren's Syndrome (pSS).

7.1.1. Data curation

7.1.1.1. Case Study 1 – Demonstration and benchmarking of the proposed medical data curation workflow

The scope of this case study is to present the objectives, functionalities, and methodological advances of an integrated framework for medical data curation in terms of data quality assessment and validate the framework across two cohort studies. The developed framework was evaluated on two cohorts with anonymized data from patients that have been diagnosed with primary Sjögren's Syndrome (pSS). The anonymized data from the first cohort include 200 patients from the University of

Athens (UoA), whereas the second includes 100 patients from the Harokopio University of Athens (HUA). The cohort data were obtained under the data protection agreement version 3.7 as of August 2018 according to the Article 35 (3) (b) of the GDPR fulfilling all the necessary ethical and legal requirements for data sharing.

The UoA dataset consists of 162 features (58 discrete, 73 continuous, 31 unknown) and 440 instances with 44.56% missing values in total (Table 9). Out of 162 features, 91 were characterized as problematic; 60 features with more than 50% missing values and 31 features with unknown data type. The HUA dataset consists of 204 features (104 discrete and 94 continuous) and 100 instances with 33.61% missing values (Table 9). Out of 204 features, 69 were characterized as problematic; 63 features with more than 50% missing values and 6 features with unknown data types.

Table 9. Cohorts' metadata.

Cohort	**UoA**	**HUA**
Number of features	162	204
Number of instances	440	100
Discrete features	58	104
Continuous features	73	94
Problematic features	91	69
Missing values (%)	44.56	33.61

An example of the boxplot for four random features is depicted in Figure 31 (A), where values higher than 75% or lower than 25% of the value range are considered as outliers. The overall LOF distribution is depicted in Figure 31 (B), where values close to 1 are considered as outliers. Since the LOF is a multivariate method, its application is constrained to features with equal number of samples (and no missing values) and thus the LOF was computed only for the "good" features, i.e., those without any missing values.

According to Figure 31 (B), the LOF distribution does not indicate the existence of outliers due to the small number of "good" features. As for the rest of the methods, the missing values were ignored during the outlier detection process since they are univariate. The z-scores have also been computed for each feature. The z-score distributions of the four features of Figure 31 (A) are depicted in Figure 32, where the

features with values larger than 3 or lower than -3 are considered as outliers. The identified outliers were derived by four randomly selected features from the UoA cohort.

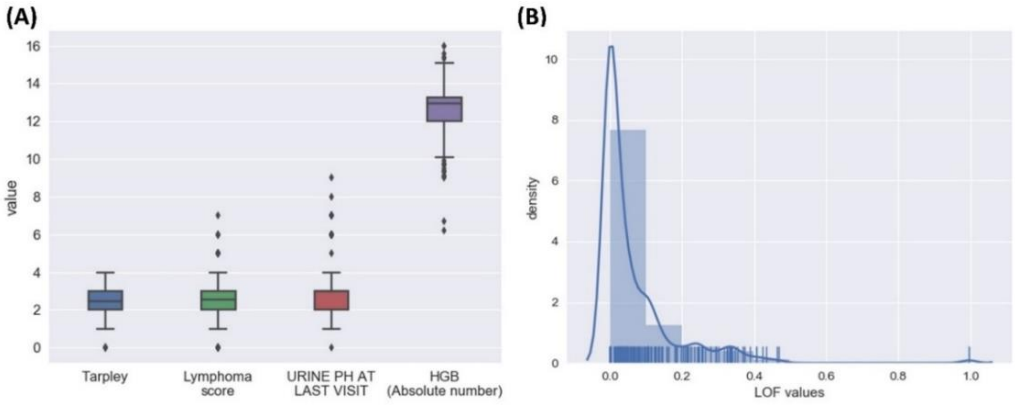


Figure 31. Results of two methods for outlier detection: (A) A boxplot for outlier detection based on the Interquartile Range (IQR) method for four randomly selected features, and (B) the overall Local Outlier Factor (LOF) distribution across a specific group of features of the dataset, where the density is the normalized frequency and the density curve is a smooth distribution over the histogram.

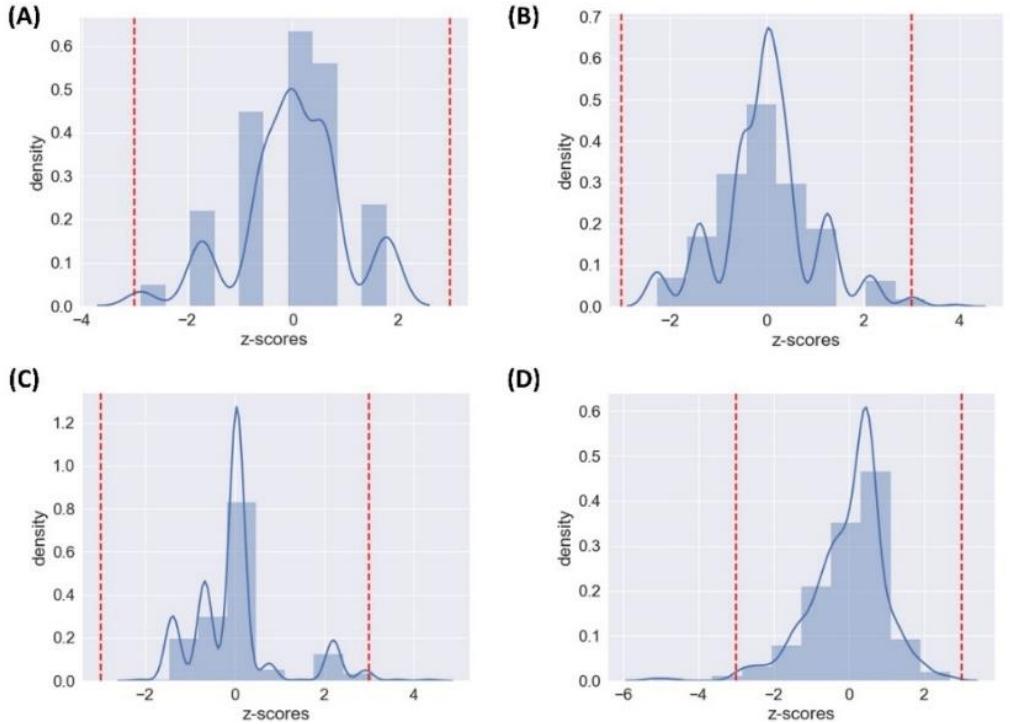


Figure 32. Z-score distributions for the four features of Figure 31 (A). Values that lie outside the red vertical lines are considered as outliers: (A) Tarpley, (B) Lymphoma score, (C) Urine pH at first visit, and (D) HGB (absolute number), where HGB stands for hemoglobin. In each plot, the density is the normalized frequency, and the density curve is a smooth distribution over the histogram.

The Spearman coefficient was computed for each pair of features resulting in a $162x162$ adjacency matrix for the UoA cohort and an $204x204$ matrix for the HUA cohort, with correlation values in the range (-1, 1), where 0 implies no correlation and +1 implies strong correlation (Figure 33 (A), (C)). In each adjacency matrix, the field (i, j) corresponds to the Spearman correlation between the features i and j. Each pair is also accompanied by a p-value which denotes the statistical significance of the correlation value (the confidence interval was set to 99%). Then, the pairs (i, j) having similarity value larger than 90% and $p < 0.01$, are highlighted as highly significant and correlated features. The Jaro distance has been also computed between each pair of feature labels to seek for potential duplicate features yielding a $162x162$ lexical distance matrix for the UoA cohort and an $204x204$ lexical distance matrix for the HUA cohort, where 0 implies no string matching and +1 implies features have the exact same labels (Figure 33 (B), (D)).

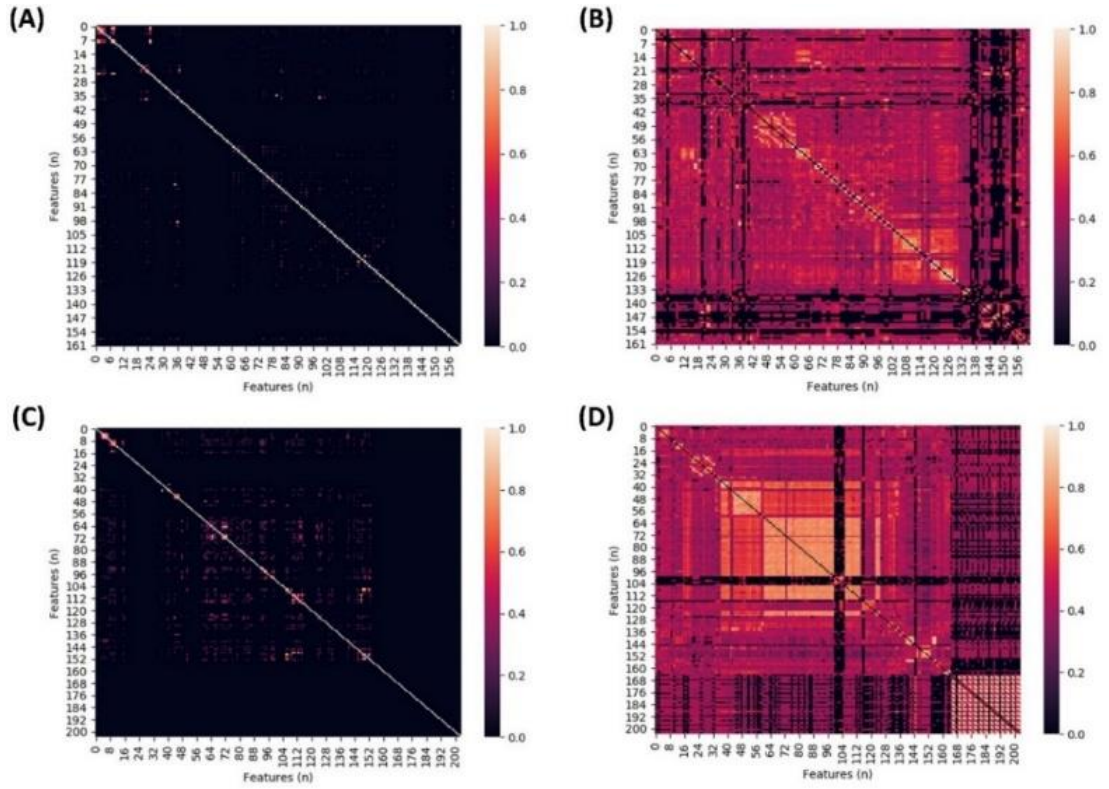


Figure 33. Correlation and lexical distance matrices for detecting highly-correlated and duplicated terms. (A) The 162x162 correlation matrix for the UoA dataset along with (B) the lexical distance matrix, (C) the 204x204 correlation matrix for the HUA dataset along with (D) the corresponding lexical distance matrix. The colorbars in the correlation and the lexical distance matrices is used to quantify the importance of the Spearman correlation and the lexical

similarity between each pair of features, respectively. A cell (i, j) that is depicted in black color denotes the absence of correlation (or lexical similarity) among the distribution of features i and j, whereas the light orange color denotes a strong correlation (> 0.9) between them.

The following pairs of features were highlighted for clinical evaluation from the UoA cohort; as highly-correlated: (i) {"Raynaud's phen (0-1)", "Rayanud"} with (rho=0.93, p<0.01), (ii) {"Ro/La", "Anti-Ro (0-1)"} with (rho=0.95, p<0.01), and (iii) {"Date of first biopsy", "Year of disease diagnosis"} with (rho=0.93, p<0.01), and as duplicate names: (i) "Lymphoma score" and "Lymphoma (0-1)" with s=0.95, (iv) "Lymphadenopathy (0-1) (fixed)" and "Lymphadenopathy" with s=1, (v) "Rose-Bengal Stain (0-1)" and "Rose-Bengal Stain" with s=0.98. As for the HUA cohort, the following pairs of features were highlighted as highly-correlated for clinical evaluation: (i) {"Antibodies to Ro(SSA) or La(SSB) antigens, or both at diagnosis or during follow-up", "Anti-Ro positive at diagnosis or during follow-up"} with (rho=1, p<0.01), (ii) {"Muscle biopsy", "Myopathy at diagnosis or during follow-up related to disease"} with (rho=1, p<0.01), and (iii) {"Elevated serum Creatinine", "Kidney Interstitial disease at diagnosis or during follow-up related to disease"} with (rho=0.91, p<0.01). In this case, no duplicate names were detected.

The data quality assessment report is one major output of the data curator which summarizes useful information regarding the value range of each feature, the type of each feature, the number of missing values, the state of each feature (based on the missing values), whether outliers were detected or not, and finally compatibility issues. An instance of the data quality assessment report can be seen in Table 10 for the UoA cohort and in Table 11 for the HUA cohort.

Table 10. An instance of the data quality assessment report for the UoA cohort.

Feature	**Value range**	**Type**	**Var. type**	**Missing values**	**State**	**Outliers[1]**	**Compatibility issues**
SEX (female=1)	[0, 1]	categorical	int	2	fair	no	no
First visit (year)	[1981, 2018]	numeric	date	0	good	yes	no
Year of disease diagnosis	[1982, 2018]	numeric	date	1	fair	no	no
Age at SS diagnosis	[14, 81]	numeric	int	1	fair	no	no
Whole salivary flow Date	[1985, 2018]	numeric	date	178	bad	no	yes, bad feature

Feature	Value range	Type	Var. type	Missing values	State	Outliers[1]	Compatibility issues
Dry eyes subj	[0, 1]	categorical	int	1	fair	no	no
Dry eyes, subjctiv Date	[1970, 2018]	numeric	date	39	fair	yes	no
Rose-Bengal Stain (0-1)	[+, 0, 1]	unknown	unknown	259	bad	not-applicable	yes, unknown type of data
Positive ocular stain score	[1/9, 5/9, 6/9, 8/9, 9/9]	categorical	string	429	bad	not-applicable	yes, bad feature
Abnormal Shirmer's	[0, 1]	categorical	int	96	fair	no	no
ANA+	[0, 1]	numeric	int	16	fair	no	no
RF (<20=0, >20=1) IU/ml	[0, 1]	categorical	int	52	fair	no	no
Anti-Ro (0-1)	[0, 1]	categorical	int	5	fair	no	no
Anti-La (0-1)	[0, 1]	categorical	int	10	fair	no	no
Date of first biopsy	[801, 19997]	numeric	date	16	fair	yes	no
Fat score	[0, 5]	numeric	int	345	bad	no	yes, bad feature
Germinal centers	[0, 1]	categorical	int	222	bad	no	yes, bad feature
MALT in MSG 1	[0, 1]	categorical	int	40	fair	no	no
SGE	[0, 1]	categorical	int	6	fair	no	no
Raynaud	[0, 1]	categorical	int	3	fair	no	no
Lymphadenopathy	[0, 1994]	numeric	int	5	fair	yes	no
Ro/La	[0, 1]	categorical	int	7	fair	no	no
RF+	[0, 1]	categorical	int	51	fair	no	no
LOW C4 (<20)	[0, 5]	numeric	int	38	fair	yes	no
Lymphoma score	[0, 7]	numeric	int	133	fair	yes	no
Type of monoclonal gammopathy	[IgA(κ), IgGk, IgGλ, IgMλ]	categorical	string	436	bad	not-applicable	yes, bad feature
MSG 2nd bxClonality Bx (0-1)	[0, 1, 1?]	unknown	unknown	416	bad	not-applicable	yes, unknown type of data
Urine pH at last visit	[0, 5, 5.5, 6, 6.5, 7, 8, 8.5, oj]	unknown	unknown	236	bad	not-applicable	yes, unknown type of data
Monocyte number	[42, 7540]	numeric	int	233	bad	yes	yes, bad feature
HGB (absolute number)	[6.2, 16]	numeric	float	76	fair	no	no
CRP (0,1)	[0, 1]	categorical	int	34	fair	no	no

Feature	Value range	Type	Var. type	Missing values	State	Outliers[1]	Compatibility issues
Anti-HCV (0-1)	[0, 1]	categorical	int	237	bad	no	yes, bad feature
Anti-HTLV-1 (0-1)	[0, 0]	numeric	int	465	bad	no	yes, bad feature
ANA (titer-1)	[0, 1/80, 1, 11280, <1/160, >1/640]	unknown	unknown	24	fair	not-applicable	yes, unknown type of data
IgG	[27.3, 5580]	numeric	float	291	bad	yes	yes, bad feature
IgM	[1.5, 1711]	numeric	float	294	bad	yes	yes, bad feature
LDH	[113, 495]	numeric	int	82	fair	no	no
AMA (titer-1)	[0, 1, 1/160, 164]	unknown	unknown	297	bad	not-applicable	yes, unknown type of data
Anti-TPO (0,1)	[0, 1]	categorical	int	211	fair	no	no
Anti-TG (titer)	[0, 1]	categorical	int	226	bad	no	yes, bad feature
Lymphoma (0-1)	[0, 1]	categorical	int	1	fair	no	no

[1]The z-score was used as the outlier detection method.

Note: The highlighted rows correspond to features where outlier detection was not-applicable (either empty features or features with unknown type of data).

Table 11. An instance of the data quality assessment report for the HUA cohort.

Feature	Value range	Type	Var. type	Missing values	State	Outliers[1]	Compatibility issues
Ethnicity	[6, 6]	numeric	int	0	good	no	no
Gender (0:F, 1:M)	[0, 1]	categorical	int	1	fair	yes	no
Year of Birth	[1927, 1988]	numeric	date	0	good	no	no
Year of diagnosis	[1987, 2018]	numeric	date	0	good	no	no
Year of first symptom	[1980, 2017]	numeric	date	0	good	yes	no
Year of first visit	[1991, 2018]	numeric	date	0	good	no	no
Year of last follow-up	[1999, 2018]	numeric	date	0	good	yes	no
Disease Duration Years	[0, 37]	numeric	date	0	good	yes	no

Feature	Value range	Type	Var. type	Missing values	State	Outliers[1]	Compatibility issues
SS CRITERIA 2002	[1, 5]	numeric	int	0	good	no	no
Oral Dryness	[0, 1]	categorical	int	1	fair	no	no
Ocular Dryness	[0, 1]	categorical	int	1	fair	no	no
Abnormal Schirmer Test	[0, 1]	categorical	int	37	fair	no	no
Abnormal Rose-Bengal	[0, 1]	categorical	int	46	fair	no	no
Abnormal BUT	[0, 1]	categorical	int	54	bad	no	yes, bad feature
Abnormal Shirmer OR_Rose Bengal	[0, 1]	categorical	int	1	fair	no	no
Lymphoma	[1, 2, NHL]	unknown	unknown	0	good	not-applicable	yes, unknown type of data
Year of Lymphoma development	[2000, 2017]	numeric	date	93	bad	no	yes, bad feature
WBC baseline (absolute number)	[2940, 10693]	numeric	int	3	fair	no	no
NEU baseline (%)	[12, 87]	numeric	int	18	fair	yes	no
LY baseline (%)	[9, 73]	numeric	int	19	fair	yes	no
PLT baseline (absolute number)	[20400, 3370000]	numeric	int	3	fair	yes	no
ESR	[5, 117]	numeric	int	9	fair	yes	no
Elevated CRP>10mg/l	[0, 1]	categorical	int	4	fair	no	no
CRYO positive	[0, 1]	categorical	int	68	bad	no	yes, bad feature
Elevated serum Creatinine	[0, 1]	categorical	int	3	fair	yes	no
Elevated SGOT>40IU/L	[0, 1]	categorical	int	2	fair	yes	no
Elevated SGPT>40IU/L	[0, 1]	categorical	int	3	fair	yes	no
Elevated γ-GT>30IU/L	[0, 1]	categorical	int	8	fair	yes	no
Elevated ALP>140IU/L	[0, 1]	categorical	int	2	fair	yes	no
Elevated LDH>230IU/L	[0, 1]	categorical	int	2	fair	yes	no
C3(mg/dL) baseline	[31, 202]	numeric	int	11	fair	yes	no
C4(mg/dL) baseline	[0, 85]	numeric	float	11	fair	yes	no

Feature	Value range	Type	Var. type	Missing values	State	Outliers[1]	Compatibility issues
Urinalysis pH at last visit	[5, 7]	numeric	float	28	fair	no	no
urinalysis/Pyuria (value WBC/hpf)	[0, 0]	numeric	int	97	bad	no	yes, bad feature
Microscopichematuria	[0, 1]	categorical	int	12	fair	yes	no
HBsAg positive 1=positive, 2=negative	[0, 2]	numeric	int	59	bad	yes	yes, bad feature
Other biopsies-sites: 0=no, 1=lymph node, 2=skin, 3=kidney, 4=Lung, 5=peripheral nerve, 6=bone marrow, 7=thyroid, 8=stomach, 9=small intestine	[3, 6, 3.0, 6.0, 8.0]	unknown	unknown	90	bad	not-applicable	yes, unknown type of data
HCV positive 1=positive, 2=negative	[0, 2]	numeric	int	62	bad	yes	yes, bad feature

[1]The z-score was used as the outlier detection method.

Note: The highlighted rows correspond to features where outlier detection was not-applicable (either empty features or features with unknown type of data).

Each variable (feature) is categorized into four different types of groups, namely: integer, float, date, and string. Moreover, there is an extra characterization into categorical and numeric. Categorical variables are those with binary values where the rest of the variables are denoted as numeric. This extra characterization can help the clinician to identify cases where categorical variables take values larger than 1 although such cases will be already detected as outliers. The state of each variable is denoted as good, fair, or bad, according to the number of missing values (see 3.3.6). For illustration purposes, we selected the z-score as a measure to detect outliers. Outlier detection is not applicable in cases where the features have unknown or string data type, as well as, in cases where the features are completely empty. For good features, outlier detection is normally applied whereas for fair or bad features, outlier detection is applied on the

non-missing values for maximizing the impact of detecting extreme values in such intensive data quality assessment cases.

Examples of features with unknown type of data for the UoA cohort, include the "Rose-Bengal Stain (0-1)", where the range of the values include a symbol "+" which is unknown and probably denotes positivity, the "Urine pH at last visit" which includes a value "oj" that is probably erroneously parsed, among others. These unknown symbols are also highlighted in the curated dataset with red color as incompatibilities.

An example of an outlier for the UoA cohort can be seen in the variable "Date of first biopsy", where the value range includes a minimum value of 801 and a maximum value of 19997, which denotes a discrepancy since it does not correspond to an ordinary year. A similar example occurs for the variable "Lymphadenopathy", where the normal range is "[0, 1]" but there exists a maximum value of 1994 which probably denotes a year that has been erroneously filled. In total, 13 features were characterized as problematic due to several discrepancies that were automatically detected (as described above) and 43 features were highlighted for the existence of potential outliers.

As for HUA, examples of features with unknown type of data, include "Lymphoma" where the normal range is "[0, 1]" but a string "NHL" exists which is unknown. Another example includes the variable "Other biopsies-sites" which takes values in the range "[0, 8]" but there are cases with patients having more than one values which denote that these patients have conducted biopsies in more than one sites, a fact that confuses the processing of data and the application of data analytics workflows.

An example of an outlier can be seen in the variable "HBsAg" where although the defined range is "1" for positive and "2" for negative, there are several zero values. The same applies for the variable "HCV". In total, 6 features were characterized as problematic due to several discrepancies and 82 features were highlighted for the existence of potential outliers (mathematically) that could lead to data contamination.

The results of the data curator REST service are depicted in Figure 34 for a pSS-related dataset. Through the REST settings, the user can define a local or global method for outlier detection (z-score, Interquartile range, Grubb's test, Local Outlier Factor). An example of the data quality assessment panel (similar to those that are presented in Table 10, Table 11) can be displayed in Figure 34 (C).

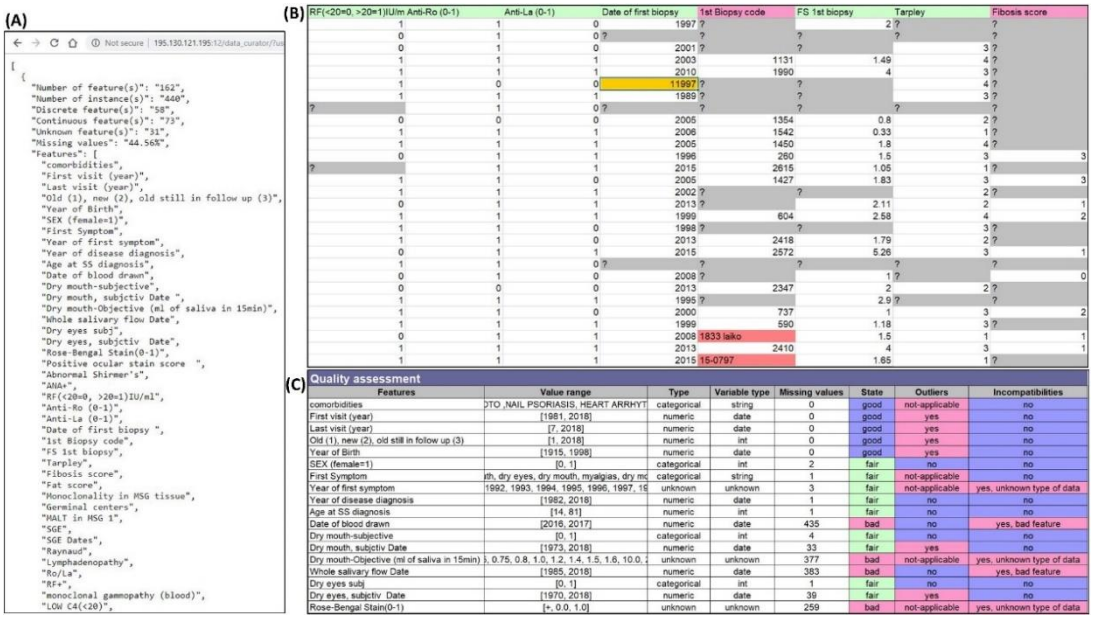


Figure 34. The results of the data curator REST service execution on the UoA cohort dataset that lies in a secure private cloud space: (A) An instance of the returned .JSON structure of the REST service call, (B) an instance of the curated dataset, and (C) an instance of the data quality assessment report.

An example of the produced curated dataset is displayed in Figure 34 (C), where: (i) the outliers are highlighted with gold color, (ii) the problematic fields are highlighted with red color, (iii) missing values with gray, (iv) fair features with green, good features with blue and bad features with rose. The XML schema of the reference model has been incorporated in the service, as well as, the NLTK language toolkit for data standardization purposes. In fact, the majority of the returned parameters from the .JSON structure are already summarized in the produced data quality assessment report for easiness.

We used an updated version of a reference model that was developed in a previous study [127], where a chart describing all the necessary requirements for defining the domain knowledge of the pSS (i.e., attributes descriptions and values) was provided by the clinical experts. The updated chart includes information regarding the ranges of the attributes and the class (category) where each attribute belongs to. Using this chart, a complete reference model was developed to reflect the meaning and range of each field. This common template includes a variety of patient-related information, such as, demographics, clinical tests, therapies. The types and ranges of each specified variable

within the template were determined during the development process according to the guidelines we received from the clinical experts.

Figure 35 depicts an example of the standardization process. The reference model is depicted as an XML schema (i.e., a semantic representation or an ontology) which describes the reference model using classes, sub-classes, and object properties. Each class consists of variables where each variable has a range which serves as a set of mapping values, a type, and its parent. Thus, the ontology can be seen as a three-level hierarchical model. In the first level lies the main class "Patient" which consists of four subclasses, i.e., (i) the "Demographics", (ii) the "Clinical tests", (iii) the "Therapies", and (iv) the "ESSDAI domain scores" (that belong to Level 2). Each class in Level 2 has further sub-classes (i.e., "Ocular tests", "Oral tests", "Laboratory tests") or variables (e.g., "C4 (mg/dL)") that belong to Level 3. For illustration purposes, the depicted schema describes only an instance of the pSS domain knowledge.

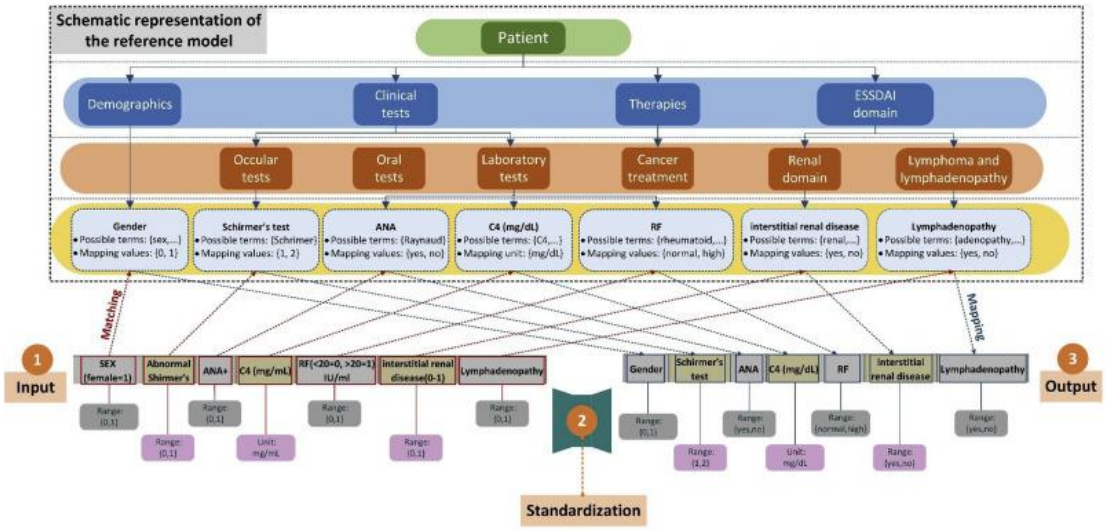


Figure 35. An illustration of the data standardization process.

A vocabulary was created using the pSS reference model. The terms of the reference model have been incorporated into an XML schema, so that the algorithm can automatically extract these terms and create the vocabulary. The classes denoted in Level 2 (Figure 35) were also specified by the clinical experts. The vocabulary consists of pairs (x, y) where x is the term of the reference model (e.g., "C3 (mg/dL)") and y is the class it belongs to (e.g., "Clinical tests"). The NLTK's WordNet corpus reader was used to enrich the existing vocabulary by computing synonymous/homonymous terms for each term. For example, "gender", and "sexual relations", are indicative examples

of synonymous sets (also referred to as synsets) for the term “sex”. The Jaro distance measure was used to calculate the similarity between each term of the raw dataset with those from the vocabulary. Matching block methods were used to match blocks among the terms and rules were developed according to standard value descriptions. The result of the standardization procedure is a tuple (xraw, xref, v, c), where xraw is the term of the raw dataset, xref is the matching term from the reference model, v is the matching score, and c is the class where xraw belongs to.

An illustration of the data standardization procedure is depicted in Figure 35, for a random instance of the UoA cohort dataset. According to Figure 35, the data standardization module receives as input the raw dataset. Then, it matches the term “SEX (female = 1)” of the input dataset with the homonymous term “gender” of the reference model and finally classifies it into the class “Demographics”. An example that involves the set of mapping values (i.e., the standard range) is depicted for the rest of the terms. The algorithm not only matches the term “Abnormal Shirmer's” with the term “Schirmer's test” and classifies it into the parent class “Clinical tests”, but also captures information related to the conversion of its value range from “0” and “1” to “1” and “2”, respectively. In addition, the term “ANA+” is matched with the term “ANA” and classified into the class “Clinical Tests” with additional information regarding the mapping of the “0” and “1” values to “yes” and “no”. Another example is shown for the term “RF (<20=0, >20 = 1) IU/mL” which is first matched with the term “RF” (Clinical Tests) and the mapping involves the conversion of the “0” and “1” values to “normal” and “high”. In a similar way, the “interstitial renal disease” and “Lymphadenopathy” terms are matched with the terms “Renal domain” and “Lymphadenopathy and Lymphoma domain” of the class “ESSDAI domain”, respectively, where the mapping involves the conversion of the values “0” and “1” to “yes” and “no”. Finally, the term “C4 (mg/mL)” is matched with the term “C4 (mg/dL)” along with additional information regarding the conversion of its measurement units from “mg/mL” into “mg/dL”.

The data standardization report provides useful information that can be used for data harmonization, such as, matching terms with similarity scores, final range of values and the classes of the ontology where the matched terms belong to. At this point, it is important to note that not all terms of the dataset are pSS-relevant. The pSS reference

model consists of pSS-related variables that are grouped into categories and describe the minimum requirements of the pSS domain and belong to different sub-domains, such as, demographics, laboratory tests, therapies, and ESSDAI domains. An example of a category is the salivary gland biopsy which consists of 10 variables (e.g., year of biopsy, age of patient at year of biopsy, focus score, etc.), the ESSDAI domain which consists of 12 sub-domains, including the constitutional domain, the lymphadenopathy domain, the glandular domain, the articular domain, etc. In total, there are 71 categories most of which involve more than one variable. According to the pSS minimum criteria that were posed by the clinical experts, the number of relevant terms for the UoA cohort was 81 out of 162 (Table 12) and for the HUA cohort, the number of relevant terms was 60 out of 204 (Table 13). Note that the HUA cohort is a rich clinical-oriented database (with detailed symptomatology) instead of a research-oriented one (UoA).

Regarding the UoA cohort, the data standardization module was able to successfully match and classify 73 out of 82 (89.02%) pSS-related terms (Table 12). As far as the HUA cohort is concerned, the data standardization module was able to successfully match and classify 52 out of 60 (86.6%) pSS-related terms (Table 13). In both tables, the matching terms are stated along with their similarity (matching) score, the final range of values and the class they belong to (1 = "Demographics", 2 = "Clinical tests", 3 = "Therapies", 4 = "ESSDAI domain"). Similarity scores were computed using the Jaro distance as a string-matching metric and the sequence matcher algorithm to identify matching blocks (patterns) among the terms of the input dataset with those from the reference model. Exact matches are those due to identical matching or due to synonymous matching and the similarity score is always equal to 1.

The rest of the matches are considered as partial. Partial matches are those that either achieve similarity score larger than 0.9 and/or when the sequence matcher detects matching blocks between the terms. If the sequence matcher identifies exact matching blocks among two terms, the similarity score is set to 1. The small number of pSS parameters that are observed in both cohorts comes from the fact that there exists a large group of variables which is related to the symptomatology of the different ESSDAI domains, including the "arthralgia" in the arterial domain, the "lung involvement" in the pulmonary domain, the "myositis" in the muscular domain, the "palpable purpura" and "non-palpable purpura" in the cutaneous domain, the "weight-

loss" and "fever" symptoms in the constitutional domain, the "kidney involvement" in the renal domain, etc. In fact, there are 41 ESSDAI-related symptoms in the HUA cohort and 31 in the UoA cohort that are not listed in the reference model, with the purpose of creating a more research-oriented data model.

Table 12. An instance of the data standardization report for the UoA cohort.

Feature	**Matched term or category from the reference model**	**Score**	**Type of match[1]**	**Captured range or measurement unit**	**Class[2]**
First visit (year)	Age at inclusion	1	partial	[1981, 2018]	1
Last visit (year)	Age at last follow-up	1	partial	[1991, 2018]	1
Year of birth	Year of birth	1	exact	[1918, 1995]	1
SEX (female=1)	Gender	1	exact	[0, 1]	1
Year of first symptom	Age at onset of first symptom	1	partial	[1971, 2016]	1
Year of disease diagnosis	Age at diagnosis of pSS	1	partial	[1983, 2018]	1
Age at SS diagnosis	Age at diagnosis of pSS	1	partial	[17, 84]	1
Dry mouth-subjective	Oral dryness	1	partial	[yes, no]	2
Dry-mouth, subjictiv Date	Oral dryness	1	partial	[1977, 2017]	2
Dry mouth-Objective (ml of saliva in 15 min)	Oral dryness	1	partial	[0, 2.5]	2
Whole salivary flow Date	Unstimulated whole saliva	1	partial	[1985, 2018]	2
Dry eyes subj	Ocular dryness	1	partial	[yes, no]	2
Dry eyes, subjctiv Date	Ocular dryness	1	partial	[1976, 2017]	2
Rose-Bengal Stain(0-1)	Rose-Bengal	1	partial	[0, 1, +]	2
Positive ocular stain score	Ocular staining score	1	partial	[1/9, 5/9, 6/9, 8/9, 9/9]	2
Abnormal Shirmer's	Schirmer's test	1	partial	[1, 2]	2
ANA+	ANA	1	partial	[yes, no]	2
RF(<20=0, >20=1) IU/ml	RF	1	partial	[normal, high]	2
Anti-La (0-1)	Anti-La	1	partial	[yes, no]	2

Feature	Matched term or category from the reference model	Score	Type of match[1]	Captured range or measurement unit	Class[2]
Monoclonality in MSG tissue	Serum monoclonal M component	1	partial	[yes, no]	2
MALT in MSG 1	Minor salivary gland biopsy	1	partial	[0, 1]	2
RF+	Rheumatoid factor	1	partial	[normal, high]	2
monoclonal gammopathy (blood)	Serum monoclonal M component	1	partial	[yes, no]	2
LOW C4 (<20)	C4	1	partial	mg/dL	2
Lymphoma score	Lymphadenopathy and lymphoma domain	1	partial	[0, 7]	4
Type of monoclonal gammopathy	Serum monoclonal M component	1	partial	[yes, no]	2
Time of 2st MSG biopsy (mm/yr)	Minor salivary gland biopsy	1	partial	[1985, 2017]	2
Code 2nd MSG Biopsy	Minor salivary gland biopsy	1	partial	[1231, …, parotid]	2
MSG 2nd bx Focus Score	Minor salivary gland biopsy	1	partial	[0.22, 12]	2
Time of 3st MSG biopsy (mm/yr)	Minor salivary gland biopsy	1	partial	[2006, 2017]	2
MSG 3nd bx Focus Score	Minor salivary gland biopsy	1	partial	[1.54, 22.84]	2
Time of 4th MSG biopsy (mm/yr)	Minor salivary gland biopsy	1	partial	[2013, 2015]	2
MSG 4th bx Focus Score	Minor salivary gland biopsy	1	partial	[1, 12]	2
Dyspareunia, subjctiv (0-1)	Dyspareunia VAS domain	1	partial	[0, 1]	4
Dyspareunia, subjctiv Date	Dyspareunia VAS domain	1	partial	[1985, 2016]	4

Feature	Matched term or category from the reference model	Score	Type of match[1]	Captured range or measurement unit	Class[2]
Abnormal Schirmer’s test (0-1)	Schirmer's test	1	partial	[1, 2]	2
Schirmer’s test date	Schirmer's test	1	partial	[1983, 19984]	2
Rose-Bengal Stain(0-1) 2	Rose-Bengal	1	partial	[0, 1]	2
Rose-Bengal Stain Date	Rose-Bengal	1	partial	[1987, 2018]	2
Rose-Bengal Stain Score	Rose-Bengal	1	partial	[3/9, 5/9, 6/9, 7/9, 8/9]	2
Chronic Fatigue (0-1)	Fatigue VAS domain	1	partial	[yes, no]	4
Chronic Fatigue date	Fatigue VAS domain	1	partial	[1988, 2017]	4
Vasculitic ulcer	Cryoglobulinemic vasculitis	1	partial	[yes, no]	2
Vasculitic ulcer date(-yr)	Cryoglobulinemic vasculitis	1	partial	[1985, 1985]	2
PNS-entrapment (0-1)	PNS domain	1	partial	[yes, no]	4
PNS-entrapment date	PNS domain	1	partial	[1975, 2011]	4
PNS-vasculitic (0-1)	PNS domain	1	partial	[yes, no]	4
PNS-vasculitic date(-yr)	PNS domain	1	partial	[1994, 2010]	4
CNS involvmt (0-1)	CNS domain	1	partial	[yes, no]	4
CNS involvmt date(-yr)	CNS domain	1	partial	[nan, nan]	4
Lymphadenopathy (0-1) (fixed)	Lymphadenopathy and lymphoma domain	1	exact	[yes, no]	4
Lymphadenopathy(fixed) date(-yr)	Lymphadenopathy and lymphoma domain	1	partial	[1992, 2013]	4
URINE SPECIFIC GRAVITY AT DIAGNOSIS	Urinalysis	1	partial	[normal, abnormal]	2
interstitial renal disease(0-1)	Renal domain	1	partial	[yes, no]	4
URINE PH AT FIRST VISIT	Urinalysis	1	partial	[normal, abnormal]	2
URINE PH AT LAST VISIT	Urinalysis	1	partial	[normal, abnormal]	2

Feature	Matched term or category from the reference model	Score	Type of match[1]	Captured range or measurement unit	Class[2]
γ-globulins(11-18=0,>18=1,<11=2)	Serum immunoglobulins	1	partial	[normal, high]	2
Anti-HCV (0-1)	anti-HCV antibody	1	exact	[yes, no]	2
ANA(titer-1)	ANA titer	1	partial	[0,1,1/1250, 1/1280, 1/160, 1/2560, 1/320, 1/5120, 1/640, 1/80, <1/160]	2
IgG	IgG	1	exact	[338, 7700]	2
IgM	IgM	1	exact	[59, 1370]	2
IgA	IgA	1	exact	[70, 1273]	2
LDH	LDH	1	exact	[113, 495]	2
C3 (mg/mL)	C3	1	exact	mg/dL	2
C4 (mg/mL)	C4	1	exact	mg/dL	2
Cryo (0,1)	Cryoglobulinemia	1	partial	[yes, no]	2
Cryo (type-II, IgMk) (0,1)	Cryoglobulinemia	1	partial	[yes, no]	2
Lymphoma (0-1)	Lymphadenopathy and lymphoma domain	1	exact	[yes, no]	4
Lymphoma diagnosis date	Lymphadenopathy and lymphoma domain	1	partial	[1986, 2018]	4
Anti-Ro (0-1)	Anti-Ro	1	partial	[1, ro]	2
TREATMENT EVER	Therapies	1	partial	[rituximab, azathioprine, pilocarpine, anti-TNFa, …]	3
TREATMENT last follow up	Therapies	1	partial	[naturale tears, corticosteroids, pilocarpine, rituximab, …]	3

Highlighted rows correspond to features with discrepancies (i.e., outliers and/or inconsistent).

[1]Class: 1 = "Demographics", 2 = "Clinical tests", 3 = "Therapies", 4 = "ESSDAI domain".

[2]Type of match: exact = identical terms, partial = highly-similar terms/terms with matching blocks.

Table 13. An instance of the data standardization report for the HUA cohort.

Feature	Matched term or category from the reference model	Score	Type of match[1]	Captured range or measurement unit	Class[2]
Ethnicity	Ethnicity	1	exact	[1, 6]	1
Gender (0:F, 1:M)	Gender	1	exact	[0, 1]	1
Year of Birth	Year of birth	1	exact	[1927, 1988]	1
Year of diagnosis	Age at diagnosis of pSS	1	partial	[1987, 2018]	1
Year of first symptom	Age at onset of first symptom	1	partial	[1980, 2017]	1
Year of first visit	Age at inclusion	1	partial	[1991,2018]	1
Year of last follow-up	Age at last follow-up	1	partial	[1999, 2018]	1
Oral Dryness	Oral dryness	1	exact	[yes, no]	2
Ocular Dryness	Ocular dryness	1	exact	[yes, no]	2
Abnormal Schirmer Test	Schirmer's test	1	partial	[1, 2]	2
Abnormal Rose-Bengal	Rose-Bengal	1	partial	[0, 1]	2
Abnormal Shirmer OR_Rose Bengal	Schirmer's test	1	partial	[1, 2]	2
Objective Evidence of Salivary Gland Involvement_Oral Tests	Oral tests	1	partial	[nan, nan]	2
Unstimulated whole salivary flow (<1.5 ml in 15 minutes)	Unstimulated whole saliva	1	partial	[nan, nan]	2
Site of FIRST salivary gland biopsy (1:lip, 2:parotid, 3:submandibular)	Salivary gland biopsy	1	partial	[1, 1]	2
FIRST MSGB Focus Score (no/4mm2)	Minor salivary gland biopsy	1	partial	[0, 13]	2
FIRST MSGB Focus Score ≥1	Minor salivary gland biopsy	1	partial	[0, 1]	2
FIRST MSGB B cell monoclonality	Serum monoclonal M component	1	partial	[yes, no]	2

Feature	Matched term or category from the reference model	Score	Type of match[1]	Captured range or measurement unit	Class[2]
SECOND MSGB Focus Score (no/4mm2)	Minor salivary gland biopsy	1	partial	[3.45, 11]	2
SECOND MSGB Focus Score ≥1	Minor salivary gland biopsy	1	partial	[1, 1]	2
SECOND MSGB B cell monoclonality	Serum monoclonal M component	1	partial	[yes, no]	2
ANA positive	ANA	1	partial	[yes, no]	2
ANA title	ANA titer	0.93	partial	[160, 2560]	2
ANA pattern	ANA pattern	1	exact	[1, 7]	2
Anti-Ro positive at diagnosis or during follow-up	Anti-Ro (autoantibodies)	1	partial	[yes, no]	2
Anti-Ro60 positive at diagnosis or during follow-up	Anti-Ro (autoantibodies)	0.93	partial	[yes, no]	2
Anti-Ro52 positive at diagnosis or during follow-up	Anti-Ro (autoantibodies)	0.93	partial	[yes, no]	2
Anti-La positive at diagnosis or during follow-up	Anti-La (autoantibodies)	1	partial	[yes, no]	2
RF positive >35IU/ml	Rheumatoid factor	1	partial	[normal, high]	2
Chronic Fatigue at diagnosis or during follow-up related to disease	Fatigue VAS domain	1	partial	[0, 1]	4
Lymphadenopathy at diagnosis or during follow-up related to disease	Lymphadenopathy and lymphoma domain	1	partial	[0, 1]	4
Vasculitic ulcer at diagnosis or during	Cryoglobulinemic vasculitis	1	partial	[yes, no]	2

Feature	Matched term or category from the reference model	Score	Type of match[1]	Captured range or measurement unit	Class[2]
follow-up related to disease					
Renal Tubular Acidosis at diagnosis or during follow-up related to disease	Renal domain	1	partial	[0, 1]	4
Renal insufficiency at diagnosis or during follow-up related to disease	Renal domain	1	partial	[0, 1]	4
Peripheral neuropathy at diagnosis or during follow-up related to disease	PNS domain	1	partial	[0, 1]	4
CNS involvement at diagnosis or during follow-up related to disease	CNS domain	1	partial	[yes, no]	4
Lymphoma	Lymphadenopathy and lymphoma domain	1	partial	[1, 2, NHL]	4
Year of Lymphoma development	Lymphadenopathy and lymphoma domain	1	partial	[2000, 2017]	4
Elevated CRP>10mg/l	Increased C-reactive protein	1	partial	[yes, no]	2
CRYO positive	Cryoglobulinemia	1	partial	[yes, no]	2
Elevated serum Creatinine	Creatinine	1	partial	[normal, high]	2
Elevated SGOT>40IU/L	AST	1	partial	[normal, high]	2
Elevated SGPT>40IU/L	ALT	1	partial	[normal, high]	2
Elevated γ-GT>30IU/L	γ-GT	1	partial	[normal, high]	2
Elevated ALP>140IU/L	ALP	1	partial	[normal, high]	2
Elevated LDH>230IU/L	LDH	1	partial	[normal, high]	2

Feature	Matched term or category from the reference model	Score	Type of match[1]	Captured range or measurement unit	Class[2]
C3(mg/dL) baseline	C3	1	partial	mg/dL	2
C4(mg/dL) baseline	C4	1	partial	mg/dL	2
Urinalysis Specific Gravity at last visit	Urinalysis	1	partial	[normal, abnormal]	2
Urinalysis pH at last visit	Urinalysis	1	partial	[normal, abnormal]	2
urinalysis/Pyuria (value WBC/hpf)	Urinalysis	1	partial	[normal, abnormal]	2
urineprotein/24h(mg/24h)	Urinalysis	1	partial	[normal, abnormal]	
GLUCOCORTICOIDS	Glucocorticoids	0.98	exact	[yes, no]	3

Note: The highlighted rows correspond to features with discrepancies (i.e., outliers and/or inconsistent).

[1]Class: 1 = "Demographics", 2 = "Clinical tests", 3 = "Therapies", 4 = "ESSDAI domain".

[2]Type of match: exact = identical or synonymous terms, partial = highly-similar terms or terms with exact matching blocks.

The REST service was implemented in Python 3.6 and was executed twice, one for each cohort through a secure virtual private network (VPN). The average execution time of the web service for the UoA cohort was 3.79 sec whereas for the HUA cohort the execution time was equal to 1.9 sec (Figure 36). More specifically, the time for fetching data was almost equal for both cohorts (< 1 sec). The average execution time for the application of the service including, data annotation and evaluation, outlier detection, similarity detection, and standardization, was 1.1 sec for the UoA cohort and 1.3 sec for the HUA cohort. The average execution time for constructing the data evaluation and data standardization reports along with the curated dataset was equal to 9 sec for the UoA cohort and 4 sec for the HUA cohort. According to Figure 36, the execution time for the data quality operations is affected by the number of features (for the HUA cohort the number of features is larger than the number of features from the UoA cohort) whereas the time for constructing the reports (data quality assessment and standardization) and the curated dataset is affected by the number of patients (for the UoA cohort the number of patients is 4.4 times larger than the number of patients from the HUA cohort). The small execution time demonstrates the dominance of automated data curation against traditional manual data curation where the time for identifying the outliers, and inconsistencies by both clinicians was large enough due to the size and

complexity of the datasets. The curated dataset was able to highlight all the cases with unknown data types, outliers and missing values, informing the clinicians that these cases would need their attention in just a few seconds. The data evaluation report was able to summarize the metadata information in the same amount of time.

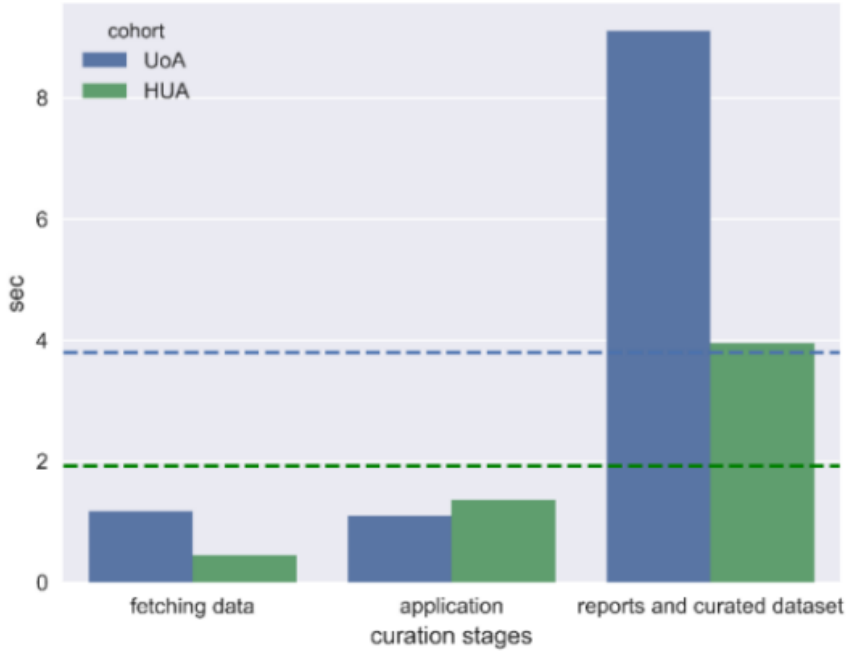
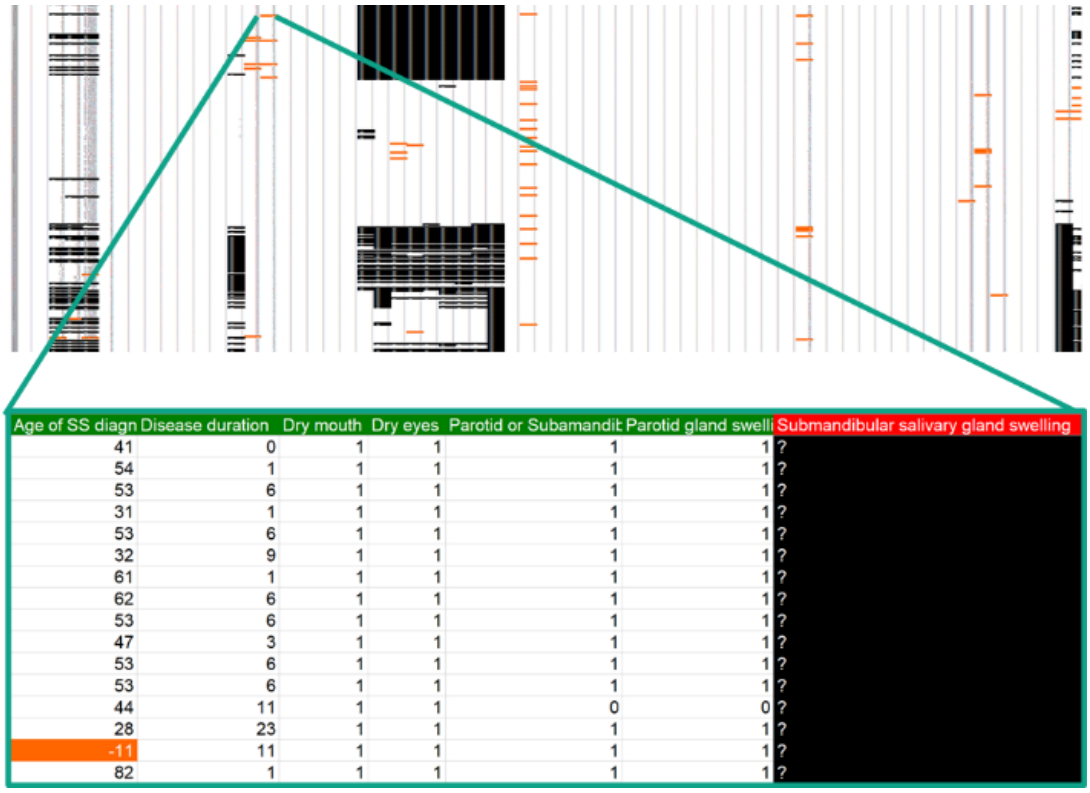


Figure 36. Execution time (in seconds) for the different stages (i.e., fetching data, application, reports and curated dataset) of the data curator's web service. The average execution times are depicted in horizontal lines (blue color: UoA cohort, green color: HUA cohort).

Our results confirm the validity of the proposed framework towards the precise identification of outliers, inconsistencies (unknown data types), and highly correlated and duplicated terms in both cohorts, as well as, the clinical usefulness and guidance of the data quality assessment report and the curated dataset towards the improvement of the overall accuracy, consistency, and relevance of the examined clinical data. The data standardization process was able to successfully capture more than 85% of the pSS-relevant terms in both datasets using lexical matching techniques combined with rules that use knowledge from a reference model. The framework uses an XML representation of the reference model which increases its overall scalability and thus can be generalized for different types of diseases, introducing the ontologies as a preliminary step for medical data harmonization. The fact that all the computational tasks were executed in a few seconds, demonstrates the dominance of automated data curation against traditional manual data curation.

7.1.1.2. Case Study 2 – Small-scale data curation

The scope of this study is to enhance the quality of 10 European cohorts in pSS. We acquired anonymized clinical data were collected from 10 databases with patients who have been diagnosed with primary Sjögren's Syndrome (pSS) under the HarmonicSS Project [343]. The 10 databases included 316 lymphoma patients (targets) and 4692 non-lymphoma patients (controls). According to the data quality diagnostics (Figure 37), a large portion of anomalies was detected in demographic- and laboratory-related measures, on each dataset, which were marked with orange color and removed from the analysis. All features were ranked based on their quality. Instances with green color have adequate quality whereas those with red color have poor quality and fields with black color denote missing values (Figure 37). A small portion (5%) of features with joint variability was identified between biopsy-related features. The flexible data harmonization approach yielded 41 features with more than 80% overlap across the 10 datasets.



Age of SS diagn	Disease duration	Dry mouth	Dry eyes	Parotid or Subamandit	Parotid gland swelli	Submandibular salivary gland swelling
41	0	1	1	1	1	?
54	1	1	1	1	1	?
53	6	1	1	1	1	?
31	1	1	1	1	1	?
53	6	1	1	1	1	?
32	9	1	1	1	1	?
61	1	1	1	1	1	?
62	6	1	1	1	1	?
53	6	1	1	1	1	?
47	3	1	1	1	1	?
53	6	1	1	1	1	?
53	6	1	1	1	1	?
44	11	1	1	0	0	?
28	23	1	1	1	1	?
-11	11	1	1	1	1	?
82	1	1	1	1	1	?

Figure 37. An instance of a selected dataset with quality diagnostics.

The data curator was able to enhance the quality in terms of accuracy, relevance, completeness, and conformity.

7.1.1.3. Case Study 3 – Large-scale data curation

The scope of this study is to enhance the quality of 21 European cohorts in pSS. A summary of the overall demographic information from the 21 European databases on pSS is presented in Table 14. The total number of eligible patients who fulfilled the inclusion criteria was 7,156, where the gender information was recorded for 7,000 patients (6,512 females, 488 males with a female to male ratio 13.34%). The average age at SS diagnosis in the female group was 51.82 (±13.96) years whereas in the male group the average age was 54.24 (±13.77) years.

Table 14. Demographic information.

Demographics	Females	Males
Gender	6,512	488
Age at SS diagnosis (mean±std)	51.82 (±13.96) years	54.24 (±13.77) years
Disease duration (mean)	7.08 years	5.59 years
Female to male ratio	13.34%	

Data curation was applied on each individual cohort database to automatically remove outliers, data inconsistencies and duplicated fields. The LOF algorithm was combined with the Isolation Forests to track down and remove outliers with 90% accuracy and the Spearman correlation coefficient was combined with the Jaro distance score to detect duplicated features. Data imputation was applied only to features with less than 30% missing values upon approval from the clinical experts. The automated data curation workflow enhanced the quality of the raw cohort data at a great extent.

7.1.2. Data harmonization

7.1.2.1. Case Study 1 – A reference model for pSS

The scope of this study is to develop a "gold" reference model for pSS. Metadata (feature/variable names, value ranges, short description) were extracted from 21 cohorts in pSS. A chart describing all the necessary requirements for defining the domain knowledge of the pSS (i.e., attributes descriptions and values) was provided by the clinical experts. Using this chart, a complete reference model was developed in order to reflect the meaning and range of each field. This common template includes a variety of patient-related information, such as, demographic, laboratory tests (e.g., oral, ocular,

OSS, etc.), therapies, etc. The types and ranges of each specified variable within the template have been determined during the development process according to the guidelines we received from the clinical experts, based on international measurement systems. Ontology mechanisms were recruited in order to represent the fundamental domains of the pSS, based on the reference model. The classes, sub-classes, and data properties between all the domains of the reference model were defined using Protégé [395]. Figure 38 presents the main class hierarchy along with indicative data properties that belong to the related subclasses. In addition, Figure 39 clearly depicts the graph model of the pSS ontology. The main class (or superclass) "Patient" is connected with its (sub-) classes, namely "Demographic", "Tests", "ESSDAI", and "Therapies", through the "has" object property (e.g., a patient has laboratory tests measures).

- Demographics: This class includes data properties related to the patient's demographic information, such as, ethnicity, gender, age of patient at pSS diagnosis, age of the onset of first symptoms of pSS, pregnancy (concerning the female population: outcome, number of twins, twin type, SS concordant), education level (i.e., none, elementary, intermediate, high school, university).
- *EULAR Sjögren's syndrome disease activity index (ESSDAI) domains* The European League Against Rheumatism (EULAR) Sjogren's syndrome disease activity index (ESSDAI) [396] is a disease activity index developed by the EULAR for patients with pSS. In this ontology, the ESSDAI is evaluated in twelve domains, namely the glandular, articular, cutaneous, renal, pulmonary muscular, constitutional, lymphadenopathy and lymphoma, hematological, central, and peripheral nervous system and biological. Each domain is described by its level of impact (i.e., no, low, moderate, high) as well as the corresponding weights/values which are necessary for the computation of the ESSDAI score.
- *Clinical tests:* This class consists of further subclasses, including (a) "Biopsies", (b) "Laboratory", (c) "Ocular", (d) "Oral", and (f) "Others". The former includes mainly salivary gland biopsy measures (e.g., year of biopsy, type of salivary gland, number of foci, focus score). In the case of other tissue biopsies, (i) the site, (ii) diagnosis, and (iii) the reason (for the assessment of SS or lymphoma) are required.
- *Laboratory:* The subclass "Laboratory" includes results and values for a variety of pathogenic clinical factors, such as, leukopenia, lymphopenia, anemia, neutropenia,

thrombocytopenia, LDH, C3, C4, CD3, CD4, CD8, CD19, serum albumin, proteinuria, cryoglobulinemia, cryocit, etc.

- *Ultrasound:* includes status indicators (positive/negative) related to the collection of parotid and submandibular ultrasound images according to the local radiologist or the scoring system in use. The subclass "Oral" tests is mainly comprised of results (i.e., positive/negative) and values for well-known oral tests, such as, salivary scintigraphy, sialography, (un-) stimulated saliva flow. In a similar way, the subclass "Ocular" tests includes results and values for three international ocular tests; (i) Schrimer's, (ii) van Bjisterveld's, and (iii) sicca occular staining score (OSS).

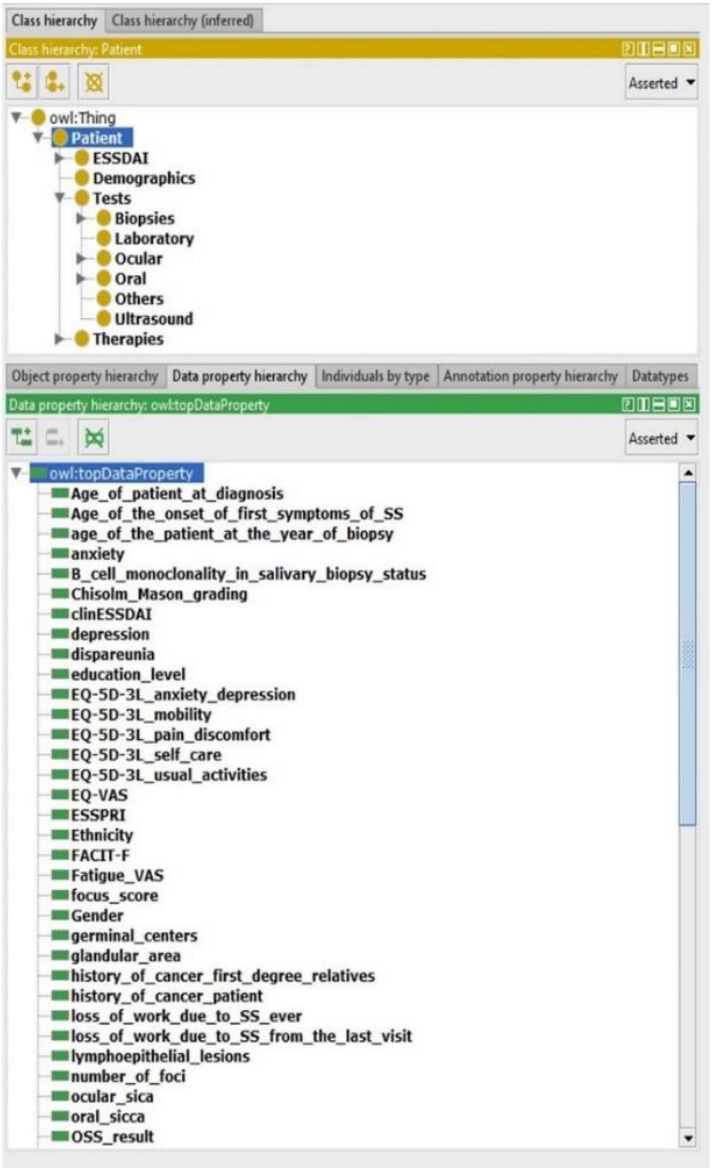


Figure 38. Visualization of the Patient's class hierarchy with a few indicative data properties from Protégé [395].

- *Others*: includes other types of tests and clinical related information, such as, the ESSPRI (EULAR Sjogren's Syndrome Patient Reported Index) [397], SSDI (Social Security Disability Index), oral/ocular/vaginal sicca visual analogue scale (VAS), smoking status, cancer history (not only for the patient but for his/her relatives as well), age-adjusted Charlston Comorbidity Indices (with 20 subfields, including dementia, rheumatic disease, myocardial infraction, peripheral vascular and cerebrovascular disease, Acquired Immunodeficiency Syndrome, hemiplegia, leukemia, lymphoma, mild and moderate liver disease, among others), five EQ-5D-3L items (i.e., mobility, self-care, usual activities, anxiety/depression, pain/discomfort), etc.

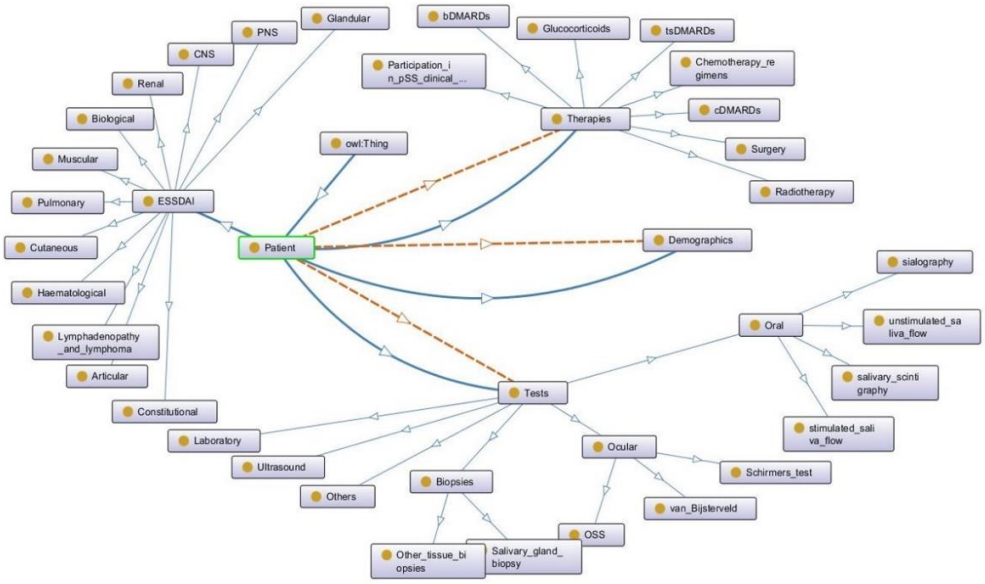


Figure 39. Graph representation of the pSS ontology using Protégé's OntoGraf [398].

- *WPAI Questionnaire:* The Work Productivity and Activity Impairment (WPAI) questionnaire [399] consists of six questions which are related to employment status, hours missed due to health problems or other reasons, working hours as well as the level of health affected productivity in regular activities and during working. WPAI provides useful information regarding the impairments in paid work and physical activities. Other related measures include the rheumatoid Arthritis Specific Work Productivity Survey (WPS-RA), as well as the Work Instability Scale for Rheumatoid Arthritis (RA-WIS) [400] among others. Here, the answers to the six WPAI questions have been included in the ontology as data properties of the "Patient" (super-) class.

- *Past/current therapies:* The class "Therapies" aims to collect and provide clinical information related to past/current therapies including the name of the therapy, the start date, the administration, the dosage and in some cases whether the drug was administered due to pSS or not. Examples of the therapies domain include various types of prescribed drugs, such as, biological disease-modifying antirheumatic drugs (bDMARDs), conventional disease-modifying antirheumatic drugs (cDMARDs), targeted synthetic disease-modifying antirheumatic drugs (tsDMARDs), glucocorticoids. In the case of chemotherapy, the regimens are recorded as well as the reason (due to pSS or not). The same stands for radiotherapy. Finally, the patient's participation in surgeries and other pSS clinical trials is also recorded with appropriate fields.

After finalizing with the classes, sub-classes, object and data properties definition process, the ontology has been published in the form of an .owl file on the following link: https://github.com/vpz4/PSS-Ontology. Each variable within the ontology is yet to be linked with international vocabularies with the purpose of enriching the ontology with information from external sources. We presented a first, complete, schematic representation of the pSS domain knowledge based on a predefined reference model provided by the clinical experts. The reference model was analyzed and transformed to a final version which allows the definition of the main pSS ontology. The innovation of this ontology lies on the fact that (a) it is a first hierarchical model that covers a large part of the pSS domain knowledge and (b) it serves as a common model for mapping heterogeneous pSS ontologies into a common one and thus enables medical data harmonization and integration. The latter enhance the statistical power of the participating cohort datasets leading to more accurate statistical models and effective outcomes. The ontology has been finally published for promoting research in the SS medical field in general. Since pSS is relevant not only due to its clinical impact but also as one of the few diseases to link autoimmunity, cancer development, as well as the pathogenetic role of infection, its examination can establish research in many areas of medicine.

7.1.2.2. Case Study 2 – Small-scale data harmonization

The scope of this study is to harmonize 4 European cohorts in pSS. We acquired anonymized clinical data from four European cohorts on primary Sjögren's Syndrome

(University of Athens (UoA); Harokopion University (HUA); University of Pisa (UNIPI); University of Udine (AOUD)). The cohort data (Table 15) were shared with the platform under the data protection agreement version 3.7 as of August 2018 according to the Article 35 (3) (b) of the GDPR fulfilling all the necessary ethical and legal requirements for data sharing.

Table 15. Demographic information.

	UoA	**HUA**	**UNIPI**	**AOUD**
Age at diagnosis	53.5±13.5	47.5±12.3	51.4±14	52.4±13.9
Gender ratio (females/males)	415/25	96/3	693/25	274/23
Lymphoma/Non-lymphoma (%)	76/364 (20.87%)	6/93 (6.45%)	31/687 (4.51%)	26/271 (9.59%)
Total number of patients	440	99	718	297

The extracted cohort metadata are presented in Table 16. In total, 31 features were inconsistent in the UoA cohort, 6 in the HUA cohort and 1 in the AOUD cohort, where the UoA and HUA cohorts had the highest number of bad features. The total percentage of missing values was 44.8% for the UoA cohort, 33.61% for the HUA cohort, 21.98% for the UNIPI cohort and 17.15% for the AOUD cohort. No outliers were detected.

Table 16. Extracted cohort metadata.

	UoA	**HUA**	**UNIPI**	**AOUD**
Number of features	167	204	102	82
Number of instances (cases)	440	100	718	297
Categorical features	76	146	85	75
Numeric features	60	52	17	6
Good features	27	62	37	51
Fair features	59	74	50	17
Bad features*	81	68	15	14
Features with outliers	0	0	0	0
Features with inconsistencies	31	6	0	1
Total % of missing values	44.8%	33.61%	21.98%	17.15%
*these features were discarded from further analysis.				

The cohort data harmonization workflow was applied on the curated cohort data. The pSS ontology was used as a gold standard to enable the terminology alignment of each

cohort dataset. The number of relevant terms with the pSS reference model was initially identified by the clinical experts. According to Table 17, the data harmonization process was able to match more than 85% of the reference model terms in all four cohorts (UoA: 92.3%; HUA: 90.47%; UNIPI: 88.88%; AOUD: 89.13%) yielding harmonized data with increased statistical power. Moreover, the number of terms requiring data standardization was significant in the AOUD cohort (14 terms) whereas in the remaining cohorts the terms were already in line with the pre-defined range values in the reference ontology.

Table 17. Cohort data harmonization results.

	Cohorts			
	UoA	**HUA**	**UNIPI**	**AOUD**
Number of terms *	82	136	87	67
Relevant terms with the pSS reference model **	39	42	54	46
Lexically similar terms with those from the ontology	36	38	48	41
Percentage of harmonized terms	92.3%	90.47%	88.88%	89.13%
Number of terms requiring data standardization ***	2	3	1	14
Common number of terms ******	19			

* after the removal of terms having more than 50% missing values (through data curation).

** the number of pSS-relevant terms for each cohort was identified by the clinical experts (after evaluation).

*** the number of terms for data transformation according to the range values in the ontology.

**** the number of common harmonized terms (not individual terms) across the cohorts.

To demonstrate the consistency of the harmonized cohort data we applied Principal Component Analysis (PCA) on each harmonized cohort dataset, separately, as well as, on the integrated dataset and extracted the first two principal components (PCs) as those that describe the largest portion of variance within the data. The distributions of the two PCs from each harmonized cohort against those from the integrated cohort, are depicted in Figure 40 and Figure 41, respectively. To offer a quantitative way to demonstrate the consistency of the data after the data harmonization process, we applied the Wilcoxon rank-sum statistical test to examine the null hypothesis that the distributions of the two PCs between the individual, harmonized cohort data and the integrated cohort data are common. In all cases, the p-values were larger than 0.05 which denotes that the distributions of the PCs between the individual harmonized cohort data and the integrated cohort data are not significantly different.

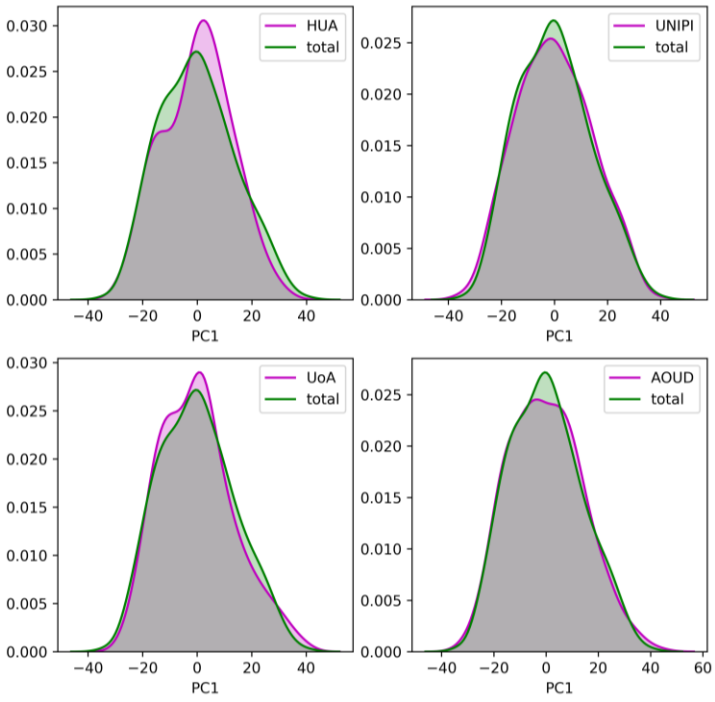


Figure 40. The distribution of the first principal component for each harmonized cohort dataset against the integrated dataset.

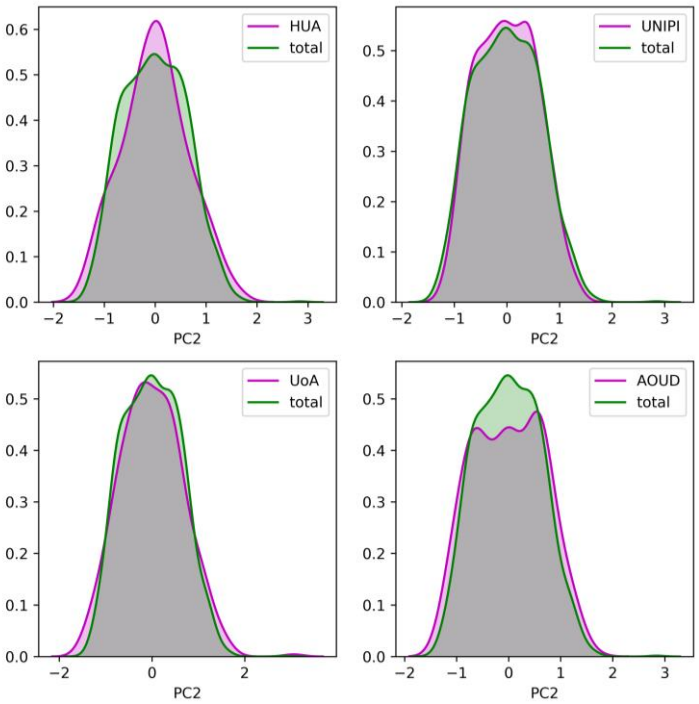


Figure 41. The distribution of the second principal component for each harmonized cohort dataset against the integrated dataset.

The data harmonization strategy is semi-automated and requires a reference model as input. This strategy is in line with the majority of the state-of-the-art data harmonization

tools, such as, the BiobankConnect software [143], the SORTA [142] and the DataSHaPER [141], [147] which require the definition of a common set of parameters for the domain of interest prior to the harmonization process. The consistency of the harmonized cohort data was demonstrated by the qualitative and quantitative comparison of the principal components between each harmonized cohort dataset and the integrated harmonized dataset, which suggest an increased homogeneity among the harmonized cohort data. The importance of the proposed approach, however, lies on the fact that it is more clinical-oriented and it is based on the definition of a disease-oriented ontology for the purposes of pSS which is a rare, autoimmune disease with a reported lack of domain knowledge instead of the genome-oriented tools, like the DataSHaPER and the BiobankConnect software which are used for the harmonization of genomic data.

7.1.2.3. Case Study 3 – Large-scale data harmonization

The scope of this study is to harmonize 21 European cohorts in pSS. A summary of the overall demographic information from the 21 European databases on pSS is presented in Table 18. The total number of eligible patients who fulfilled the inclusion criteria was 7,156, where the gender information was recorded for 7,000 patients (6,512 females, 488 males with a female to male ratio 13.34%). The average age at SS diagnosis in the female group was 51.82 ($\pm$13.96) years whereas in the male group the average age was 54.24 ($\pm$13.77) years.

Table 18. Demographic information.

Demographics	Females	Males
Gender	6,512	488
Age at SS diagnosis (mean$\pm$std)	51.82 ($\pm$13.96) years	54.24 ($\pm$13.77) years
Disease duration (mean)	7.08 years	5.59 years
Female to male ratio	13.34%	

Ontologies were constructed for each curated cohort database based on the extracted metadata. Semantic mapping rules were defined between the individual ontologies and the pSS reference ontology. As shown on Table 19, the cohort data harmonization process resulted in 48 common concepts (or terminological concepts) which constitute the pSS minimal criteria (minimal common data elements) across the 21 federated cohort databases.

Table 19. The common set of terminologies.

Feature	Presence/ Abnormal	Absence/ Normal	Mean	Median
Gender	6879 (females)	513 (males)	-	1
Age at SS diagnosis	-	-	51,95	53
Disease duration	-	-	7,23	6
Dry Mouth (aka Xerostomia)	6101	699	-	1
Dry Eyes	6046	748	-	1
Parotid or Submandibular swelling	1897	3186	-	0
Parotid Gland swelling	1616	2227	-	0
Submandibular salivary gland swelling	139	2648	-	0
Raynaud's Phenomenon	1577	4365	-	0
Fatigue	2840	2416	-	1
Arthritis	993	5006	-	0
Renal Disease	162	6021	-	0
Tubulointerstitial Nephritis	66	4663	-	0
Glomerulopathy	37	4293	-	0
Membranoproliferative Glomerulonephritis (MPGN)	15	4309	-	0
Membranous Glomerulonephritis (MGN)	3	4163	-	0
Mesangioproliferative Glomerulonephritis (MPGN)	9	4157	-	0
Other Glomerulonephritis	4	4501	-	0
Pulmonary Disease	415	5421	-	0
Small Airway Disease	157	5012	-	0
Lymphocytic Interstitial Pneumonia (LIP)	47	4414	-	0
Nonspecific Interstitial Pneumonia (NSIP)	34	4034	-	0
Usual Interstitial Pneumonia (UIP)	31	4041	-	0
Cryptogenic Organizing Pneumonia (COP)	0	4077	-	0
Liver Disease	131	5300	-	0
Autoimmune Hepatitis (AIH)	40	4682	-	0
Primary Biliary Cholangitis (PBC)	81	5535	-	0
Sclerosing cholangitis	11	4639	-	0
Nervous System Disease	560	5577	-	0
Peripheral Nervous System Disease	267	4712	-	0
Central Nervous System Disease (CNS)	125	4487	-	0
PalpablePurpura	396	5257	-	0
CutaneousDisease	458	4158	-	0
Muscular System Disease	357	4633	-	0

Feature	Presence/ Abnormal	Absence/ Normal	Mean	Median
IdiopathicInflammatoryMyopathy (IIM)	10	3842	-	0
Inclusion Body Myositis (IBM) documented with Biopsy	150	4429	-	0
B-cell Mucosa-associated Lymphoid Tissue (MALT) Lymphoma	245	5324	-	0
Diffuse Large B-cell Lymphoma (DLBCL)	45	5326	-	0
B-cell Nodal Marginal Zone Lymphoma (NMZL)	24	5346	-	0
B-cell Splenic Marginal Zone Lymphoma (SMZL)	6	4789	-	0
Other mature B-cell neoplasms	21	5305		0
Anti-La-SSB [presence]	2670	3499	-	0
Anti-Ro-SSA [presence]	4565	1703	-	1
Rheumatoid Factor (RF) [Units-volume]	2282	2362	-	0
Antinuclear Antibodies (ANA) [presence]	4354	1096	-	1
C4 levels (Serum complement) [Mass-volume]	2485	1725	-	1
Cryoglobulins [presence]	266	4409	-	0
Lymphoma*	354	5653	-	0
* The records of patients with missing lymphoma status were ignored from the analysis.				

The data harmonization workflow uses lexical and semantic matching to identify terminologies with common lexical and conceptual basis, where the pSS reference model is expressed into a .RDF/.OWL format. The harmonization process yielded 48 common concepts (or terminological concepts) which constitute the pSS minimal criteria (minimal common data elements) across the 21 federated cohort databases.

7.1.3. Data augmentation

7.1.3.1. Case Study 1 – Small-scale data augmentation

The scope of this study is to enhance the performance of the existing lymphoma classification models in pSS through data augmentation on a single European cohort. To this end, we acquired an anonymized dataset which consists of 449 patients who have been diagnosed with primary Sjögren's Syndrome (pSS) at the University of Athens (UoA) cohort. The number of lymphoma pSS patients was 70 with an average age 48.77 (±12.54) whereas the number of controls was 140 with an average age 52.47 (±13.86). There were 162 features, including demographics, medical conditions (e.g.,

dry eyes), and laboratory measures (e.g., C3), among others. All clinical data were shared according to the EU GDPR requirements.

The performance of the virtual data generators in the UOA cohort is presented in Table 20 for the tree ensembles and the RBF-based ANNs, while in Table 21 for the Bayesian networks and the log-MVND. The performance of the virtual generation methods was favorable. According to Table 20, the average GOF was 0.021 for the unsupervised tree ensembles, 0.022 for the supervised tree ensembles, 0.068 for the RBF-based ANNs, 0.37 for the Bayesian networks and 0.133 for the Log-MVND. In addition, the average KL-divergence was 0.0289 for the unsupervised tree ensembles, 0.034 for the supervised tree ensembles, 0.033 for the RBF-based ANNs, 5e-05 for the Bayesian networks and 0.085 for the Log-MVND. The unsupervised tree ensembles generated virtual distributions with high similarity and convergence with the real data.

Table 20. Summary of the average performance evaluation measures for assessing the quality of the virtual data generated by each virtual population generation method for the pSS domain.

Virtual population generation method	**Quality of the virtual data**		
	GOF	**KL-divergence**	**Correlation coefficient**
Unsupervised tree ensembles	0.021	0.0289	0.1 ± 0.22
Supervised tree ensembles	0.022	0.034	0.102 ± 0.23
Supervised RBF-based ANNs	0.068	0.033	0.103 ± 0.23
Bayesian networks	0.37	0.000005	0.06 ± 0.07
Log-MVND	0.133	0.085	0.5 ± 0.47

The absolute correlation difference between the real and virtual data by the unsupervised tree ensembles is depicted in Figure 42, with an average correlation difference 0.1 ± 0.22. The white horizontal and vertical lines in the features “Renal disease” and “Kidney infiltrates” denote the existence of strong correlation differences. This occurs because only 4 patients had positive Renal disease while only 7 patients had positive kidney infiltrates among the 449 patients and thus the virtual distributions included only negative samples. The average correlation difference was 0.102 ± 0.23 for the supervised tree ensembles, 0.103 ± 0.23 for the RBF-based ANNs, 0.5 ± 0.47 for the Log-MVND, and 0.06 ± 0.07 for the Bayesian networks. The latter had the smallest correlation difference but lower GOF values than the unsupervised tree ensembles.

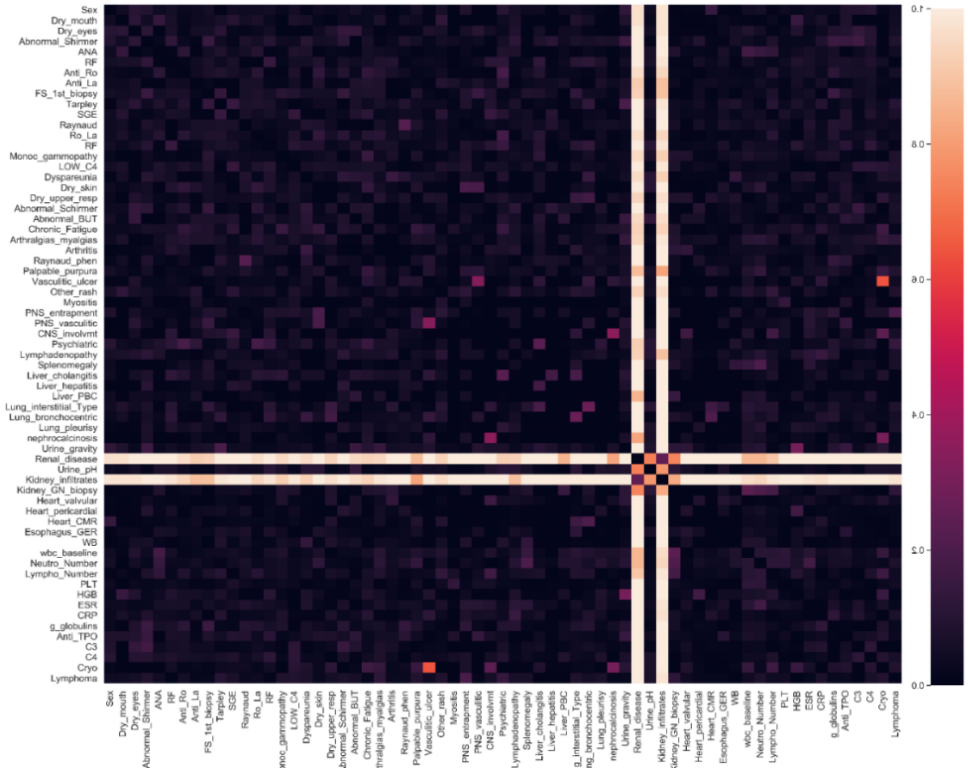


Figure 42. The absolute difference between the real and virtual correlation matrices for the UoA dataset, in the case of the unsupervised tree ensembles generator. The features are ordered according to their appearance in Supplementary ***Table 1***. Values with dark and purple color denote low variations among the real and virtual data whereas values with orange/white color denote otherwise.

The application of the XGBoost on the real data yielded: accuracy 0.724; sensitivity 0.679; specificity 0.814; AUC 0.802. On the other hand, according to Table 21, the average performance of the XGBoost on the aggregated real and virtual data from the unsupervised tree ensembles achieved the best classification performance, yielding accuracy 0.833, sensitivity 0.786, specificity 0.929, and AUC 0.924. The performance of the XGBoost using the augmented data from the supervised tree ensembles, and the supervised RBF-based ANNs come next. Finally, the performance of the XGBoost using the augmented data from the Log-MVND, and the Bayesian networks was lower than in the previous case (using the real data only).

In a similar manner, the performance of the lymphoma classification models from the AdaBoost and Random Forests using the augmented data from the tree ensembles was higher than in the case of the real data. The application of the AdaBoost on the real data yielded accuracy 0.719, sensitivity 0.675, specificity 0.807, AUC 0.749. On the other

hand, according to Table 21, the average performance of the AdaBoost on the aggregated real and virtual data from the unsupervised tree ensembles achieved better classification performance, yielding accuracy 0.79, sensitivity 0.732, specificity 0.907, and AUC 0.814. In the case of the Random Forests, the application on the real data yielded: accuracy 0.729, sensitivity 0.657, specificity 0.871, AUC 0.81. On the other hand, according to Table 21, the average performance of the Random Forests on the aggregated real and virtual data from the unsupervised tree ensembles achieved better classification performance, yielding accuracy 0.824, sensitivity 0.746, specificity 0.979, and AUC 0.922.

Table 21. A summary of the lymphoma classification results from the XGBoost, AdaBoost and Random Forests before and after data augmentation using the virtual data from each generator.

Virtual population generation method for data augmentation	**Lymphoma classification performance**			
	accuracy	**sensitivity**	**specificity**	**AUC**
XGBoost				
Before data augmentation	0.724	0.679	0.814	0.802
Unsupervised tree ensembles	0.833	0.786	0.929	0.924
Supervised tree ensembles	0.814	0.757	0.929	0.912
Supervised RBF-based ANNs	0.819	0.764	0.929	0.914
Bayesian networks	0.752	0.707	0.843	0.787
Log-MVND	0.8	0.754	0.893	0.824
AdaBoost				
Before data augmentation	0.719	0.675	0.807	0.749
Unsupervised tree ensembles	0.79	0.732	0.907	0.814
Supervised tree ensembles	0.79	0.725	0.921	0.82
Supervised RBF-based ANNs	0.824	0.764	0.943	0.87
Bayesian networks	0.69	0.593	0.886	0.76
Log-MVND	0.767	0.696	0.907	0.784
Random Forests				
Before data augmentation	0.729	0.657	0.871	0.81
Unsupervised tree ensembles	0.824	0.746	0.979	0.922
Supervised tree ensembles	0.767	0.661	0.979	0.877
Supervised RBF-based ANNs	0.757	0.636	1	0.901
Bayesian networks	0.762	0.661	0.964	0.839
Log-MVND	0.757	0.668	0.936	0.852

The ROC curves are shown in Figure 43, highlighting the performance of the unsupervised tree ensembles (increase by 10.9% in the accuracy, 10.7% in sensitivity, 11.5% in specificity, and 12.2% in AUC) in the case of the XGBoost which suggests a notable performance enhancement. A similar increase is also observed in the case of the AdaBoost (7.1% in the accuracy, 5.7% in sensitivity, 10% in specificity, and 6.5% in AUC), as well as, in the case of the Random Forests (9.5% in the accuracy, 8.9% sensitivity, 10.8% in specificity, and 11.2% in AUC).

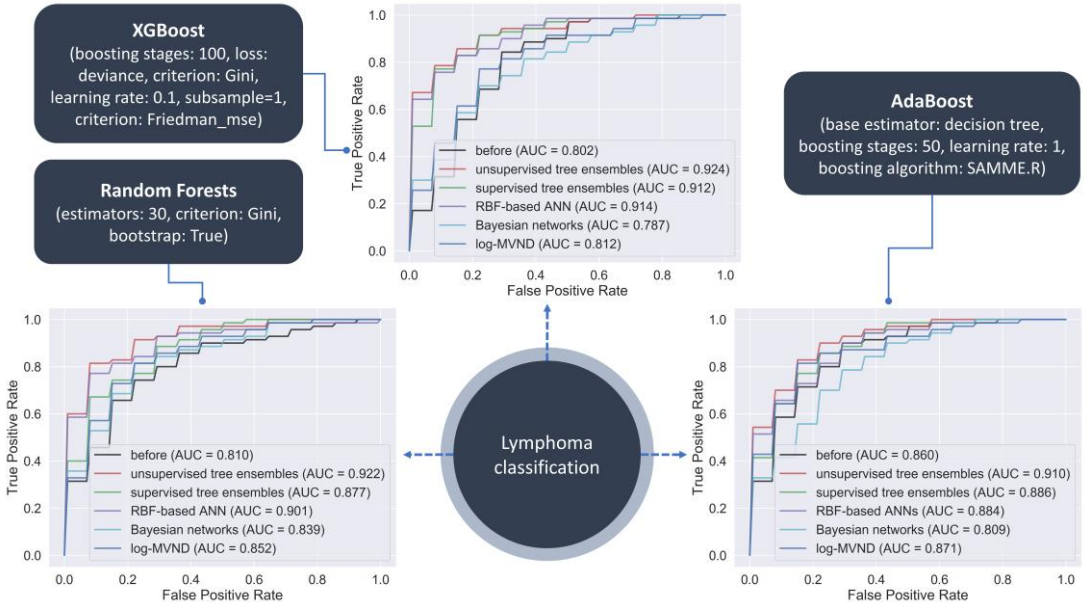


Figure 43. ROC curves depicting the classification performance of the XGBoost, the AdaBoost and the Random Forests for lymphoma classification with and without data augmentation.

Our results highlight the favorable performance of the tree ensembles towards the generation of high-quality virtual data with goodness-of-fit (GOF) 0.021 and Kullback Leibler (KL)-divergence 0.029 in the pSS domain. The aggregation of the real and the virtual data from the tree ensembles revealed a notable increase in the classification accuracy, sensitivity, and specificity for lymphoma classification, where the XGBoost yielded an increase by 10.9% in accuracy, 10.7% in sensitivity, and 11.5% in specificity. A similar increase is also observed in the case of the AdaBoost for lymphoma classification (5.5% in the accuracy, 5.3% in sensitivity, 6.3% in specificity, and 10.1% in AUC), as well as, in the case of the Random Forests for lymphoma classification (9.4% in accuracy, 10.1% in sensitivity, 7.2% in specificity, and 12.2% in AUC). The outcomes of the proposed pipeline are promising since the existing lack of population size in pSS obscures the development of robust disease classification and

risk stratification models. To our knowledge, this is the first computational pipeline which aggregates high-quality virtual with real curated clinical data to address crucial clinical unmet needs in pSS.

7.1.3.2. Case Study 2 – Large-scale data augmentation

To enhance the performance of the existing lymphoma classification models in pSS through data augmentation across 10 European cohorts. Anonymized clinical data were collected from 10 databases with patients who have been diagnosed with primary Sjögren's Syndrome (pSS) under the HarmonicSS Project [343]. The 10 databases included 316 lymphoma patients (targets) and 4692 non-lymphoma patients (controls). The density forest ensembles were applied on each dataset to augment the real population yielding 10,016 high-quality virtual patients (586 targets, 9430 controls), in total, with average gof 0.01, KL divergence less than 0.001, and correlation difference 0.02. The distributed learning pipeline was then utilized, using the hybrid loss function (Figure 44), where the steepness of the logcosh and the wideness of the modified Huber loss were combined for different values. The value was defined as in the proposed distributed MART with dropouts, where r is the dropout rate.

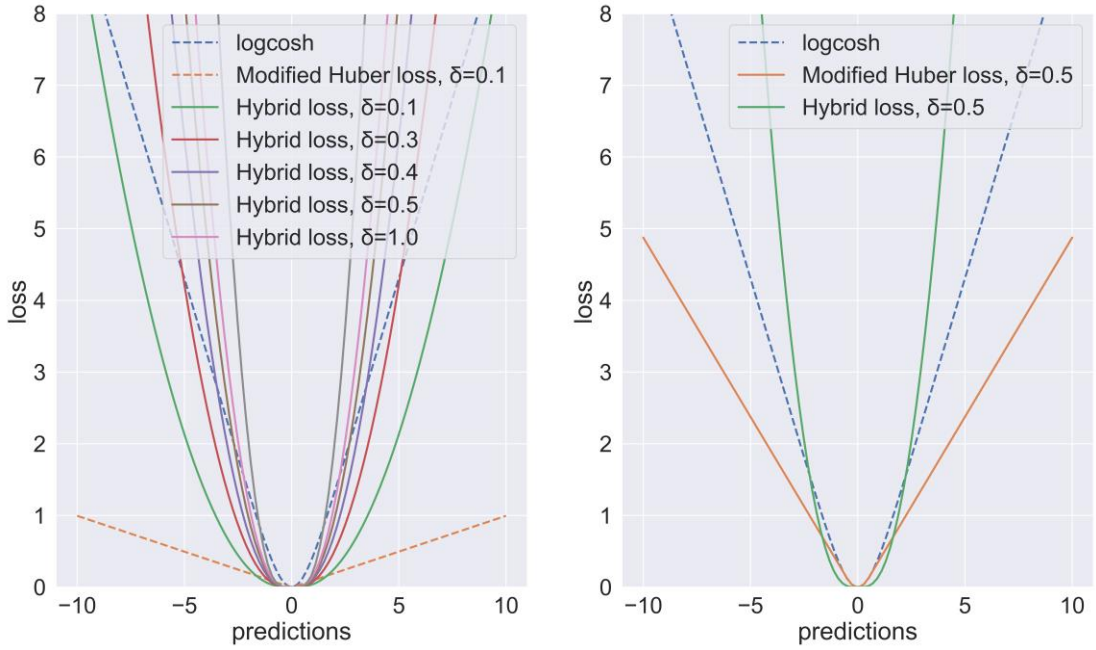


Figure 44. Distribution of the hybrid loss function compared to the modified Huber loss and the logcosh for different δ values.

In the real case, the 9 datasets having the highest number of targets were used for distributed training (300 targets and 4,411 controls, in total), whereas in the data augmentation case, the 9 training datasets included 9,422 patients (546 targets, 8,876 controls), in total. In both cases, the remaining (real) dataset was used for testing (16 targets, 281 controls). Random down-sampling with replacement was also applied on each case for class imbalance handling.

The overall performance of the distributed algorithms was better on the augmented data, where the distributed MART achieved accuracy 0.852, sensitivity 0.833 and specificity 0.854 against the one trained on the real data with accuracy 0.808, sensitivity 0.722 and specificity 0.818. A notable increase was observed in the case of the proposed distributed MART with δ = 0.4 (r = 0.3) which achieved accuracy 0.865, sensitivity 0.84, and specificity 0.868 whereas in the real case the algorithm achieved accuracy 0.791, sensitivity 0.772, and specificity 0.794.

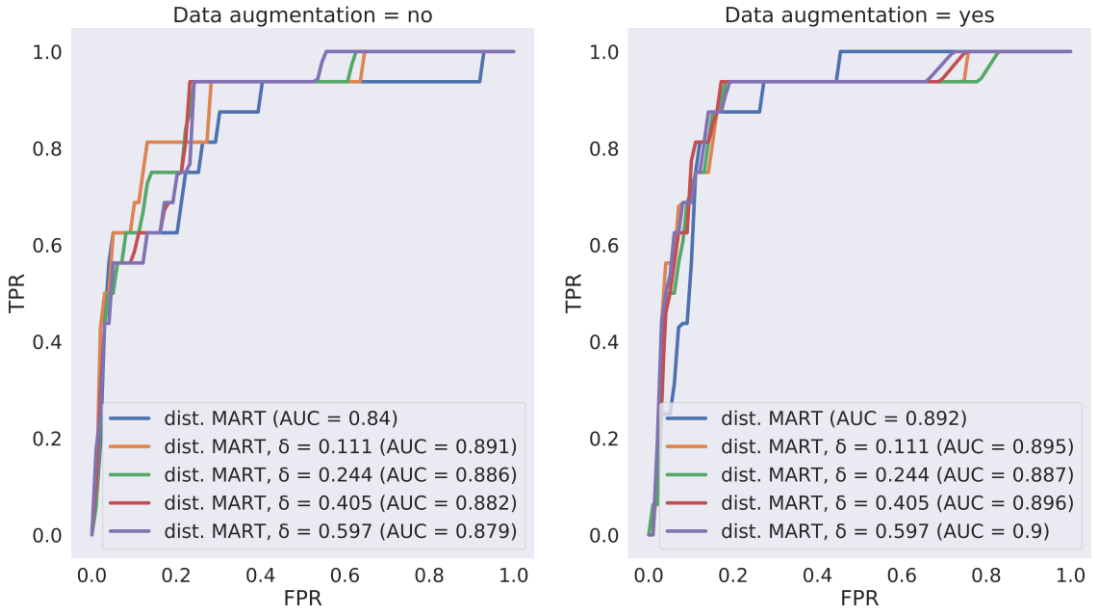


Figure 45. Receiver Operating Characteristic (ROC) curves for distributed classification with and without augmentation.

A similar increase occurs for δ = 0.6 (r = 0.4) with accuracy 0.862, sensitivity 0.868, and specificity 0.861 against the real case where the algorithm achieved accuracy 0.835, sensitivity 0.854, and specificity 0.833. According to Figure 45, the area under the curve scores in the distributed MART yielded an average increase by 5.2%, as well as, by 1.4% in the proposed distributed MART with δ = 0.4 and 2.1% with δ = 0.6. The

positive impact of data augmentation is also reflected by the detection error tradeoff (DET) curves which are depicted in Figure 46 in logarithmic scale. The DET score was defined as the median absolute ratio of the false positive rate over the false negative rate. According to Figure 46, an average decrease by 2.6% in the DET score is observed in the proposed distributed MART with δ = 0.4 (r = 0.3) and 4.5% with δ = 0.6 (r = 0.4). In this work, we presented a pipeline for additive training across augmented and harmonized clinical data in distributed environments through the utilization of distributed multiple additive regression trees (MART) with a hybrid loss. The pipeline includes data pre-processing routines for the precise detection of data anomalies, as well as, features with joint variability. Both flexible and stringent lexical analysis were applied to detect terminologies with increased coherence among the distributed data. Density forest ensembles were finally developed for the generation of high-quality virtual distributions which were used for data augmentation.

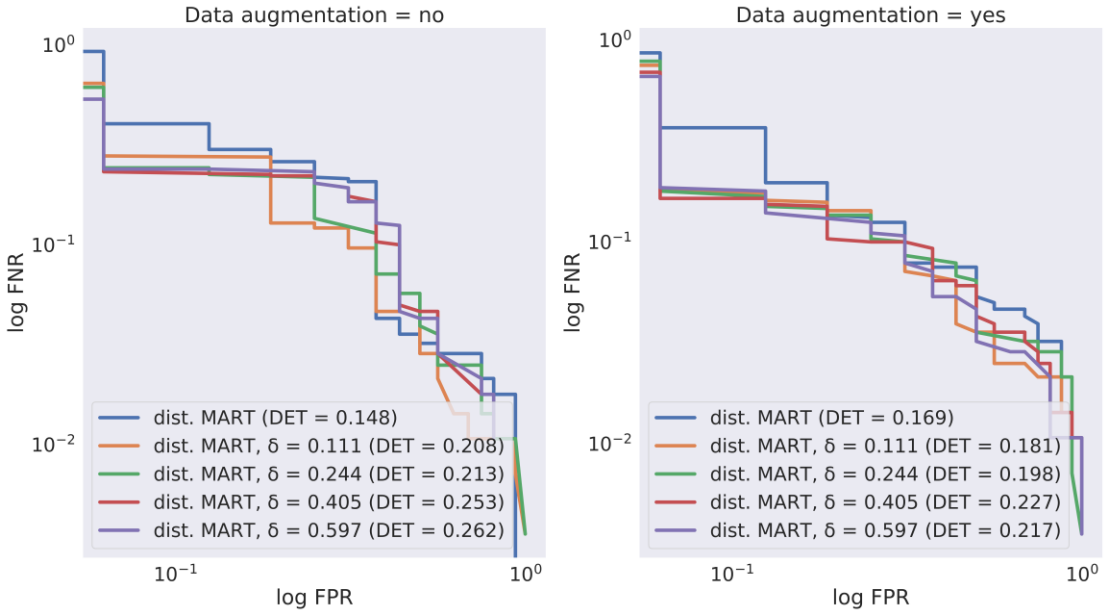


Figure 46. Detection error tradeoff (DET) curves for distributed classification with and without augmentation.

The density forest ensembles were able to generate virtual data for data augmentation with decreased divergence with the real data (average gof 0.01, KL divergence less than 0.001, and correlation difference 0.02). The proposed pipeline was able to yield robust distributed learning models from the augmented data with an average increase by 6.8% in sensitivity, and 10.4% in specificity for δ = 0.4. The proposed loss function avoids

overfitting effects which are caused by the early inclusion of regression trees in the ensemble. To our knowledge, this is the first case study which combines data augmentation and distributed regression tree ensembles with hybrid loss yielding robust disease classification models through a case study in autoimmune diseases.

7.1.4. Federated/distributed learning

7.1.4.1. Case Study 1 – Incremental learning

The scope of this study is to validate an incremental learning framework across multiple distributed subsets of a single cohort in pSS towards lymphoma classification. To demonstrate the efficacy of the presented framework we acquired anonymized clinical data from the University of Athens (UoA) cohort which consists of patients that have been diagnosed with primary Sjögren's Syndrome (pSS), with the purpose of developing a binary lymphoma prediction model. Three datasets, assume, A, B, and C, were used for training purposes and four datasets, assume, $T1$, $T2$, $T3$, and $T4$ were used for validation.

Each training dataset consists of 70 pSS patients (50 non-lymphoma; 20 lymphoma) whereas each testing dataset consists of 24 pSS patients (20 non-lymphoma; 4 lymphoma). Each dataset includes a set of 68 pSS-related features [135]. For demonstration purposes and considering the imbalance between the two groups, we worked on the extreme gradient boosting [357] classifier to construct an optimized, distributed supervised learning model for predicting binary lymphoma outcomes (i.e., "0": no lymphoma, "1": lymphoma) based on tree ensembles. The simple regression models (Section 6.3.2.1), the Multinomial Naïve Bayes (Section 6.3.2.3) and the neural networks (Section 6.3.2.4) were also constructed for comparison purposes. Regarding the gradient boosting trees, the maximum depth was set to 6 levels using a step size value of 0.3 to prevent overfitting along with a binary logistic loss function as the learning objective. The simple regression models, the neural networks, and the Multinomial Naïve Bayes, were developed based on the specifications that have been described in 6.3.2.1, 6.3.2.3 and 6.3.2.2, respectively.

The initial prediction model was trained on dataset A, updated on B, and re-updated on C, yielding the final one. For each supervised learning configuration, the lymphoma prediction model was evaluated on each testing dataset.

According to Figure 47, the gradient boosting trees exhibit the highest performance with an average AUC score 0.94 across the four testing datasets along with an accuracy 0.916 and sensitivity 0.875. The performance of the SGD-based methods of Section 6.3.2.1 and 6.3.2.2 was poor (< 60% accuracy) since these methods are unable to deal with the class imbalance and the associations between the features during the training stage. The same issue was observed for the neural networks (due to the small-scale datasets) and the NB model.

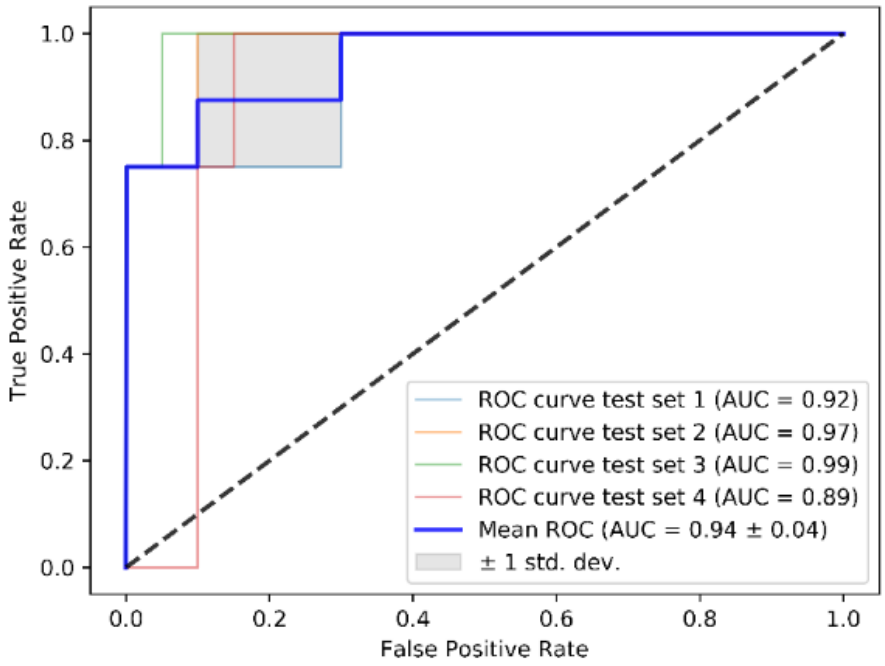


Figure 47. Prediction performance across the four testing datasets using the GBT.

We took advantage of the mathematical basis of incremental learning to deploy a computationally efficient and secure strategy for disease prediction modeling across data that are stored in multiple sites by assuming that: (i) the clinical data are stored in private cloud spaces and are harmonized through semantic interlinking methods, and (ii) the model for predicting disease outcomes is continuously adjusted to the data that lie across private cloud spaces. A case study is conducted to highlight the applicability of the framework by constructing a distributed binary lymphoma prediction model across private cloud spaces consisting of clinical data from patients with primary Sjögren's Syndrome (pSS); a chronic autoimmune disease that exhibits salivary gland dysfunction, with 5% of the pSS patients being prone to lymphoma development. Our results demonstrate the superiority of the distributed trees towards the precise

discrimination of lymphoma cases, against the conventional methods, with an average accuracy 91.6%, area under the curve score 94%, and sensitivity 87.5%.

7.1.4.2. Case Study 2 – Distributed learning

The scope of this study is to validate a distributed learning framework across four European cohorts in pSS towards lymphoma classification. We acquired anonymized clinical data from four European cohorts on primary Sjögren's Syndrome (University of Athens (UoA); Harokopion University (HUA); University of Pisa (UNIPI); University of Udine (AOUD)). The cohort data (Table 22) were shared with the platform under the data protection agreement version 3.7 as of August 2018 according to the Article 35 (3) (b) of the GDPR fulfilling all the necessary ethical and legal requirements for data sharing. To deal with the high imbalance between the lymphoma and non-lymphoma groups (Table 22), the number of controls was set as twice the number of lymphoma cases per training cohort by downsampling the majority class so that a 1:2 ratio, with age and sex matched controls, is maintained between the case group and the control group (Table 22).

Table 22. Demographic information.

Features	**UoA**	**HUA**	**UNIPI**	**AOUD**
Age at diagnosis	53.5±13.5	47.5±12.3	51.4±14	52.4±13.9
Gender ratio (females/males)	415/25	96/3	693/25	274/23
Lymphoma/Non-lymphoma	76/364	6/93	31/687	26/271
Total number of patients	440	99	718	297

Due to the small population and to take advantage of the statistical power of the whole population on each cohort, the incremental learning process was repeated 5 times using different subsets of controls for the training process each time.

Table 23. Overall population characteristics for distributed lymphoma prediction.

Descriptive statistics	**Cohorts**			
	Set of training cohorts			**Testing cohort**
	AOUD	**UoA**	**UNIPI**	**HUA**
Number of lymphoma cases	26	76	31	6
Number of controls *	52	152	62	93
Total population	78	228	93	99
* the number of controls was randomly selected 5 times to "cover" each cohort's population.				

The incremental learning workflow was applied on the harmonized cohort data to develop a distributed lymphoma prediction model using three cohorts for training and one cohort for testing. In order to make the analysis complete, each cohort was included in the testing process by repeating the process four times. For each combination, the XGBoost, Support Vector Machines, Logistic regression, Multinomial Naïve Bayes, and Multi-layer Perceptron algorithms were applied in an incremental manner. Then, the optimal combination, i.e., the one with the highest performance in all five algorithms was selected for demonstration purposes, according to which the AOUD, UoA, and UNIPI cohorts were used for training and the HUA cohort for testing. The results are depicted in Table 24. The lymphoma presence was set as the target to solve a binary classification problem.

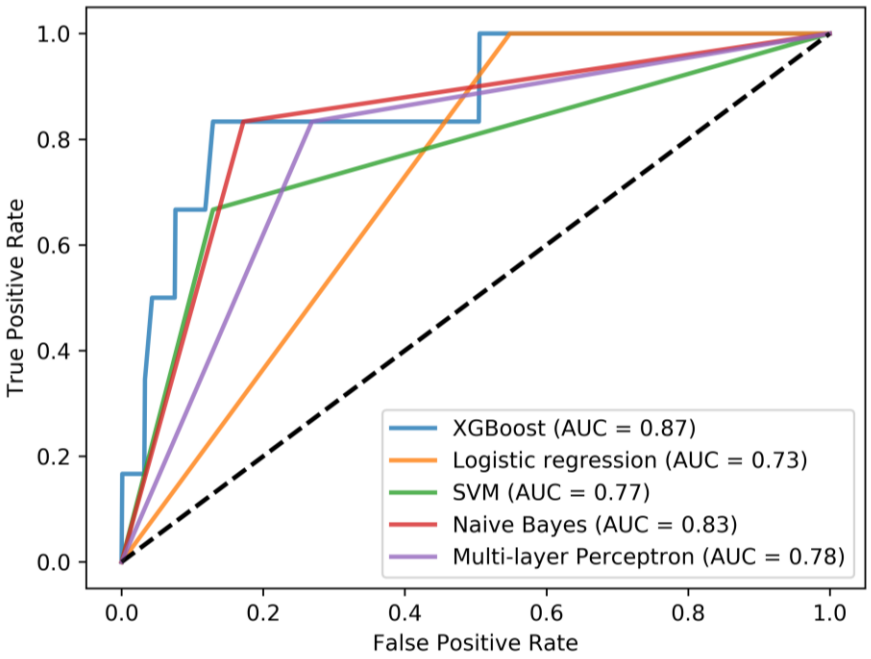


Figure 48. ROC curves for each incremental learning algorithm.

The performance evaluation measures include the accuracy, sensitivity, specificity, and AUC scores, which are depicted in Table 24, averaged across 5 runs. According to Table 24, the XGBoost algorithm (AUC 0.871, accuracy 0.859, sensitivity 0.833, specificity 0.86) outperforms the rest of the incremental learning algorithms. The Multinomial Naïve Bayes has the second-best performance along with the Multi-layer Perceptron. The performance of the Logistic regression and the Support Vector

Machines was significantly low in terms of sensitivity. The ROC curves per incremental learning algorithm are depicted in Figure 49, which confirm XGBoost's superiority.

Table 24. Performance evaluation scores per incremental learning algorithm and testing cohort combination.

Algorithm	Testing cohort	Accuracy	Sensitivity	Specificity	AUC
XGBoost	AOUD	0.835	0.692	0.849	0.849
	UoA	0.834	0.526	0.898	0.898
	UNIPI	0.872	0.484	0.889	0.889
	HUA	0.859	0.833	0.860	0.871
Logistic regression	AOUD	0.926	0.462	0.970	0.970
	UoA	0.805	0.184	0.934	0.934
	UNIPI	0.85	0.290	0.875	0.875
	HUA	0.899	0.167	0.946	0.556
Support Vector Machines	AOUD	0.902	0.692	0.923	0.923
	UoA	0.714	0.684	0.720	0.720
	UNIPI	0.948	0.032	0.990	0.990
	HUA	0.768	0.667	0.774	0.720
Multinomial Naïve Bayes	AOUD	0.869	0.846	0.871	0.871
	UoA	0.823	0.500	0.890	0.890
	UNIPI	0.799	0.484	0.814	0.814
	HUA	0.828	0.833	0.828	0.831
Multilayer Perceptron	AOUD	0.875	0.808	0.882	0.882
	UoA	0.839	0.263	0.959	0.959
	UNIPI	0.879	0.387	0.901	0.901
	HUA	0.747	0.833	0.742	0.788

To further enhance the clinical findings of the case study we have induced the decision tree from the XGBoost schema which includes the features that highly participated in the decision-making process (Figure 49). The features with the highest contribution across the splits are represented by a node along with the decision rules, and the rule outcomes (i.e., "yes/no") are depicted as branches. At the first level lies the "C4" as the root node. The features "lymphadenopathy" and "salivary gland swelling" come next along with the "Anti-La", and "gender". The leaf values on each branch denote the conditional probability of a data point falling in class 1 on that branch. We extended two previous studies [135], [250] through the curation, subsequent harmonization and

federated analysis of three European cohort data (instead of a single cohort dataset) to deal with open issues and clinical unmet needs in the domain of primary Sjögren's Syndrome (pSS). We combine lexical and ontology matching to detect common terms among the cohort data according to a pre-defined reference ontology. Then, we apply a federated learning pipeline to develop a lymphomagenesis prediction model across the cohort data to avoid physical data integration by storing the data in private cloud databases. Our results confirm the dominance of the federated XGBoost schema with accuracy 0.848, sensitivity 0.833, specificity 0.849, and area under the curve 0.868, along with a prominent pathway that is induced by the decision tree highlighting four prominent features and one prominent combination for decision-making.

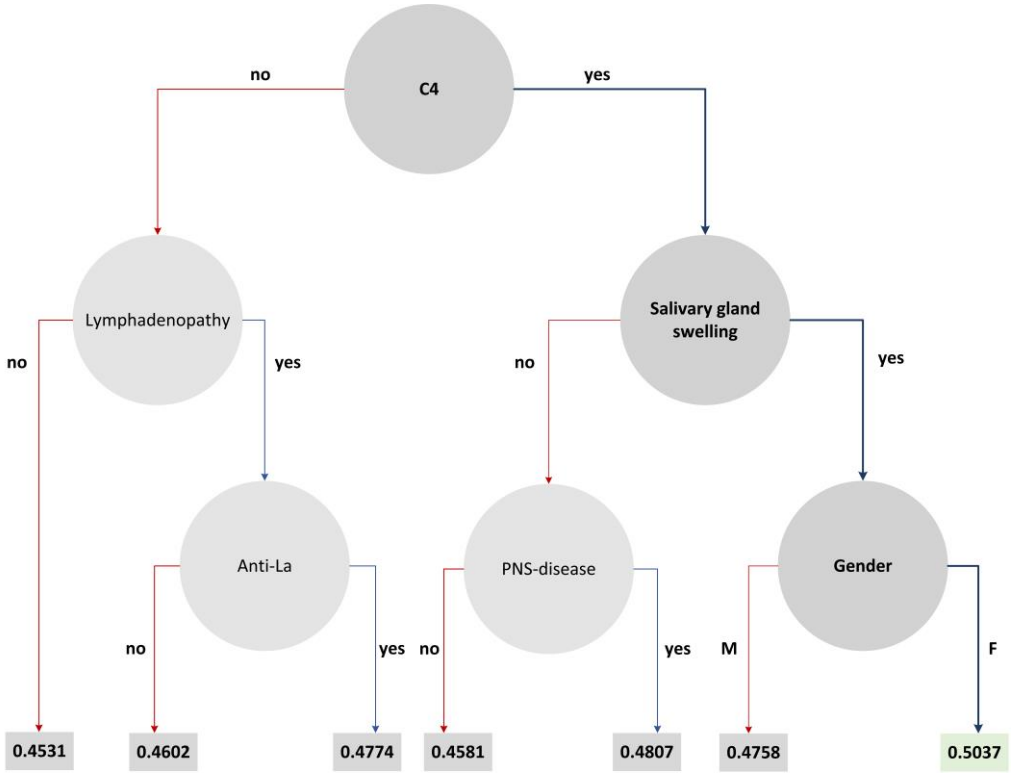


Figure 49. The decision tree that is induced by the XGBoost schema.

7.1.4.3. Case Study 3 – Federated learning across 21 European cohorts

The scope of this study is to validate a federated learning framework across 21 European cohorts in pSS to address the unmet needs (lymphomagenesis modeling, biomarkers). A summary of the overall demographic information from the 21 European databases on pSS is presented in Table 25. The total number of eligible patients who fulfilled the inclusion criteria was 7,156, where the gender information was recorded

for 7,000 patients (6,512 females, 488 males with a female to male ratio 13.34%). The average age at SS diagnosis in the female group was 51.82 (±13.96) years whereas in the male group the average age was 54.24 (±13.77) years.

Table 25. Demographic information.

Demographics	**Females**	**Males**
Gender	6,512	488
Age at SS diagnosis (mean±std)	51.82 (±13.96) years	54.24 (±13.77) years
Disease duration (mean)	7.08 years	5.59 years
Female to male ratio	13.34%	

The lymphoma types include the B-cell Mucosa-associated Lymphoid Tissue (MALT) Lymphoma, the Diffuse Large B-cell Lymphoma (DLBCL), the B-cell Nodal Marginal Zone Lymphoma (NMZL), the B-cell Splenic Marginal Zone Lymphoma (SMZL), and other mature B-cell neoplasms. These lymphoma types were merged into a single lymphoma type with 354 positive lymphoma patients and 6,802 non-lymphoma (or missing) patients (lymphoma to non-lymphoma ratio 5.2%). The lymphoma distribution per cohort is summarized in Table 26.

Table 26. Distribution of lymphoma and non-lymphoma patients per cohort.

Cohort acronym	**Cohort full name**	**Number of lymphoma patients**	**Number of non-lymphoma (or missing) patients**
IDIBAPS	Consorci Institut D'Investigacions Biomediques August Pi I Sunyer	0	300
UNIPG	Università degli Studi di Perugia	10	166
UPSUd PARIS	Université Paris-Sud (database 1)	24	483
UoB	University of Birmingham	3	156
UNIVAQ	Università degli Studi dell'Aquila	3	97
ULB	Université libre de Bruxelles	1	726
HUA	Harokopion University of Athens	8	151
UMCG	University Medical Center Groningen	20	166
UiB	University of Bergen	3	138
UOI	University of Ioannina	7	279
UU	Utrecht University	14	108
UNIRO	Universita' Degli Studi Di Roma La Sapienza	14	532

Cohort acronym	Cohort full name	Number of lymphoma patients	Number of non-lymphoma (or missing) patients
QMUL	Queen Mary University of London	1	47
UMCU	Universitair Medisch Centrum Utrecht	27	313
MHH	Medizinische Hochschule Hannover	5	178
UNIPI	Universita di Pisa	31	687
CUMB	Charité – Universitätsmedizin Berlin	0	71
UBO	Université de Bretagne Occidentale	4	77
UOA	National and Kapodistrian University of Athens	101	488
AOUD	Azienda Sanitaria Universitaria Integrata di Udine	16	281
UNEW	University of Newcastle	62	1358

Data curation was applied on each individual cohort database to automatically remove outliers, data inconsistencies and duplicated fields. The LOF algorithm was combined with the Isolation Forests to track down and remove outliers with 90% accuracy and the Spearman correlation coefficient was combined with the Jaro distance score to detect duplicated features. Data imputation was applied only to features with less than 30% missing values upon approval from the clinical experts. Upon the completion of the cohort data curation process, ontologies were constructed for each curated cohort database based on the extracted metadata. Semantic mapping rules were defined between the individual ontologies and the pSS reference ontology. As shown on Table 27, the cohort data harmonization process resulted in 48 common concepts (or terminological concepts) which constitute the pSS minimal criteria (minimal common data elements) across the 21 federated cohort databases.

Table 27. Set of features which represent the minimal criteria of the pSS domain knowledge.

Feature	Presence/ Abnormal	Absence/ Normal	Mean	Median
Gender	6879 (females)	513 (males)	-	1
Age at SS diagnosis	-	-	51,95	53
Disease duration	-	-	7,23	6
Dry Mouth (aka Xerostomia)	6101	699	-	1
Dry Eyes	6046	748	-	1
Parotid or Submandibular swelling	1897	3186	-	0

Feature	Presence/ Abnormal	Absence/ Normal	Mean	Median
Parotid Gland swelling	1616	2227	-	0
Submandibular salivary gland swelling	139	2648	-	0
Raynaud’s Phenomenon	1577	4365	-	0
Fatigue	2840	2416	-	1
Arthritis	993	5006	-	0
Renal Disease	162	6021	-	0
Tubulointerstitial Nephritis	66	4663	-	0
Glomerulopathy	37	4293	-	0
Membranoproliferative Glomerulonephritis (MPGN)	15	4309	-	0
Membranous Glomerulonephritis (MGN)	3	4163	-	0
Mesangioproliferative Glomerulonephritis (MPGN)	9	4157	-	0
Other Glomerulonephritis	4	4501	-	0
Pulmonary Disease	415	5421	-	0
Small Airway Disease	157	5012	-	0
Lymphocytic Interstitial Pneumonia (LIP)	47	4414	-	0
Nonspecific Interstitial Pneumonia (NSIP)	34	4034	-	0
Usual Interstitial Pneumonia (UIP)	31	4041	-	0
Cryptogenic Organizing Pneumonia (COP)	0	4077	-	0
Liver Disease	131	5300	-	0
Autoimmune Hepatitis (AIH)	40	4682	-	0
Primary Biliary Cholangitis (PBC)	81	5535	-	0
Sclerosing cholangitis	11	4639	-	0
Nervous System Disease	560	5577	-	0
Peripheral Nervous System Disease	267	4712	-	0
Central Nervous System Disease (CNS)	125	4487	-	0
PalpablePurpura	396	5257	-	0
CutaneousDisease	458	4158	-	0
Muscular System Disease	357	4633	-	0
IdiopathicInflammatoryMyopathy (IIM)	10	3842	-	0
Inclusion Body Myositis (IBM) documented with Biopsy	150	4429	-	0
B-cell Mucosa-associated Lymphoid Tissue (MALT) Lymphoma	245	5324	-	0
Diffuse Large B-cell Lymphoma (DLBCL)	45	5326	-	0
B-cell Nodal Marginal Zone Lymphoma (NMZL)	24	5346	-	0
B-cell Splenic Marginal Zone Lymphoma (SMZL)	6	4789	-	0
Other mature B-cell neoplasms	21	5305		0

Feature	Presence/ Abnormal	Absence/ Normal	Mean	Median
Anti-La-SSB [presence]	2670	3499	--	0
Anti-Ro-SSA [presence]	4565	1703	-	1
Rheumatoid Factor (RF) [Units-volume]	2282	2362	-	0
Antinuclear Antibodies (ANA) [presence]	4354	1096	-	1
C4 levels (Serum complement) [Mass-volume]	2485	1725	-	1
Cryoglobulins [presence]	266	4409	-	0
Lymphoma*	354	5653	-	0
* The records of patients with missing lymphoma status were ignored from the analysis.				

According to Table 26, the lymphoma over non-lymphoma ratio was 5.2% which implies a significant population imbalance. To deal with this, random downsampling with replacement was applied on each individual training cohort database among the lymphoma (target group) and the non-lymphoma (control group) patients. The process was repeated ten times to avoid biases during the downsampling process. On each iteration, the downsampled control group was matched with the target group according to the age, gender, and disease duration using a ratio 1:1 to yield equally balanced populations. The Wilcoxon Mann-Whitney rank-sum test was used to evaluate whether the distributions of the age and disease duration did not significantly deviate between the target group and the downsampled control group whereas the chi-square test was used for gender matching. The classification performance of the federated AI models was assessed based on the accuracy, sensitivity, specificity, and area under the ROC curve (AUC).

Four large scale federated lymphoma classification scenarios were conducted; three scenarios including a common set of training harmonized cohort databases and three different testing databases, as well as, one scenario with a different set of training databases and a single testing database. The training set in federated scenarios 1-3 is {UOA, UNIPI, UNEW, UNIPG, PARIS, UoB, UNIVAQ, HUA, UOI, UU, UNIRO, UMCU, MHH, UBO} and the testing set is {AOUD (scenario 1), UNIPG (scenario 2), HUA (scenario 3)} whereas the training set in federated scenario 4 is {AOUD, UOA, UNIPI, UNIPG, UNEW, PARIS, UoB, UNIVAQ, UOI, UU, UNIRO, UMCU, MHH, UBO, UMCU} and the testing set is HUA. According to Table 28, the federated tree ensembles achieved better performance against the FSGD-based methods, such as, the

FMNB and the FMLP, since the latter focus on the direct update of the weights of a linear loss function, without controlling for overfitting effects, their performance tends to be lower than in the case of the federated tree ensembles which utilize boosting to avoid overfitting.

Table 28. A summary of the performance evaluation results across the four federated scenarios.

Federated learning schema	Performance evaluation metrics			
	Accuracy	**Sensitivity**	**Specificity**	**AUC**
Federated scenario 1				
FGBT	0.84	0.81	0.85	0.89
FDART, rd = 0.1	0.86	0.75	0.87	0.87
FDART, rd = 0.2	0.84	0.62	0.85	0.86
FDART, rd = 0.3	0.83	0.81	0.84	0.89
FDART, rd = 0.4*	0.85	0.81	0.85	0.89
FDART, rd = 0.5	0.83	0.87	0.83	0.88
FMNB	0.51	0.94	0.49	0.71
FMLP	0.64	0.75	0.63	0.69
Federated scenario 2				
FGBT	0.71	0.70	0.71	0.73
FDART, rd = 0.1	0.69	0.70	0.69	0.76
FDART, rd = 0.2*	0.74	0.80	0.73	0.79
FDART, rd = 0.3	0.71	0.70	0.71	0.71
FDART, rd = 0.4	0.71	0.70	0.71	0.75
FDART, rd = 0.5	0.71	0.70	0.71	0.76
FMNB	0.63	0.70	0.63	0.66
FMLP	0.68	0.70	0.68	0.69
Federated scenario 3				
FGBT	0.75	0.99	0.74	0.89
FDART, rd = 0.1*	0.78	0.99	0.76	0.90
FDART, rd = 0.2*	0.78	0.99	0.76	0.91
FDART, rd = 0.3	0.76	0.99	0.74	0.90
FDART, rd = 0.4	0.71	0.87	0.69	0.86
FDART, rd = 0.5	0.74	0.75	0.74	0.86
FMNB	0.71	0.87	0.70	0.79
FMLP	0.85	0.62	0.87	0.74
Federated scenario 4				
FGBT	0.81	0.75	0.81	0.91
FDART, rd = 0.1	0.78	0.87	0.78	0.92

Federated learning schema	Performance evaluation metrics			
	Accuracy	**Sensitivity**	**Specificity**	**AUC**
FDART, rd = 0.2	0.80	0.75	0.80	0.91
FDART, rd = 0.3*	0.80	0.87	0.79	0.91
FDART, rd = 0.4	0.80	0.87	0.80	0.90
FDART, rd = 0.5	0.78	0.75	0.78	0.91
FMNB	0.62	0.87	0.61	0.74
FMLP	0.85	0.62	0.86	0.74
* With light blue color: The federated schema with the best performance, rd: dropout rate.				

According to Figure 50, the ROC curves confirm the favorable performance of the FDART along with the FGBTs, in all cases, where the FDART with dropout rate 0.4 achieved the best performance in federated scenario 1 (accuracy 0.85, sensitivity 0.81, specificity 0.85). Regarding federated scenario 2, the FDART with dropout rate 0.2 achieved the best performance (accuracy 0.74, sensitivity 0.8, specificity 0.73). In federated scenario 4, the FDART with dropout rates 0.1 and 0.2 achieved the best performance (accuracy 0.78, sensitivity1, specificity 0.76) like the FGBT (accuracy 0.75, sensitivity 1, specificity 0.74). In the final scenario, the FDART with dropout rate 0.3 achieved the best performance (accuracy 0.8, sensitivity 0.87, specificity 0.79) yielding better sensitivity than the FGBTs, where the average execution time was 30 seconds for data access and training/testing on each harmonized cohort database. The results of the Shapley additive explanation analysis are depicted in Figure 51 for the FGBT classifier and in Figure 52 for the FDART classifier with dropout rates 0.1-0.5, where the features are ranked based on their positive or negative impact on lymphoma development. Each panel in Figure 51 reflects the mean Shapley value (i.e., the average of the marginal contributions across all permutations) for a feature, in descending order, as well as, whether the impact of a feature has a positive (left) or a negative (right) value for lymphoma development.

In Figure 51, Figure 52, the color in the distribution plots denotes whether the importance of the Shapley value is either low or high and the vertical line corresponds to the base score of the AI model centered around zero. According to Figure 51 and Figure 52, the feature “Parotid or Submandibular swelling” has the highest impact in lymphoma classification, where its absence has a negative predictive value and thus decreases the risk for lymphoma development whereas its presence has a positive predictive value on lymphoma development (i.e., the positive samples shift the ground

truth to the right). Features “Rheumatoid factor”, “Fatigue”, “Age of SS diagnosis”, “Cryoglobulinemia”, and “Disease duration” come next with favorable impact on lymphoma classification. Features “Low C4”, “Palpable purpura”, “Raynaud's phenomenon”, “Arthritis” also appear to be significant in the decision-making process. The importance of these features is also confirmed by the average coverage of each federated tree ensemble algorithm during the lymphoma decision-making process (Supplementary Figure 1).

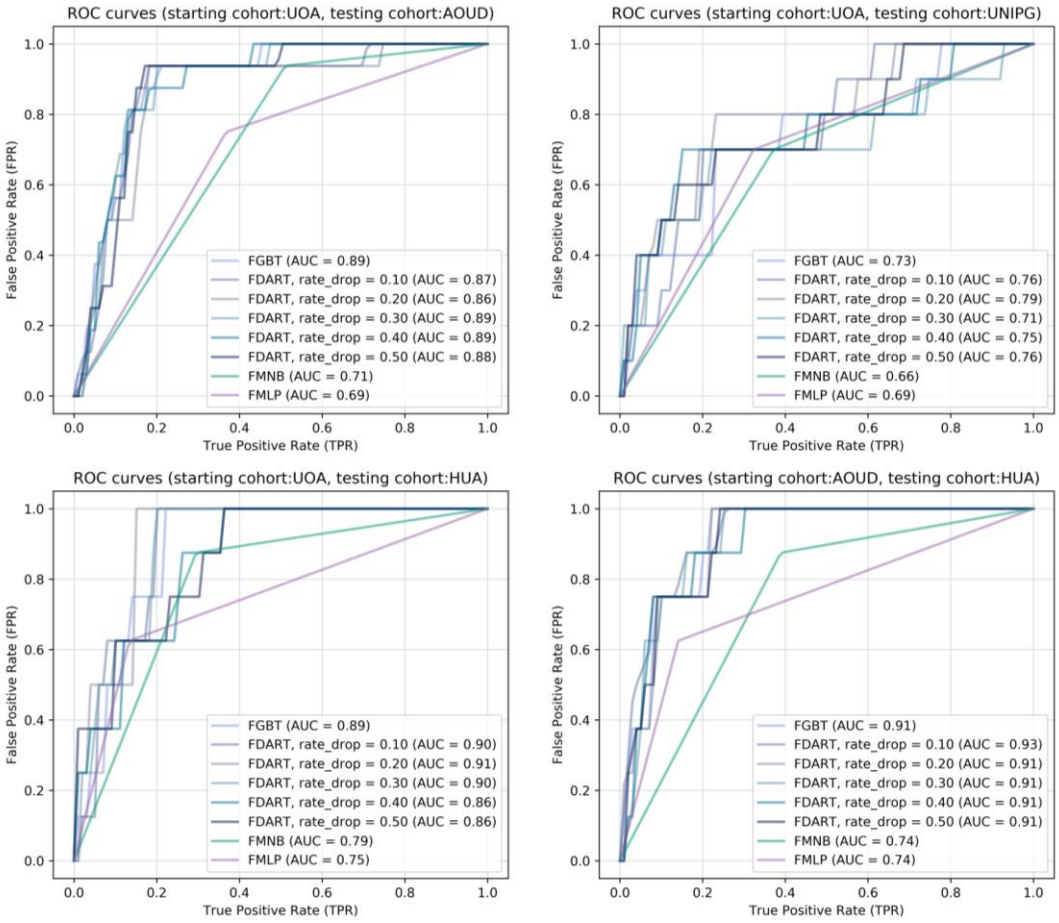


Figure 50. Receiver Operating Characteristic (ROC) curves for each federated algorithm across the two federated scenarios. From top to bottom: on the left for federated scenario 1 with testing cohorts AOUD, UNIPG, HUA and for federated scenario 2 testing cohort HUA.

The Shapley explanation analysis results for the federated learning scenarios 2, 3, and 4 are depicted in Supplementary Figure 2, Supplementary Figure 3, Supplementary Figure 4, Supplementary Figure 5, Supplementary Figure 6, and Supplementary Figure 7 for the FDART and FGBT, respectively. According to Figure 51 and Figure 52, the features “Parotid or Submandibular swelling”, “Rheumatoid factor”,

"Cryoglobulinemia", "Age at SS diagnosis", "Fatigue", and "Low C4" appear to be prominent for lymphoma classification. In all cases, patients with parotid or submandibular swelling, rheumatoid factor, cryoglobulinemia, fatigue and Low C4 tend to have higher impact for lymphoma development since the positive samples shift the ground truth to the right, thus yielding a positive contribution to lymphoma development. The same effect occurs in the case where the pSS patients exhibit palpable purpura, Raynaud's phenomenon, and arthritis, as well (Figure 51).

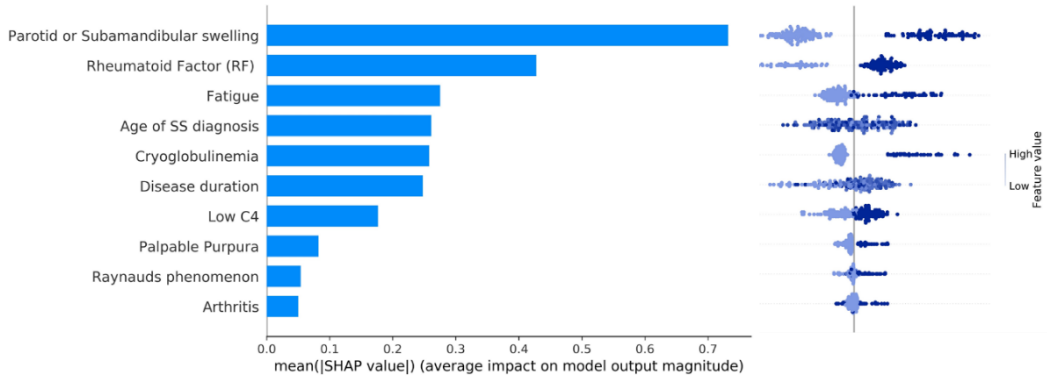


Figure 51. An illustration of the SHAP plot in federated scenario 1 for the FGBT.

The platform which was used for the purposes of the study was developed under the HarmonicSS EU funded project (HARMONIzation and integrative analysis of regional, national and international Cohorts on primary Sjögren's Syndrome (pSS) towards improved stratification, treatment and health policy making) [343] and removes the need for the installation of local servers or any type of software in each site through the adoption of a federated data management platform which supports a large family of federated AI algorithms yielding interpretable and explainable AI models. The biomarkers for lymphoma development include parotid or submandibular swelling, cryoglobulinemia, rheumatoid factor, and low C4 levels, among others, which have been validated in previous studies [228], [240] highlighting the significance of parotid or submandibular gland swelling, low C4, rheumatoid factor and cryoglobulinemia for lymphoma development. The FDART outperformed the rest of the algorithms yielding lymphoma classification models with average AUC 0.87 across the scenarios. The dropout rates introduced by the FDART yielded slightly better performance than the FGBT which confirms that the dropout elimination can enhance the decision-making process. The execution time of the federated AI workflows was 30 seconds (in average) per database which confirms the small execution time complexity.

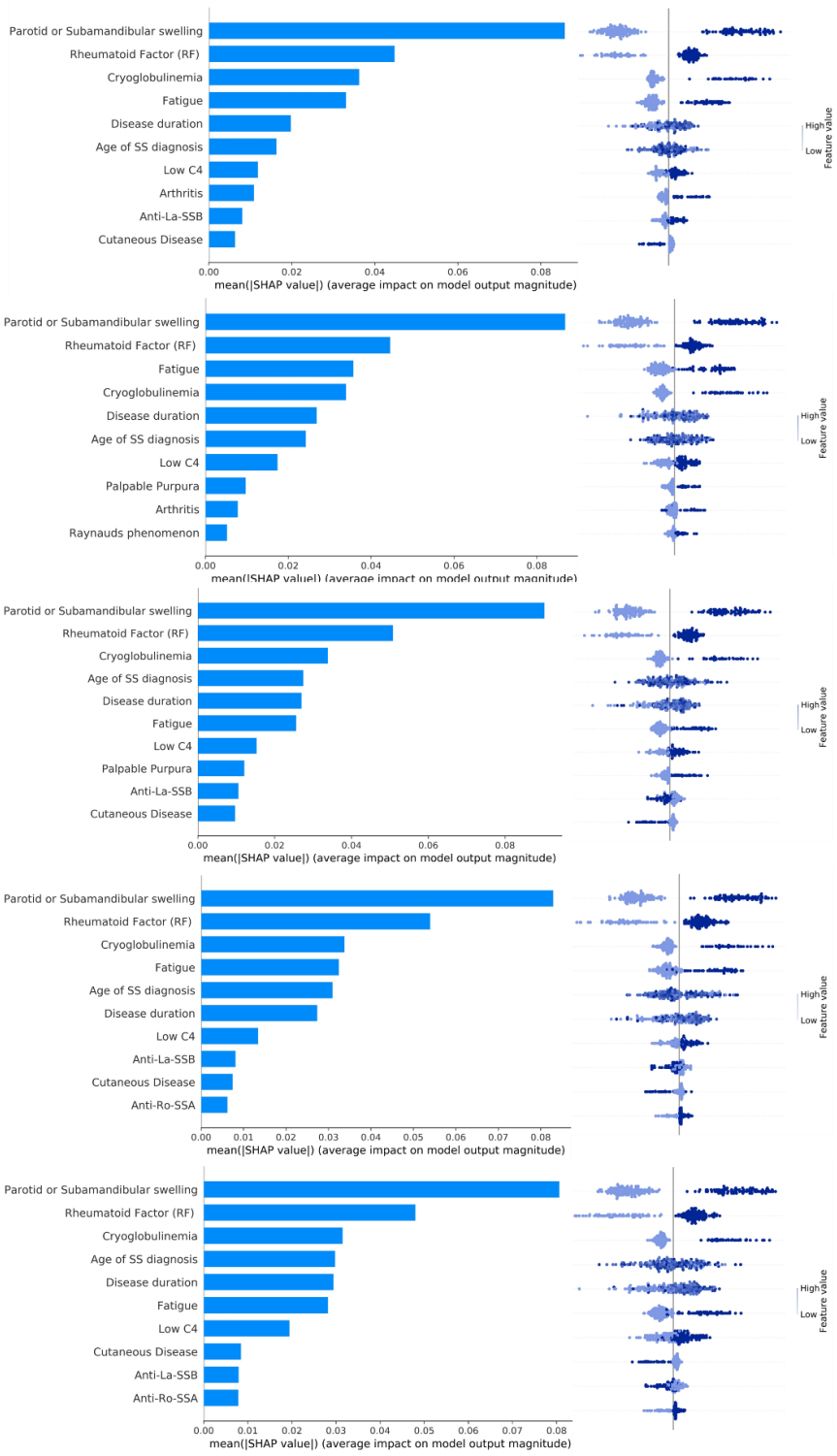


Figure 52. An illustration of the SHAP plot in federated scenario 1 for the FDART schemas.

7.1.4.4. Case Study 4 - Evaluation of the proposed FHBF algorithm

The scope of this study is to validate the classification performance and resilience against overfitting effects of a proposed federated learning algorithm across 21 European cohorts in pSS. To evaluate the performance of the FHBF against the existing

high-performance federated learning schemas (i.e., the FGBT and FDART as already reported in [65], [389]), we gained access to a Pan-European data hub on rare autoimmune diseases including 21 databases as part of the HarmonicSS Project [343]. The patient data were shared, curated, and harmonized as described in [65] and distributed in private spaces within a cloud environment. Socio-demographic information was present only in 6060 patient records which were used for the analysis. Different training and testing sequences were evaluated, involving either two or more cohorts for the training process and either one or more cohorts for testing. Two experimental phases were conducted to examine the behavior of the FHBF across highly imbalanced data structures.

To do so, the target feature in experimental design phase 1 was set to lymphoma which has a 5% occurrence in the overall population. To this end, the final number of eligible databases was reduced to 18 databases, since 5 databases had no reported lymphoma patients and thus were discarded from the experiment. The final number of harmonized patients was reduced to 4905 with 32 overlapping features (Table 27). In experimental design phase 2, the classification problem was more difficult with lower-class imbalance, where the target feature was set to MALT (mucosa-associated lymphoid tissue) lymphoma which is a lymphoma subtype with occurrence less than 3%. In this case, the final number of eligible databases was 17 (databases with no MALT patients were excluded from the analysis) with 4805 patients.

Eight case studies were defined in experimental phases 1 and 2 with random training order and different testing databases to extensively evaluate the classification performance and the average training loss of the FHBF compared to the FGBT and FDART implementations. In the first experimental phase: (i) case 1 involves the federated training across 18 databases and testing in a single database, (ii) case 2 involves the federated training in a different combination and testing in the same database as in case 1, (iii) case 3 involves the training across 18 databases and testing in a different database than in cases 1 and 2, and (iv) case 4 involves the training across 18 databases and testing in a different database than cases 1, 2 and 3. In the second experimental phase: (i) case 5 involves the federated training across 18 databases and testing in a single database ('AOUD'), (ii) case 6 involves the federated training in a different combination and testing in a different database, (iii) case 7 involves the

training across 18 databases and testing in a different database, and (iv) case 8 involves the training across 18 databases and testing in a different database than cases 5, 6 and 7. The consistency of each harmonized database was first evaluated prior to the application of the FHBF to avoid biases during the incremental weight update process. To this end, Principal Component Analysis (PCA) was applied on each individual harmonized database to extract the first four principal components as those that describe most of the variance in each database.

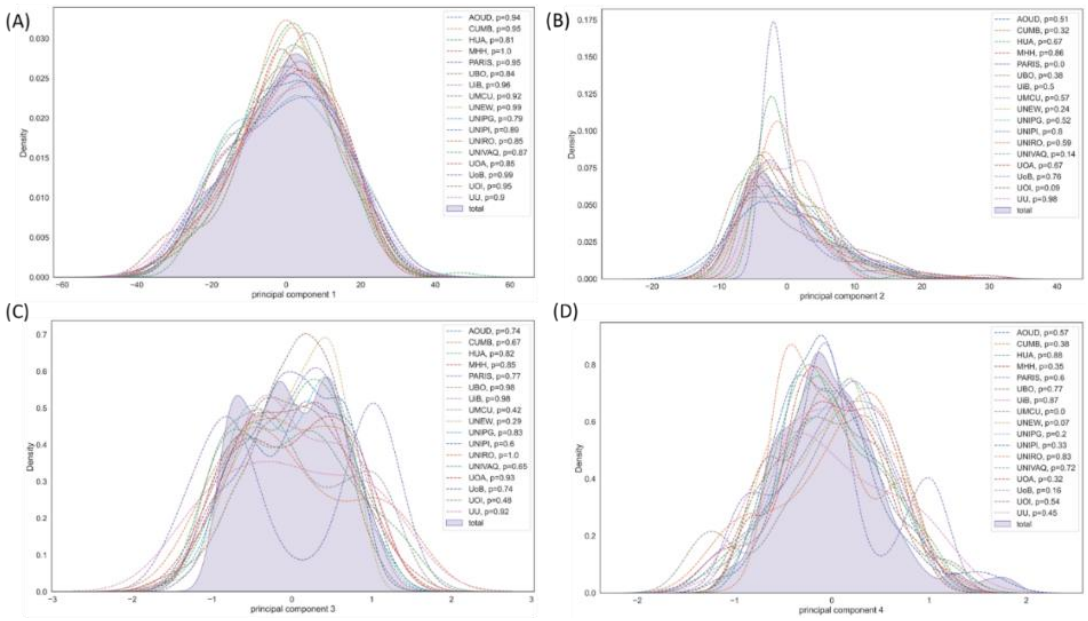


Figure 53. Distribution of the first four principal components (A-D) per database along with the first four incremental principal components (IPCs) across all databases (shaded area).

To compare the consistency of the four PCs from each individual database with the PCs across the total databases, we extended the PCA algorithm, where an empty PCA object was first fitted on a federated database A to yield a low rank approximation which was incrementally adjusted on the rest of the databases yielding the incremental PCs (IPCs) 1-4. The IPCs 1-4 were compared against PCs 1-4 from each individual database using either the Student's t-test or the Wilcoxon rank-sum test based on the normality estimations that were obtained by the Shapiro-Wilk test for normality. According to Figure 53, no statistically significant differences were observed between IPC1 and PC1 per database which confirms the consistency of the harmonized data. The same stands for IPC3 and PC3. Only one statistically significant difference was observed between PC2 and IPC2 and between IPC4 and PC4 in the PARIS ($p < 0.05$) and UMCU ($p < 0.05$) databases, respectively. The topology of the proposed hybrid loss function for

different δ values (i.e., $\delta \in [0.1, 0.3]$) is depicted in Figure 54. For comparison purposes, the topologies of the logcosh loss and the Modified Huber loss (for the same δ values) are also presented. For demonstration purposes, the horizontal axis was set to the range (-10, 10). According to Figure 54, the hybrid loss function combines the steepness of the Modified Huber loss (Figure 54 (A)) and the wideness of the logcosh loss (Figure 54 (B)) into a new loss with a smoother topology (Figure 54 (C)), where the scale of the topology is controlled by the δ value to control for overfitting effects. Since the δ value is directly linked to the dropout rate, larger dropout rates lead to a steeper loss topology thus yielding higher penalties during the weight update function.

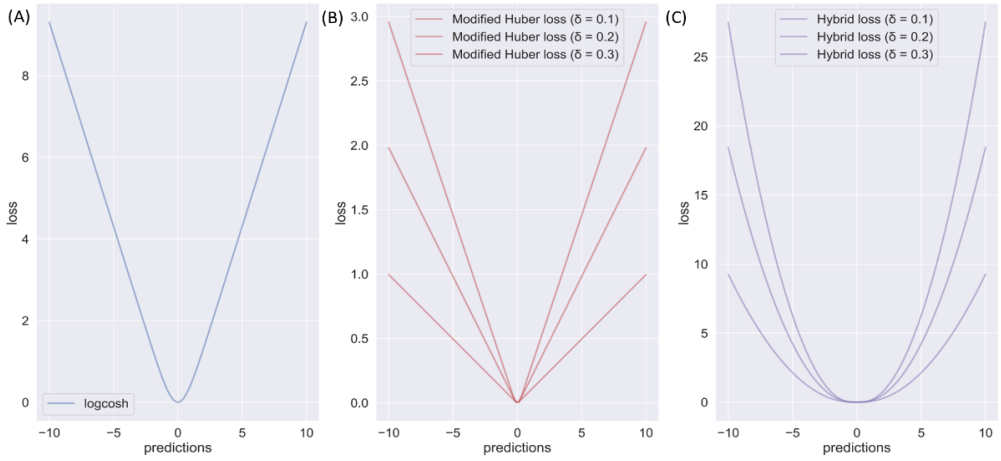


Figure 54. Topology of (A) the logcosh loss, (B) the modified Huber loss for δ values in the range 0.1 to 0.3, and (C) the proposed hybrid loss function for δ values in the same value range.

The FHBF, FGBT and FDART with dropout rates 0.1 and 0.2 were used to solve a sequence of intensive supervised learning problems across the eight cases from experimental phases 1, 2. In the FGBT, the booster was set to the 'gbtree', the objective to 'binary:logistic' and the 'eval_metric' to logloss. The parameters were updated in an incremental way, where the model that was trained in database N was updated in database $N + 1$. In the FDART, the booster was set to 'dart', the objective to 'binary:logistic', the 'eval_metric' to logloss, and the dropout rate to 0.1 and 0.2, respectively. Regarding thee FHBF, the hybrid loss scale was set to 0.1 (same as the dropout rate), the number of rounds to 20 (for evaluation purposes), and the booster to the 'HFDART' which corresponds to the FDART (with the 'dart' booster) but with the customized hybrid loss. According to Table 29, the FHBF yielded similar or better performance in both experimental phases 1 and 2 against the FGBT and the FDART.

Even in cases where the FDART with rd = 0.1 yielded poor performance (e.g., in cases 1, 5, 7), the FHBF yielded improved performance (sensitivity 0.938, specificity 0.732 in case 1; sensitivity 0.786, specificity 0.714 in case 5; sensitivity 0.833, specificity 0.659 in case 7). In other cases, the performance was similar (e.g., in cases 4, 8). It is notable that the increased performance of the FHBF is preserved in experimental phase 2 where the class imbalance ratio was even smaller, where in case 5 the FGBT weight update process is affected by the ratio yielding poor specificity. The same occurs in case 7 regarding the FDART with rd = 0.1, where the FHBF manages to prevent the weight update process from yielding zero or infinite weights.

Table 29. Performance evaluation results in experimental phases 1 and 2.

Experimental Phase	Case	Algorithm	Accuracy	Sensitivity	Specificity	AUC
Phase 1	1	FGBT	0.689	0.625	0.693	0.679
		FDART (rd = 0.1)	0.547	0.875	0.529	0.768
		FDART (rd = 0.2)	0.693	0.625	0.696	0.745
		FHBF	**0.743**	**0.938**	**0.732**	**0.871**
	2	FGBT	0.645	0.875	0.632	0.774
		FDART (rd = 0.1)	0.649	0.750	0.643	0.737
		FDART (rd = 0.2)	0.645	0.875	0.632	0.846
		FHBF	**0.743**	**0.938**	**0.732**	**0.892**
	3	FGBT	0.656	0.750	0.651	0.778
		FDART (rd = 0.1)	0.631	0.750	0.624	0.727
		FDART (rd = 0.2)	0.707	1.000	0.691	0.869
		FHBF	**0.707**	**0.875**	**0.698**	**0.885**
	4	FGBT	0.637	0.750	0.631	0.829
		FDART (rd = 0.1)	0.745	0.875	0.738	0.839
		FDART (rd = 0.2)	0.656	1.000	0.638	0.839
		FHBF	**0.758**	**0.875**	**0.752**	**0.916**
Phase 2	5	FGBT	0.497	0.857	0.479	0.693
		FDART (rd = 0.1)	0.599	0.643	0.596	0.737
		FDART (rd = 0.2)	0.667	0.786	0.661	0.796
		FHBF	**0.718**	**0.786**	**0.714**	**0.831**
	6	FGBT	0.638	0.625	0.639	0.705
		FDART (rd = 0.1)	0.733	0.750	0.731	0.746
		FDART (rd = 0.2)	0.724	0.750	0.722	0.757
		FHBF	**0.750**	**0.875**	**0.741**	**0.786**

Experimental Phase	Case	Algorithm	Accuracy	Sensitivity	Specificity	AUC
	7	FGBT	0.682	0.667	0.683	0.804
		FDART (rd = 0.1)	0.565	0.833	0.555	0.855
		FDART (rd = 0.2)	0.671	0.667	0.671	0.778
		FHBF	**0.665**	**0.833**	**0.659**	**0.812**
	8	FGBT	0.545	1.000	0.530	0.827
		FDART (rd = 0.1)	0.649	1.000	0.638	0.817
		FDART (rd = 0.2)	0.630	1.000	0.617	0.913
		FHBF	**0.675**	**1.000**	**0.664**	**0.914**
***** with **bold** color: The algorithm with the best classification performance.						

The distribution of the average log loss during the training and testing procedures across the cases 1-4 in experimental phase 1 and cases 5-8 in experimental phase 2 is depicted in Figure 55 (A) and in Figure 55 (C), respectively. In both phases, the FHBF had the lowest average training loss across cases 1-8 which highlights its resilience against overfitting. To have a concrete view of the overall loss distribution, the training and testing loss was extracted by each database and averaged across the cases in experimental phase 1 (Figure 55 (B)) and experimental phase 2 (Figure 55 (D)). Once more, the average loss of the FHBF was either lower or similar to the FGBT and FDART. The increased loss that appears in the 'UoB', 'UOI', 'MHH', and 'UBO' databases from phase 1 is leveraged by the FDART and FHBF (Figure 55 (B)). The same occurs for databases 'UNIVAQ', and 'UOI' in phase 2 (Figure 55 (D)).

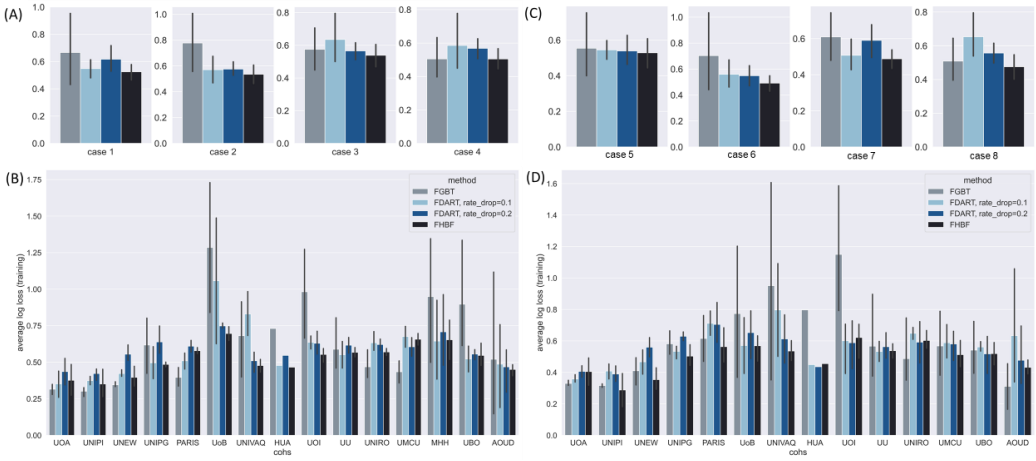


Figure 55. Average training and testing loss distribution across (A) cases 1-4, (B) cases 5-8, (C) training and testing databases involved in cases 1-4, (D) training and testing databases involved in cases 5-8.

The increased performance of the FHBF is also confirmed by the increased AUC in the ROC curves which are depicted in Figure 56.

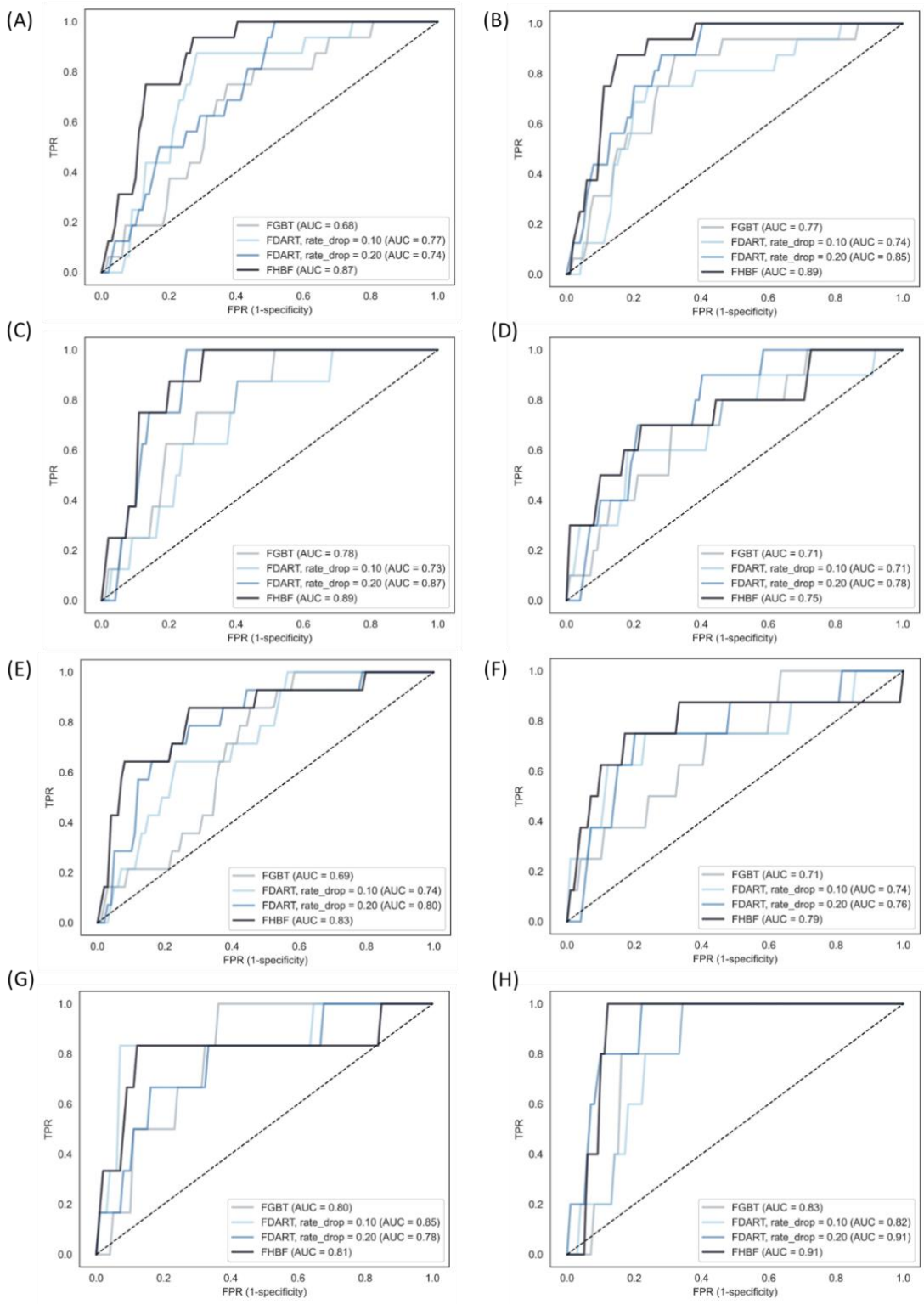


Figure 56. ROC curves for the FHBF, FGBT, and FDART (with $rd \in [0.1, 0.2]$) across cases 1-4 from experimental phase 1 (A-D) and cases 5-8 from experimental phase 2 (E-H).

Although SHAP (SHapley Additive exPlanations) analysis does not currently support the ‘dart’ booster [401], we applied the FHBF algorithm on the best case from experimental phase 2 using the HFGBT as a booster to obtain explainable outcomes and evaluate their clinical impact.

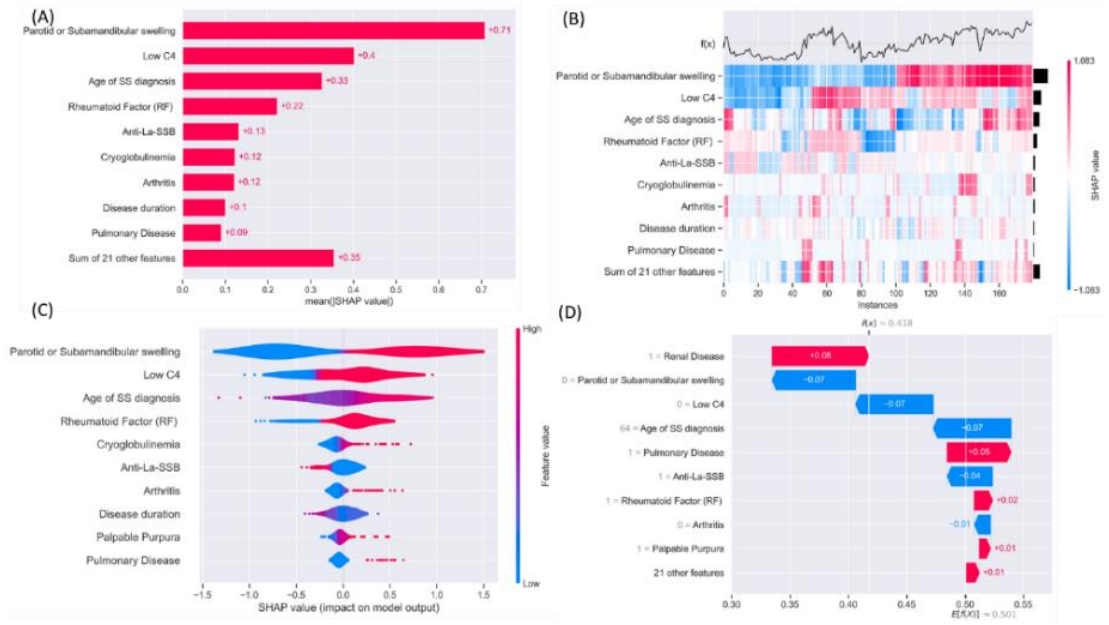


Figure 57. SHAP analysis results on a randomly selected case from experimental phase 2. (A) Global importance of each feature. (B) The population substructure is clustered by their explanations. (C) Distribution of importance for each variable. (D) Explanations for individual predictions.

More specifically, the FHBF provides: (i) global importance plots, where the global importance of each feature is expressed as the mean absolute value for that feature over all the given samples (Figure 57 (A)), (ii) heatmaps to display the population substructure of a database where data points are clustered by their explanations and not by the original feature values (Figure 57 (B)), (iii) violin plots which display the distribution of importance for each variable (Figure 57 (C)), and (iv) waterfall plots which display explanations for individual predictions (Figure 57 (D)). The bottom of a waterfall plot starts as the expected value of the model output, and then each row shows how the positive (red) or negative (blue) contribution of each feature moves the value from the expected model output over the background database to the model output for this prediction. According to Figure 57, all identified risk factors are in complete line with literature findings reported in previous study [65].

According to Figure 62(A), the average execution time of the FHBF was comparable to the FDART (and lower than the FGBT) although the total execution time (Figure 62 (B)) of the FHBF is directly affected by the number of trees in the forest. More specifically, the execution time was 59.11 sec for 20 trees, 143.09 sec for 50 trees, 315.73 sec for 100 trees, 430.16 sec for 150 trees, and 499.43 sec for 200 trees. A detailed distribution of the time execution is also presented in Figure 62(C). The runs were performed on a central node with Intel(R) Core (TM) i7-10750H CPU @ 2.60GHz.

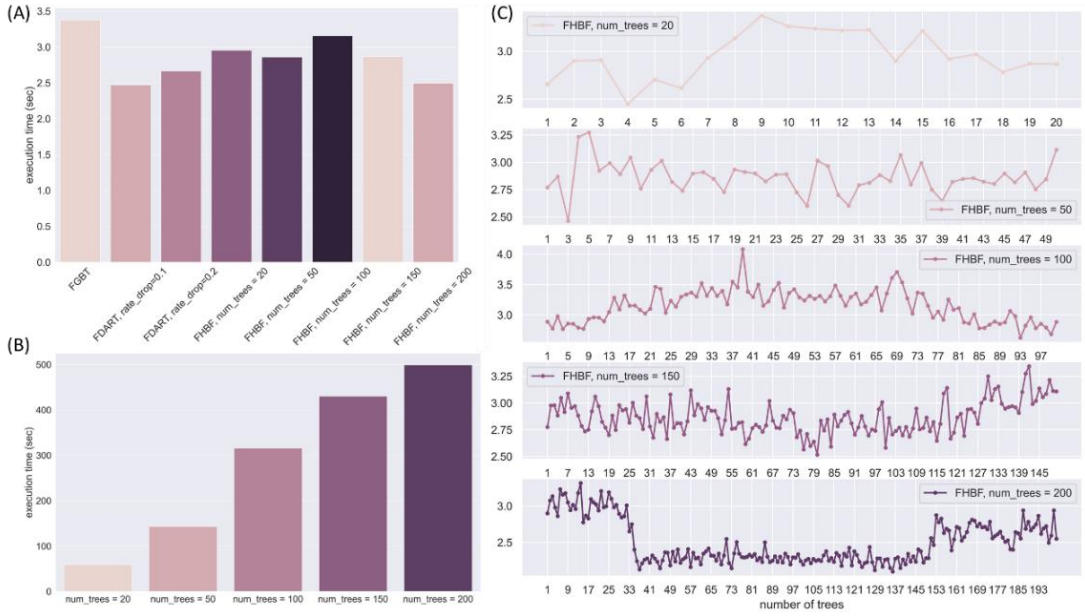


Figure 58. Computational performance in terms of execution time (sec). (A) The execution time (sec) of the FGBT and the FDART and the average execution time (sec) for the FHBF with 20, 50, 100, 150 and 200 trees. (B) The total execution time (sec) of the FHBF with 20, 50, 100, 150 and 200 trees. (C) The distribution of the individual execution times (sec) of the FHBF with 20, 50, 100, 150 and 200 trees.

The FHBF algorithm implements a hybrid weight update approach to deal with ill-posed problems that arise from overfitting effects during the training across complex and highly imbalance data in federated databases. A scale parameter is introduced to control the shape of the hybrid loss function based on the dropout rate to avoid overfitting effects. The FHBF currently supports both the hybrid FGBT (HFGBT) and the hybrid FDART (HFDART) as boosters. Class imbalance handling functionalities are incorporated to develop clusters of HFDARTs, where each cluster is formulated based on a random subset of the federated training instances.

Then, a log loss score is used to isolate the weak sets of regression trees for the classification task under investigation to further boost the classification performance of the algorithm and increase its resilience against weak decisions. SHAP analysis is applied as a final step to yield explainable outcomes. Eight case studies were conducted to demonstrate the superiority of the FHBF to solve demanding classifications tasks involving 18 federated databases against existing federated learning implementation. Our results highlight the robustness and resilience of the FHBF against overfitting effects during training and testing, yielding an average loss 0.527 across the eight cases compared to the FGBT (0.611) and the FDART (0.65 for rd 0.1, 0.582 for rd 0.2) along with increased classification performance which reached 0.938 sensitivity and 0.732 specificity supporting explainable outcomes based on the HFGBT booster.

7.2. Hypertrophic cardiomyopathy

This case study involves the application of the beyond the state-of-the-art methods that were developed for data curation (CHAPTER 3), synthetic data generation and augmentation (CHAPTER 5) to address open issues and clinical unmet needs (Section 1.4) in the domain of the hypertrophic cardiomyopathy (Section 2.3.2).

7.2.1. Data curation

7.2.1.1. Case Study 1 – Curation across two timepoints

The scope of this study is to enhance the quality of a single database in HCM across two timepoints. Anonymized clinical data were obtained from 2454 patients under the SILICOFCM project which included 69 features in total [402].

An instance of the curated clinical dataset is depicted in Figure 59. According to Figure 59, features having less than 50% missing values are depicted in blue color, features with no missing values are depicted in green color and features with more than 50% missing values are depicted in red color and are characterized as “bad” feature. The missing values are depicted in black color using the “NaN” flag and the outliers are depicted in orange. All “bad” features along with the outliers were automatically removed from further analysis and the missing values were replaced according to the mean/most frequent approach yielding the final dataset to be used for the virtual population generation.

Anonymized data were obtained from 1227 patients at two timepoints (2454 records in total) under the SILICOFCM project [402]. The dataset included 29 features (17 discrete, 12 continuous) related to demographic (e.g., age, gender), laboratory measures (e.g., diastolic pressure), and gene-related information (e.g., ACTC1, CSRP3). All the data came from the same center, the Cardiomyopathies Unit at Careggi Hospital, Florence, and were collected by a very limited number of clinicians over more than 20 years.

ST_segment_abnormali	Negative_T_wave	LA	LAVs	MvmaxPG	MVmeanPG
1	1	45	97	3,5	1,1
0	1	37	63	2	0,6
0	0	42	91	3,1	1,4
0	1	46	54	6	2,5
0	1	36	55	4,2	2,1
0	0	34	99	6,1	1,9
0	0	30	NaN	2,8	NaN
1	0	25	NaN	3,3	1,6
0	0	27	NaN	1,5	NaN
0	0	35	NaN	5,6	1,9
0	0	37	NaN	2,8	1,3
0	0	40	51	3,8	1,8
0	1	33	64	13,1	4,7
0	0	42	65	5,4	1,5
0	0	37	NaN	NaN	NaN
0	1	39	46	4,1	2,4
0	0	32	88	7,5	2,5
0	1	30	45	4,2	2,7
0	0	34	55	1,9	0,7

Figure 59. An instance of the curated dataset.

The total number of missing values within the data was 10.12% (13 features had less than 50% missing values and 16 features had no missing values at all). Neither outliers nor duplicated features or inconsistent fields were detected in the anonymized data. The data curator was able to enhance the quality of the single database in HCM.

7.2.1.2. Case Study 2 – Evaluation of the proposed “smart” data imputer

The scope of this study is to address data completeness in HCM. Anonymized clinical data were acquired from 648 patients with HCM from the Cardiomyopathies Unit at Careggi Hospital, Florence under the SILICOFCM project [402]. The total number of features was 192 (quality state: 31 “good”, 51 “fair”, 110 “bad”). Out of 192 features, 54 were discrete, 133 were continuous and 5 were unknown (i.e., mixed data types). Features with “bad” quality state (i.e., 110), missing records and anomalies were removed from the analysis yielding a final dataset with 82 features and 290 instances. Out of the remaining 82 features, 20 were selected by the clinical experts as critical for HCM development, including the “age”, “sex”, “nyhaClass” (New York Heart

Association) class, “systolicPressure”, “diastolicPressure”, “syncope”, “heart murmur”, “eflv” (ejection fraction left ventricle), “lvids” (left ventricle internal diameter end systole, “lvidd” (left ventricle internal diameter end diastole), “ivsd” (interventricular septal end diastole), “plwd” (posterior wall thickness at end diastole), “svlv” (systolic volume of left ventricle), “lvot” (left ventricular outflow tract), “lvotMaxPg” (maximal instantaneous pressure gradient), “ee” (Early diastolic mitral annular tissue velocity), “BMI” (body mass index), “la” (left atrium), “alt” (alanine aminotransferase), “ao” (aortic root diameter), and “aorticValve”.

In total 10000 virtual patient profiles were produced by each generator to enhance the number of candidate virtual profiles for the “smart” imputation process. The BGMM generator achieved the smallest average correlation difference (0.04) among the real and the virtual patient profiles with 0.02 KL divergence and VMR smaller than 1. The tree ensembles, the RBF-based ANNs and the BN come next with 0.06, 0.08, and 0.14 average correlation difference. The BN and RBF-based ANNs did not have acceptable VMR (larger than 1). The real dataset was randomly contaminated with missing values ratios, say $r \in [0.1, 0.5]$, where 0.1 and 0.5 denote 10% and 50% missing values, respectively. For each ratio, the search algorithm was applied on every set of 10000 virtual profiles from the four generators to identify the most prominent matches. The PMS was estimated for each virtual patient in each set through (3.10).

According to Table 30, the BGMM generator achieved the smallest dissimilarity between the proposed values for imputation and the original ones, in all cases, yielding average correlation difference 0.02, 0.04, 0.04, 0.05, and 0.05 for $0.1 \leq r \leq 0.5$ (the average SSAD values were 0.77, 1.7, 3.79, 5.78, 6.87). This is also confirmed in Figure 60 (C), (D) by the black patterns in the heatmaps of the average CD and SSAD values (for $r = 0.2$). According to Figure 60 (A), the kurtosis of the PMS distribution was lower for $0.1 \leq \mathrm{r} \leq 0.4$. This implies that the search algorithm was able to identify the best virtual profiles across a relatively large set of candidate virtual profiles. On the other hand, in the case r = 0.5, the PMS distribution had higher kurtosis due to the increased ratio of missing values per patient (i.e., 50%) and thus the reduced number of candidate virtual profiles. This is also confirmed by the 290x10000 heatmap of the PMS values for $\mathrm{r} = 0.2$ (Figure 60 (B)) which provide explainable scores (the best matching profiles are depicted with intense colors).

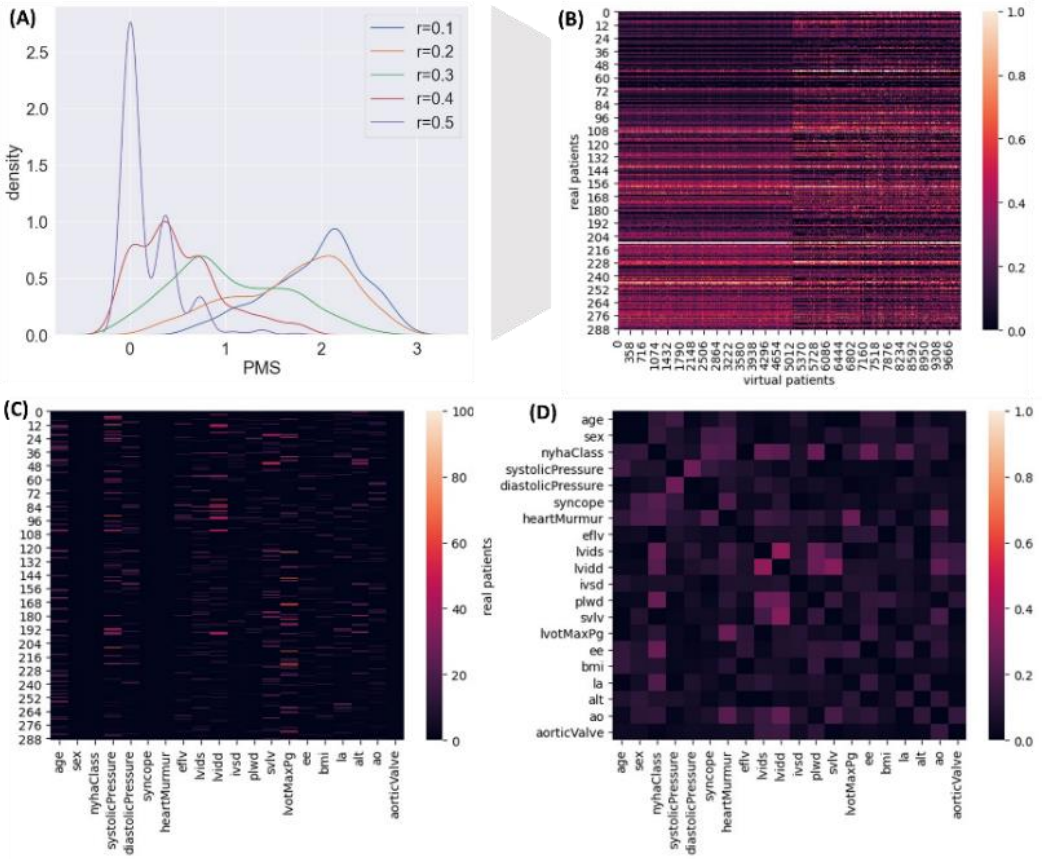


Figure 60. PMS distributions and heatmaps from the BGMM generator. (A) The PMS distribution for $0.1 \leq r \leq 0.5$. (B) The 290 (real patients) x 10000 (virtual patients) heatmap of the PMS values for r=0.2. (C) The heatmap with the average SSAD between the real and the "smart" imputed patients for r=0.2. (D) The corresponding heatmap with average CD.

The remaining generators achieved favorable performance, where the tree ensembles achieved similar results to the BGMM but at a lesser extent (average CD: 0.03, 0.04, 0.05, 0.05, and 0.05 for ratios 0.1, 0.2, 0.3, 0.4, 0.5, respectively; average SSAD: 0.84, 1.84, 3.85, 5.75, and 6.86) (Table 30).

Table 30. Average correlation difference (CD) and average scaled squared absolute differences (SSAD) between the real and the randomly imputed values per virtual data generator across different ratios of missing values.

Metric	**Virtual profile generation schemas**			
	BGMM	**Bayesian Network**	**Tree ensembles**	**RBF-based ANN**
Contamination ratio: 10% missing values ($r = 0.1$) *				
Avg. CD	0.02	0.03	0.03	0.02
Avg. SSAD	0.77	1.10	0.84	0.98

Contamination ratio: 20% missing values ($r = 0.2$) *				
Avg. CD	0.04	0.05	0.04	0.04
Avg. SSAD	1.70	2.43	1.84	2.09
Contamination ratio: 30% missing values ($r = 0.3$) *				
Avg. CD	0.04	0.06	0.05	0.05
Avg. SSAD	3.79	4.83	3.85	4.54
Contamination ratio: 40% missing values ($r = 0.4$) *				
Avg. CD	0.05	0.07	0.05	0.06
Avg. SSAD	5.78	6.84	5.75	6.84
Contamination ratio: 50% missing values ($r = 0.5$) *				
Avg. CD	0.05	0.09	0.05	0.05
Avg. SSAD	6.87	8.39	6.86	8.51
* Results were averaged across ten iterations.				

To address this demanding challenge, we propose a "smart" imputation workflow that effectively deals with missing data across complex clinical data structures. The workflow utilizes virtual population generators to produce high-quality virtual patient profiles. A profile matching score (PMS) was used to find the optimal virtual patient profiles for imputing the real ones through a search algorithm.

A case study was conducted to evaluate the performance of the proposed method towards data imputation for *in silico* clinical trials in the domain of hypertrophic cardiomyopathy (HCM). To this end, the real dataset was randomly contaminated with missing values for multiple ratios (e.g., from 10% to 50%).

Four state-of-the art data generators (i.e., Bayesian networks, tree ensembles, Gaussian Mixture Models with variational inference and artificial neural networks) were used to produce 10000 virtual patient profiles with 0.02 Kullback-Leibler divergence. The PMS distribution was utilized in the search algorithm to extract the optimal virtual patient profiles with explainable heatmaps. The BGMM generator provided imputed values with the lowest average squared absolute difference (0.4) and average correlation difference (0.02) with the real dataset. To our knowledge, this is the first "smart" approach that provides explainable virtual patients profiles for real data imputation.

In this work, we developed a "smart" imputation workflow to address missing data across complex clinical data structures. The proposed workflow utilizes virtual population generators to produce high-quality virtual patient profiles.

The profile matching score (PMS) is then estimated to quantify the similarity of the virtual patient profiles with the real ones. A search algorithm is finally applied to seek for optimal virtual patient profiles that match with the real patient profiles and thus to provide imputed values with reduced correlation difference and squared absolute difference.

A case study was conducted in the context of *in silico* clinical trials for the HCM domain, where the real patient dataset was randomly contaminated with missing values for multiple ratios varying from 10% to 50%. The BGMM generator yielded 10000 virtual patient profiles compared against the rest of the generators with less than 0.02 KL divergence and average correlation difference. The PMS was calculated for each virtual patient profile and the best matching profiles were extracted as those with the smallest dissimilarity between the proposed values for imputation and the original ones. The BGMM generator provided imputed values with the lowest average SSAD (0.4) and average CD (0.02) with the real dataset which can be confirmed by the extracted heatmaps since the optimal profiles are separated by intense color coding. To our knowledge, this is the first "smart" method that offers explainable heatmaps compared against the existing frameworks [67], [68], [70], [136], [139], [140] which provide either manual or semi-automated workflows based on pre-defined semantic data models to address data completeness.

7.2.2. Synthetic data generation

7.2.2.1. Case Study 1 – Statistically optimized synthetic data generation

The scope of this study is to generate synthetic data for *in silico* clinical trials in HCM using a statistically optimized synthetic data generator. Anonymized clinical data were obtained from 2454 patients under the SILICOFCM project [402] which included 69 features in total. The clinical experts examined the resulting curated dataset and selected a subset of 10 features for generating 300 virtual patients.

The subset of features includes the "BMI" (Body Mass Index), "age", "sex", "syncope", "NYHA class", "systolic", "diastolic", "heart murmur", "LVIDs (Left Ventricle Internal Diameter in systole phase)", and "LVIDd (Left Ventricle Internal Diameter in dystole phase)". The mean vector and the covariance matrix of the original population were then estimated and provided as input into the MVND formula.

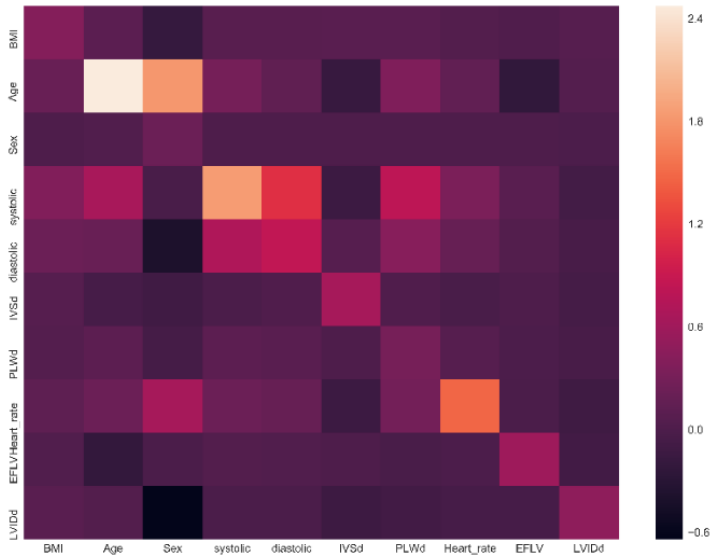
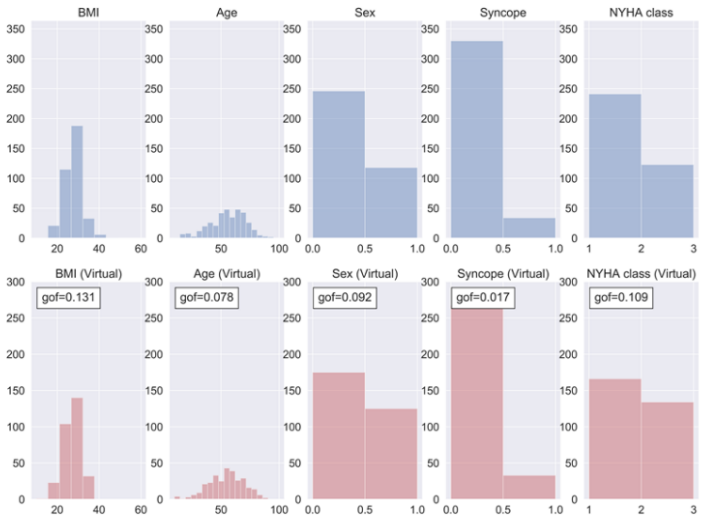


Figure 61. The covariance matrix of the real population.

The level of agreement between the multi-dimensional real distribution and the virtual one was controlled by the gof value. The 10x10 covariance matrix of the real dataset is depicted in Figure 61, where the non-diagonal cell (i, j) corresponds to the covariance between features i and j and the diagonal elements correspond to the variance of each feature. The pairs of features with high covariance are depicted in orange color whereas the pairs of features with dark color correspond to features which are independent. The results of the virtual population generation are depicted in Figure 62. In all cases, the gof values were less than (or equal to) 0.2 yielding virtual distributions similar to the real ones. The number of executions needed for the virtual population generation was approximately 5000 requiring a short amount of time (~5 sec) considering the number of virtual patients.



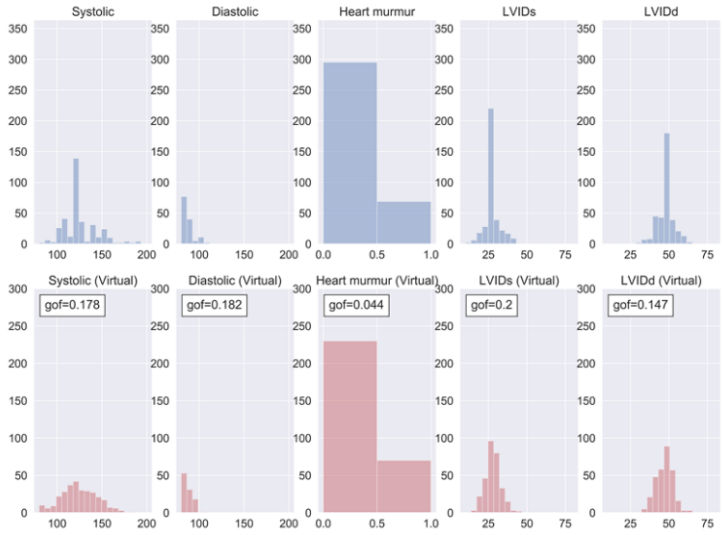


Figure 62. The histogram distribution between the features of the real (blue) and the virtual (red) populations.

The proposed method is based on a recursive execution of the MVND formula using the gof measure to control for the randomness of the virtual distributions, where the number of iterations required to control it in the previous case was equal to 5000 runs. To demonstrate the value of the proposed approach towards the generation of high-quality virtual data, the proposed method was evaluated against 10 random executions of the MVND without the gof factor as a criteria. According to Table 31, the average gof values across 10 random runs are larger than 0.2 for the systolic, diastolic, and LVIDs (Left Ventricle Internal Diameter in systole phase) features and increased for the BMI and LVIDd (Left Ventricle Internal Diameter in dystole phase). Furthermore, in these cases, the deviation of the mean values is larger than in the original population.

Table 31. Comparison results between the proposed method and 10 random executions of the MVND.

Features	**Mean/median**			**Goodness-of-fit**	
	Real	**MVND with gof opt.**	**random runs**	**MVND with gof opt.**	**random runs**
BMI	27.35	27.22	27.36	0.131	0.16
Age	55.68	54.51	56.43	0.078	0.082
Sex	0	0	0	0.092	0.078
Syncope	0	0	0	0.017	0.017
NYHA class	1	1	1	0.109	0.119
Systolic blood pressure*	126.24	124.36	125.56	0.178	0.255

Features	Mean/median			Goodness-of-fit	
	Real	MVND with gof opt.	random runs	MVND with gof opt.	random runs
Diastolic blood pressure*	76.97	75.63	76.22	0.182	0.2
Heart murmur	0	0	0	0.044	0.06
LVIDs*	28.26	28.11	28.93	0.2	0.271
LVIDd	47.39	47.19	47.43	0.147	0.193
* features with significant differences between the two cases.					

All in all, we presented a virtual population generator from a real clinical dataset based on the parametric MVND methodology optimized by an iterative process through the Kolmogorov-Smirnov goodness-of-fit test. The developed VP generator is integrated into the multi-repository virtual population model of SILICOFCM [402] which is an advanced cloud based *in silico* platform offering simulation capabilities and advanced tools for testing and development of drugs targeting the familial cardiomyopathies (FCM).

7.2.2.2. Case Study 2 - ML-based synthetic data generation

The scope of this case study is to generate high-quality synthetic data for *in silico* clinical trials in HCM using machine learning based synthetic data generators. Anonymized data were obtained from 1227 patients at two timepoints (2454 records in total) under the SILICOFCM project [402]. The dataset included 29 features (17 discrete, 12 continuous) related to demographic (e.g., age, gender), laboratory measures (e.g., diastolic pressure), and gene-related information (e.g., ACTC1, CSRP3). All the data came from the same centre, the Cardiomyopathies Unit at Careggi Hospital, Florence, and were collected by a very limited number of clinicans over more than 20 years. The clinical experts examined the resulting curated dataset and selected a subset of 10 features for generating 1000 virtual patients. The subset of features consists of the following features: “LVOTO_Rest” (Left ventricular outflow tract obstruction during resting state) (mean = 19.79, std = 21.79), “Evel” (E wave velocity) (mean = 74.8, std = 22.46), “lat_Eprime” (lateral e’ wave) (mean = 9.8, std = 3.02), “sep_Eprime” (septal e’ wave) (mean = 6.86, std = 2.14), “LA” (Left Atrium) (mean = 44.23, std = 7.39), “LVEF” (Left Ventricular Ejection Fraction) (mean = 64.25, std = 8.6), “Max_LVT” (maximum Left Ventricular Thickness) (mean = 19.24, std = 5.88),

"PW" (mean = 10.68, std = 2.08), "NYHA" (New York Heart Association class) (mean = 1.65, std = 0.71) and "Age" (mean = 51.33, std = 18.19). The performance evaluation results for each type of virtual population generation method are presented in Table 32, where the gof values where the smallest for "Max_LVT" and "Age" in the log-MVND, for the features "sep_Eprime", "LA", "PW", and "Age" in the supervised tree ensembles and for the remaining features in the unsupervised tree ensembles.

Table 32. Performance evaluation results for each type of virtual population generation method.

Feature	**Performance evaluation measures**		
	mean	**std**	**gof***
Multivariate log-normal distribution			
LVOTO_Rest	22.93	16.76	0.28
Evel	75.10	22.36	0.18
lat_Eprime	9.79	3.04	0.25
sep_Eprime	6.89	2.09	0.24
LA	43.89	7.30	0.10
LVEF	64.08	8.69	0.18
Max_LVT	19.08	6.00	**0.10**
PW	10.51	1.99	0.19
NYHA	1.63	0.76	0.27
Age	51.49	18.36	**0.04**
Supervised tree ensembles			
LVOTO_Rest	18.67	24.50	0.20
Evel	78.54	34.06	0.16
lat_Eprime	9.38	3.56	0.23
sep_Eprime	6.64	2.59	**0.18**
LA	43.60	7.26	**0.09**
LVEF	65.25	6.09	0.14
Max_LVT	17.60	5.00	0.14
PW	10.48	2.25	**0.16**
NYHA	1.53	0.68	0.09
Age	52.30	17.55	**0.04**
Unsupervised tree ensembles			
LVOTO_Rest	19.85	26.52	**0.19**
Evel	78.25	31.71	**0.15**
lat_Eprime	9.63	3.64	**0.19**
sep_Eprime	6.56	2.74	0.24
LA	42.97	7.26	0.14

LVEF	64.28	7.54	**0.07**
Max_LVT	17.70	4.73	0.13
PW	10.39	2.03	0.17
NYHA	1.51	0.62	**0.07**
Age	52.03	17.59	0.06
*The smallest gof values on each method are filled with gray color.			

The distribution of the gof values for each type of virtual population generation method is depicted in Figure 63 for the set of features which is presented in Table 32.

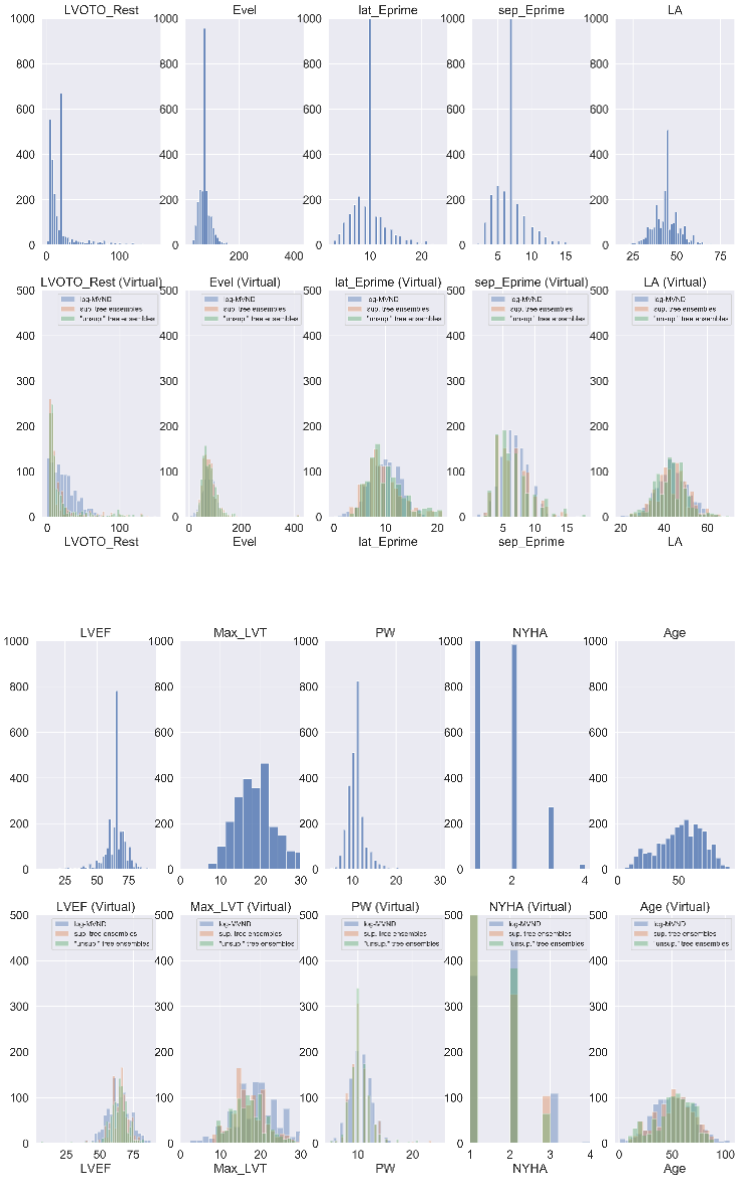


Figure 63. Distribution plots for the real (upper panel; blue color) and the virtual data (lower panel; light blue: log-MVND, green: supervised tree ensembles, orange: unsupervised tree ensembles).

The correlation matrix of the real data is depicted in Figure 64 whereas the correlation matrix for the virtual data that were derived by the unsupervised tree ensembles (as the method that achieved the highest number of "optimal" gof values) is depicted in Figure 65, showing similar association patterns within the data. High correlation values are depicted in deep blue color whereas low correlation values are depicted in yellow.

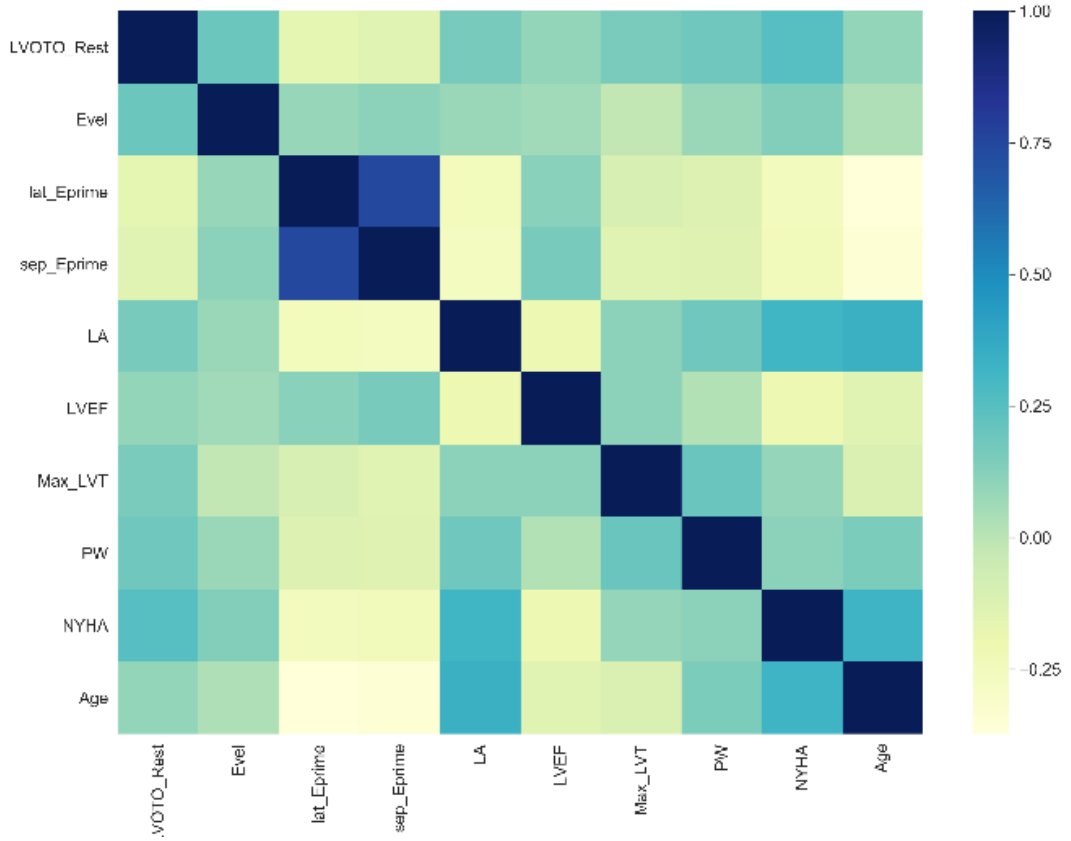


Figure 64. Correlation matrix of the real data.

Since this is a qualitative way to view the correlation between the features in the real and the virtual populations, we have also computed the absolute value of the difference between the average correlation values from the upper (or lower) triangular part of the matrices, for quantitative purposes. The difference in the average correlation values was 4.45% for the log-MVND, 8.28% for the supervised tree ensembles and 2.71% for the unsupervised tree ensembles. In this work, we deployed three computational methods to generate virtual patient data from real clinical data. We extend a previous study [132], where the MVND method was used to generate 300 virtual patients by comparing the multivariate log-normal distribution with the tree ensembles to generate 1000 virtual patients for *in-silico* clinical trials targeting the drug development for familiar cardiomyopathies (FCM).

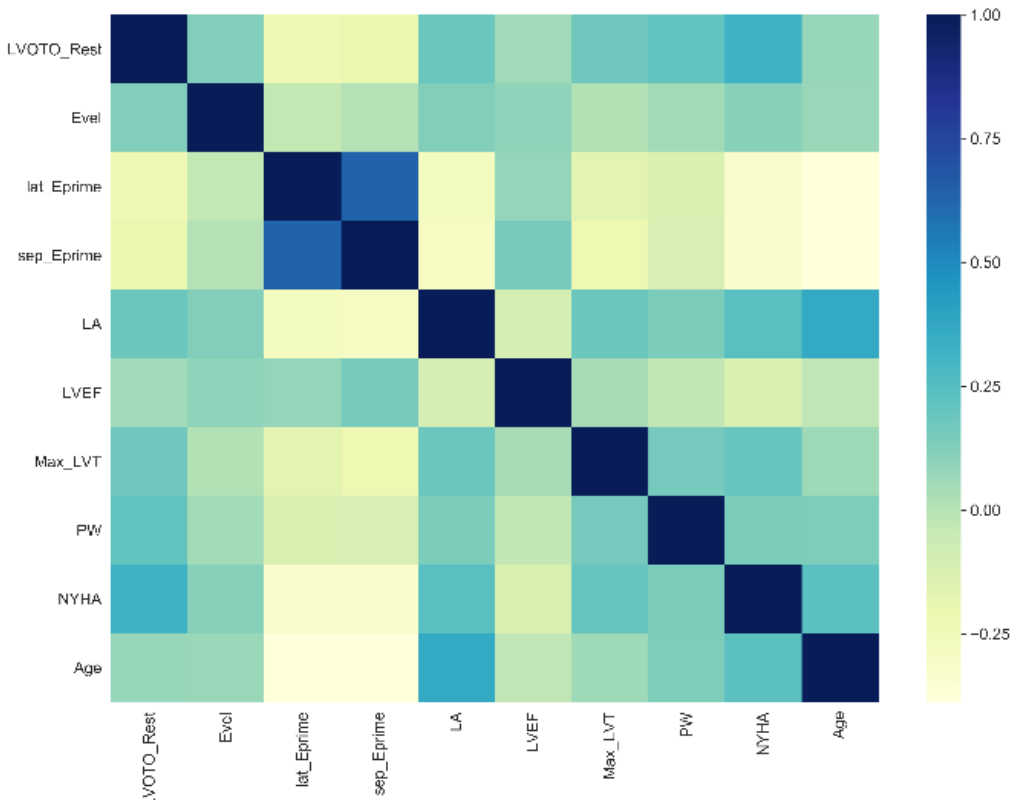


Figure 65. Correlation matrix of the virtual data from the "unsupervised" tree ensembles.

These methods have been integrated into the *in-silico* clinical trial SILICOFCM cloud-based platform [402]. The latter incorporates straightforward simulation tools for FCM drug development. Our results confirm the dominance of the tree ensembles as a prominent method for virtual population generation yielding an increased level of agreement between the real and the virtual data with average gof values less than 0.2 and correlation patterns like the real data (2.71% difference in the average correlation values). We compared three computational methods towards the generation of 1000 high-quality, virtual patient data for *in silico* clinical trials in cardiomyopathies drug development using the goodness of fit (gof) and the correlation matrix for performance evaluation purposes. Our results suggest the dominance of the tree ensembles for virtual population generation yielding virtual patient data with an increased level of agreement (distributions with less than 0.2 gof) and at the same time maintaining the correlation patterns (associations) among the features in the real clinical data.

More specifically, the "unsupervised" tree ensembles achieved the lowest goodness-of-fit values for five out of ten features according to Table 32 and Figure 63 (i.e., for the clinical features "LVOTO_Rest", "Evel", "lat_Eprime", "LVEF", and "NYHA"), the supervised tree ensembles for four out of ten features (i.e., the "sep_Eprime", "LA",

"PW", and "Age") using "NYHA" as the target feature and finally the log-MVND for only two out of ten features (i.e., the "Max_LVT", and "Age"). The correlation matrix that was generated by the "unsupervised" tree ensembles was close to the original one, a fact that enhances the level of agreement between the virtual and the real data. For example, the strong association between the lateral e` wave ("latEprime") and the septal e` wave ("sepEprime") (Figure 64) which is high (more than 75%) is clearly preserved in the virtual population (Figure 65). The proposed methods could potentially provide significant insight in the field of virtual population generation to re-adjust the perspective of Clinical Trials (CTs) in other domains. As a future work, we also plan to deploy artificial neural networks (ANNs) that make use of radial basis functions (RBFs) as activation functions towards the generation of even more robust clinical data for *in-silico* clinical trials.

7.2.2.3. Case Study 3 - Evaluation of the BGMM with robust priors

The scope of this case study is to generate high-quality synthetic data for *in silico* clinical trials in HCM using Bayesian Gaussian Mixtures with robust Dirichlet priors and weight concentration values. Anonymized data were obtained from 776 patients under the SILICOFCM project [402]. The dataset included 20 features related to demographic and echocardiographic measurements. A data curation pipeline presented in a previous study [66] was applied on the clinical data to remove outliers, duplicated fields, and inconsistent data types using both univariate and multivariate methods. All detected outliers and duplicated fields, as well as, features with high number of missing records were removed from further analysis. The final curated dataset included 11 features, namely the: (i) "Ech_Echo_LA" (Left Atrium), (ii) "Ech_Echo_LVIDs" (Left ventricular internal dimension), "ABNORMAL_HOLTER" (Abnormal Holter indicator), "Ech_Echo_Aortic_Root", "NYHA" (New York Heart Association class), "ARRHYTHMIA_NSVT" (Non sustained ventricular tachycardia), "Ech_Echo_PW" (Pulse Wave Doppler), "BMI" (Body Mass Index), "BSA" (Body Surface Area), "Height", "High_Risk". These features were used to evaluate the generators across multiple virtual patients in the range [1000, 20000] with a step 1000. The average goodness of fit and inter-correlation values are depicted in Figure 66 for components in the interval [1, 30]. For illustration purposes, the number of virtual patients has been restricted in the interval [1000, 10000].

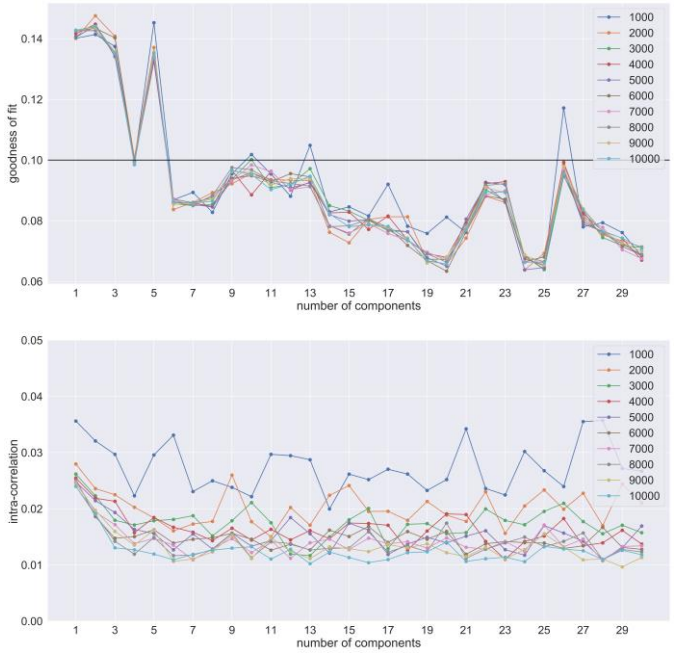


Figure 66. Performance evaluation of the proposed BGMM across multiple virtual patients in range [1000,10000].

According to Figure 66, the average gof value was less than 0.1 for more than 5 Gaussian components. The average inter-correlation difference was less than 0.04 across the multiple virtual populations' and in some executions even less than 0.03.

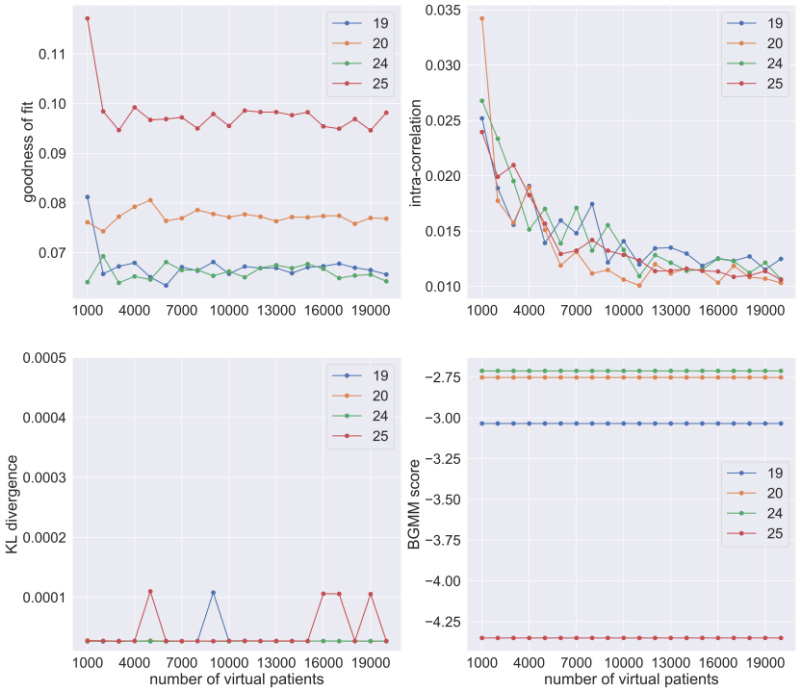


Figure 67. Performance evaluation of the proposed BGMM for the four best components across multiple virtual patients.

The average goodness of fit and correlation values from the four most prominent Gaussian components of Figure 66 (i.e., for 19, 20, 24, and 25 components) are depicted in Figure 67, along with the corresponding KL divergence and log-likelihood scores (which are referred to as BGMM scores). According to Figure 67, the number of components that yielded virtual data with the smallest goodness of fit, KL divergence scores, correlation values, and the highest BGMM scores, across all executions, was 24. This number was combined with the Dirichlet concentration (gamma) value to generate multiple virtual patients.

For comparison purposes, the number of virtual patients was set to 20000. According to Table 33, the proposed BGMM approach achieved the lowest gof (less than 0.1) along with the UTE and the STE compared to the RBF-based ANN and the Bayesian networks. In addition, the proposed BGMM method yielded the lowest inter- and intra-correlation differences between the features in the virtual data (0.0133 inter-correlation and 0.0121 intra-correlation). In all cases, the average KL divergence was less than 0.001 highlighting the increased similarity of the synthetic with the real dsitributions.

Table 33. Performance evaluation results.

Method	**Average performance evaluation measures**			
	Goodness of fit	**Inter-correlation difference**	**Intra-correlation difference**	**KL divergence**
BGMM	0.0667	0.0133	0.0121	<0.001
UTE	0.0211	0.0309	0.0281	<0.001
STE	0.0261	0.0433	0.0393	<0.001
ANN	0.1872	0.0829	0.0753	<0.001
Bayesian	0.1864	0.0824	0.0749	<0.001

According to Figure 68, the average execution time of the proposed BGMM approach was faster than the UTE and the STE methods, yielding multiple virtual populations in 4.321 sec against the UTE and the STE which required 46.537, and 34.096 sec, respectively. The gap in the proposed BGMM during the generation of 10000 patients is related to the fast convergence of the BGMM. The average execution times of the ANNs and the Bayesian methods were ignored due to their reduced performance against the previous methods. Gaussian Mixture Models (GMMs) with variational Bayesian inference (BGMM) were developed to generate large-scale virtual populations. The proposed method utilizes Dirichlet process mixtures as the BGMM's

prior structure, where the concentration of each component on the weight distribution is an exponential function of the number of components. Our approach was compared against state-of-the-art virtual data generators, including the Bayesian networks, the STE, the UTE, and the ANN for the generation of 20000 virtual patients for *in-silico* clinical trials in HCM yielding the lowest inter- and intra-correlation differences (0.013 and 0.012), in lower execution time (4.321) than the STE (46.537 sec) which had the second-best performance.

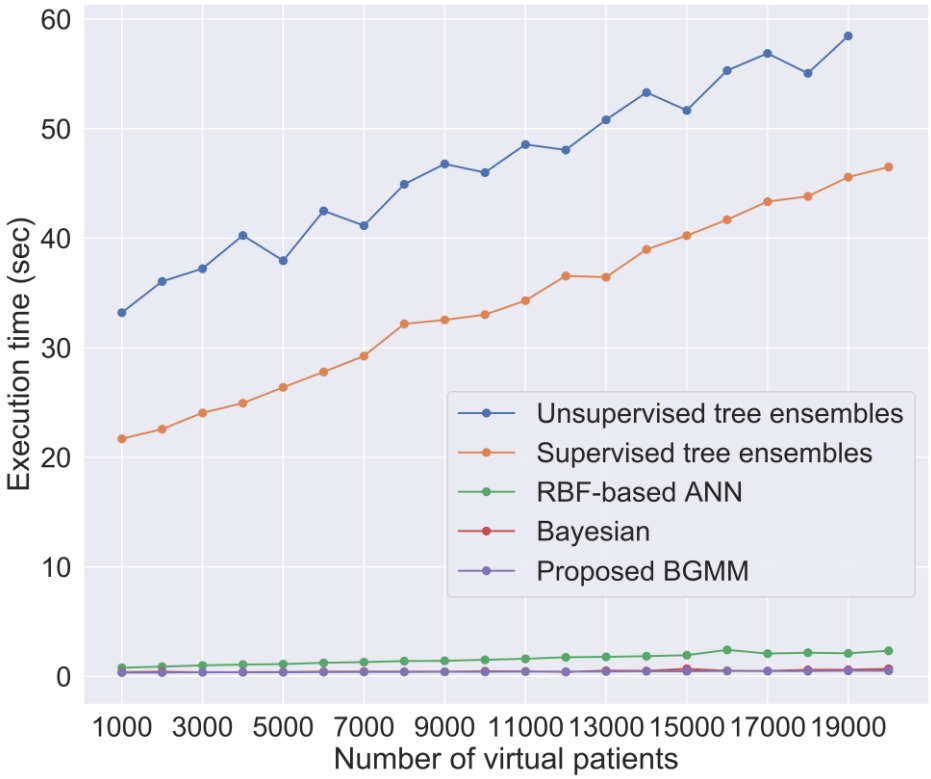


Figure 68. Execution time (sec) per virtual data generator.

We utilized probabilistic Gaussian Mixture Models with variational Bayesian inference (BGMM) for the generation of large-scale virtual populations for *in-silico* clinical trials in HCM. The proposed approach uses weight concentration values for variational inference which are based on an exponentially decaying transformation of the number of Gaussian components. The proposed approach was compared against state-of-the-art virtual data generators, including, the Bayesian networks, the supervised tree ensembles (STE), the unsupervised tree ensembles (UTE), and the ANN yielding better inter- and intra- correlation differences in less execution time than the UTE which achieved the second-best performance. The proposed method for the estimation of the

Dirichlet concentration of each component on the weight distribution yielded a stable number of components (24 components) across multiple virtual populations executions, where the prior structure of the GMM was defined according to the Dirichlet process mixture. The proposed BGMM with the optimal number of Gaussian components achieved the lowest goodness of fit values (less than 0.1) along with the UTE and the STE compared to the RBF-based ANN and the Bayesian networks (with average gof value larger than 0.15).

In addition, the proposed BGMM method yielded the lowest inter- and intra-correlation differences between the features in the virtual data (almost 0.01), in less execution time (0.4319 sec) than the STE (46.5373 sec), which had the second-best performance. This confirms the computational efficiency of the proposed BGMM approach towards the generation of large-scale virtual populations for *in-silico* clinical trials in HCM.

7.2.2.4. Case Study 4 - Evaluation of the BGMM-OCE

The scope of this case study is to generate high-quality synthetic data for *in silico* clinical trials in HCM using a computationally efficient, large scale synthetic data generator. Anonymized clinical data were obtained from 648 patients who have been diagnosed with hypertrophic cardiomyopathy under the SILICOFCM project [402]. The dataset included 188 features (71 discrete, 116 continuous, 1 unknown). Out of 188 features 85 were automatically annotated as “eligible”, whereas the remaining 103 features as “non-eligible”. Outliers were detected in 61 “eligible” features and were resolved. Imputation was applied in the “eligible” features based on the kNN approach. The non-eligible features were removed from further analysis.

The following set of 20 features was included in the analysis, upon clinical inspection as potential risk factors for HCM: age, sex, NYHA class, systolic pressure, diastolic pressure, syncope, heart murmurs, left ventricular ejection fraction (LVEF or EFLV), left ventricular internal dimension at end-diastole (LVIDd), left ventricular internal dimension at end systole (LVIDs), intraventricular septal thickness at end-diastole (IVSd), posterior wall thickness at end-diastole (PLWd), end-systolic volume of left ventricle (SVLV), left ventricular outflow tract maximum pressure gradient (maxLVOTPG), Doppler E/E’ ratio (EE), body to mass index (BMI), left atrium size (LA), Alanine aminotransferase (ALT), aorta size (AO), and aortic valve (AV).

Spectral clustering was first applied to estimate the number of clusters using the LOBPCG eigensolver with the gamma ratio set to $exp(-opt)$, where opt refers to the number of clusters and varies between 2 and 20. The DBS was computed for each cluster to quantify the consistency of the resulting clusters. According to Figure 69 (A), the number of clusters having the highest DBS was 10. The process was repeated for multiple virtual populations varying from 1000 to 30000 virtual patients with a step 1000. In each case, the BGMM-OCE algorithm was trained using 10 Gaussian components to estimate the covariance matrix and the mean vector of the Gaussian distribution. The latter were used for the training process, where the weight concentration prior was set to $\exp(-opt)$, with opt 10. According to Figure 69 (B) the distribution of the average intra-correlation differences appears to be decaying over the increasing number of virtual patients, with differences less than 0.018 for more than 14000 virtually generated patients.

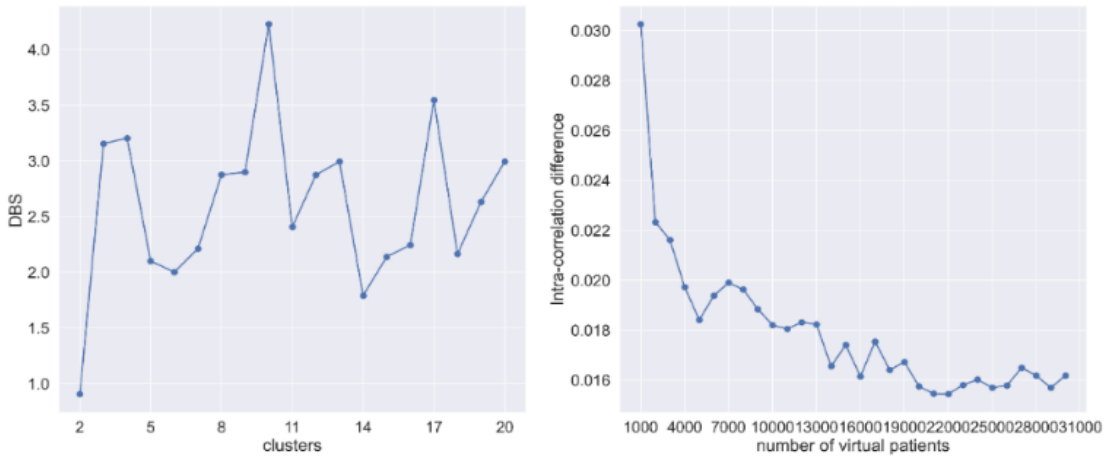


Figure 69. BGMM-OCE testing across 20 components/clusters. (A) The DBS distribution, (B) average intra-correlation difference between the real and the virtual data for multiple virtual patients.

The virtual data quality results for each data generator are depicted in Figure 70 across multiple virtual patient scenarios varying from small case population generation (i.e., 1000 virtual patients) to large-scale virtual population generation (i.e., 30000 virtual patients). According to Figure 70, the BGMM-OCE achieved the best performance yielding the lowest average intra-and inter-correlation difference, the lowest GOF and the lowest cV with non-significant variations in the average KL divergence difference (less than 0.05).

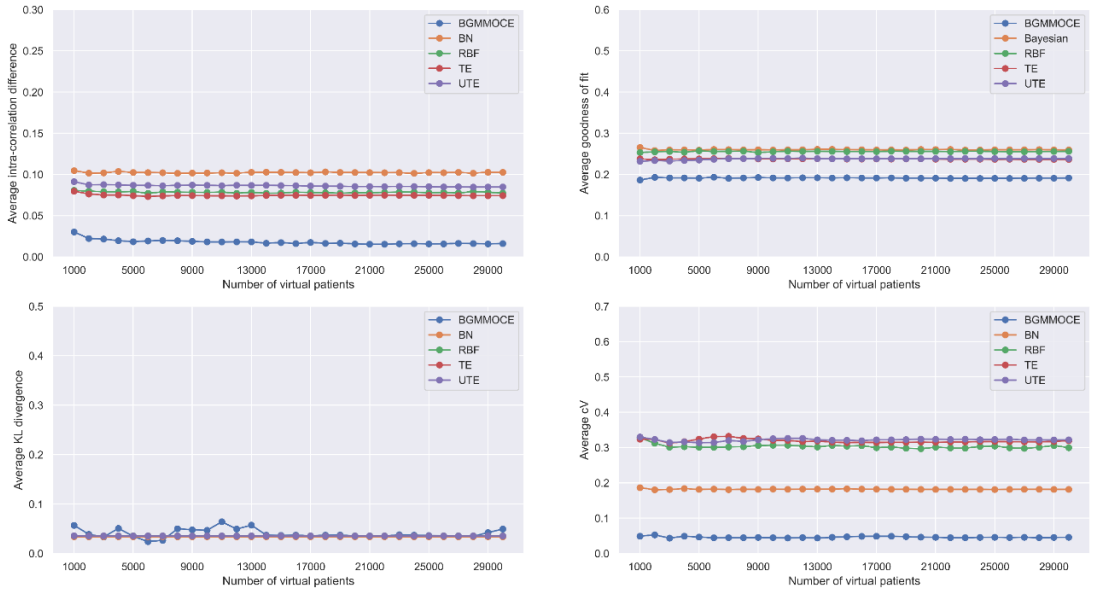


Figure 70. Virtual data quality results. (A) Average intra-correlation, (B) GOF, (C) KL-divergence, and (D) cV differences across multiple virtual patients per data generator.

Gaussian kernel density estimation was applied to estimate the density of the real and virtual data (Figure 71). According to Figure 71, the KL values were less than 0.1 in all cases except for the diastolic pressure due to its increase heterogeneity with the rest of the features.

Table 34. Data quality evaluation results for N = 1000, 10000, 20000, 30000 virtual patients.

	BGMM-OCE	**BN**	**RBF-based ANNs**	**UTE**	**STE**
N = 1000					
KL-divergence	0.057	0.034	0.036	0.036	0.036
GOF	0.186	0.265	0.253	0.232	0.238
Inter-corr. diff	0.032	0.110	0.085	0.096	0.084
Intra-corr. diff	0.030	0.105	0.081	0.091	0.080
cV diff.	0.049	0.186	0.328	0.330	0.324
N = 5000					
KL-divergence	0.035	0.034	0.036	0.035	0.035
GOF	0.186	0.259	0.257	0.235	0.238
Inter-corr. diff	0.019	0.108	0.084	0.091	0.078
Intra-corr. diff	0.018	0.102	0.079	0.091	0.074
cV diff.	0.046	0.181	0.300	0.314	0.324
N = 10000					
KL-divergence	0.047	0.034	0.036	0.035	0.035
GOF	0.192	0.259	0.255	0.239	0.238

	BGMM-OCE	BN	RBF-based ANNs	UTE	STE
Inter-corr. diff	0.019	0.107	0.082	0.092	0.078
Intra-corr. diff	0.018	0.102	0.078	0.087	0.074
cV diff.	0.045	0.182	0.306	0.325	0.320
N = 20000					
KL-divergence	0.035	0.034	0.036	0.035	0.035
GOF	0.191	0.26	0.256	0.238	0.238
Inter-corr. diff	0.017	0.108	0.082	0.090	0.078
Intra-corr. diff	0.016	0.102	0.078	0.085	0.074
cV diff.	0.046	0.181	0.297	0.324	0.315
N = 30000					
KL-divergence	0.049	0.034	0.036	0.035	0.035
GOF	0.191	0.26	0.256	0.239	0.237
Inter-corr. diff	0.017	0.108	0.081	0.089	0.078
Intra-corr. diff	0.016	0.103	0.077	0.085	0.074
cV diff.	0.046	0.181	0.300	0.322	0.319

The densities of the best data generator (i.e., the BGMM-OCE) for 1000 virtual patients are depicted in Figure 71 along with the average coefficient of variation (cV) difference between the real and the virtual distributions. In all cases, the average cV difference was less than 0.1 highlighting the reduced dispersity of the virtually generated data.

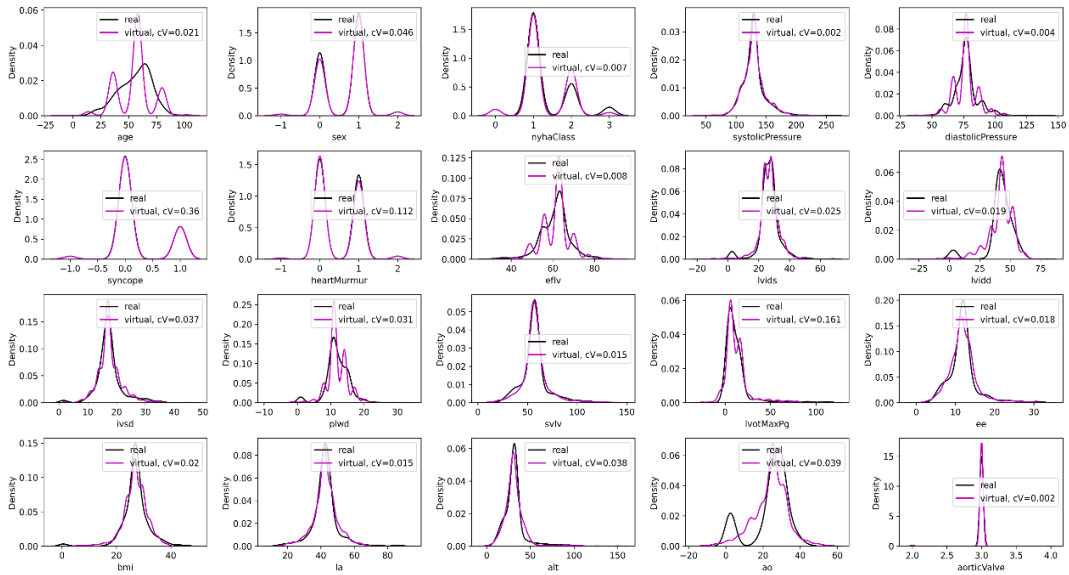


Figure 71. Real (black) and virtual (magenta) distributions for the 20 features under evaluation, where the number of virtual patients was set to 1000. The cV values refer to the absolute coefficient of variation difference between the real and the virtual ones.

According to Figure 72, the BGMM-OCE required 23 seconds on average for the optimal component initialization step, including the sequential application of spectral clustering for 2, 3, …, 20 clusters, the estimation of the DBS, the extraction of the best clustering solution and the BGMM training procedure. In the case where the sequential application of spectral clustering involved 2, 3, ..., 10 clusters, the execution time was reduced to 16 sec. However, the execution time for random sampling (upon BGMM training) across different virtual populations (1000, 2000, …, 30000 virtual patients) was less than 1 second (0.031 sec on average). On the other hand, the TE, BN, and UTE had the largest average execution time (53 sec, 63 sec, and 75 sec, respectively). It is interesting to note that the RBF-based ANNs achieved the lowest average execution time (approximately 16 sec), but its increased computational tendency for virtual populations beyond 23000 (or 17000 in the cases where the BGMM-OCE is applied across 2, 3,…, 10 clusters under evaluation) indicates a higher computational complexity than BGMM-OCE.

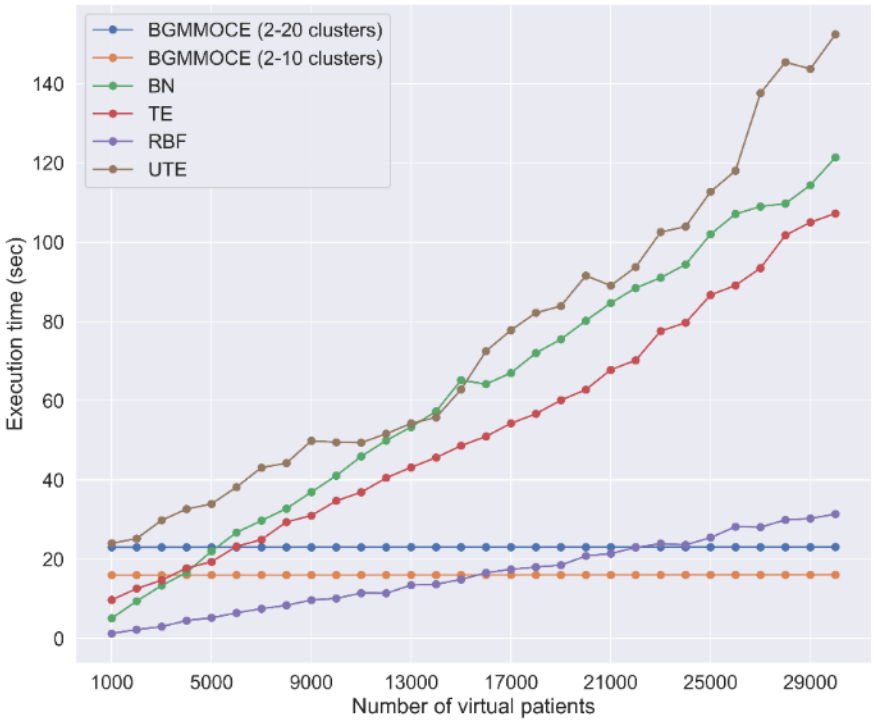


Figure 72. Execution time comparison results.

We focused on the optimal estimation of the number of Gaussian components in the conventional BGMM algorithm to yield concrete (non-arbitrary) estimations of the VI and at the same time reduced computational complexity towards large-scale virtual

population generation (we refer to this new approach as Bayesian Gaussian Mixture Models with optimal component estimation - BGMM-OCE). To do so, we first apply spectral clustering on curated and transformed input data based on the LOBPCG method to identify the best clustering solution as the one with the highest Davies Bouldin (DB) score at small computational complexity. Then, we utilize the optimal number of clusters as the number of Gaussian components in the BGMM. Since low gamma values make the model put most of the weight on few components and set the remaining weights close to zero, we define an exponential relationship between the optimal number of components to yield gamma values with smooth decay over increasing number of components. A case study was conducted to demonstrate the robustness of the proposed approach, where the BGMM-OCE was compared against state-of-the-art virtual data generators (BN, UTE, STE, ANNs) for the generation of diverse virtual populations (from 1000 to even 30000 virtual patients) in the context of in *in silico* clinical trials for hypertrophic cardiomyopathy (HCM). According to our results, the BGMM-OCE was able to generate 30000 virtual patients with the lowest coefficient of variation (0.046), goodness of fit (0.191), KL divergence (0.049), and inter- and intra- correlation differences (0.017, 0.016) at stable execution time (16.12 sec) for increased virtual populations.

7.2.3. Data augmentation

7.2.3.1. Case Study 1 – Augmentation in a single database

The scope of this case study is to enhance the performance of the existing HCM risk stratification models through data augmentation. To this end, we required an anonymized dataset which includes 2,454 records of patients who have been diagnosed with hypertrophic cardiomyopathy (HCM), at two timepoints, from the Cardiomyopathies Unit at Careggi Hospital, Florence (UNIFI cohort) [402]. The number of high-risk patients was 300 with an average age 50.13 ($\pm$17.67) and the number of low-risk patients was 476 with an average age 43.95 ($\pm$18.42). There were 123 features, including demographics, laboratory measures (e.g., Left ventricular internal diameter end systole), and physical measures (e.g., systolic pressure), among others. All clinical data were shared according to the EU GDPR requirements. The performance evaluation outcomes of the five virtual population generators in the HCM dataset are presented in Table 35 for the unsupervised and supervised tree ensembles,

and for the supervised RBF-based ANNs, while the Bayesian networks and the Log-MVND are shown in Table 36. According to Table 37, the average GOF was 0.029 for the unsupervised tree ensembles, 0.031 for the supervised tree ensembles, 0.23 for the RBF-based ANNs, 0.32 for the Bayesian networks and 0.198 for the Log-MVND. The average KL-divergence was 0.027 for the unsupervised tree ensembles, 0.031 for the supervised tree ensembles, 0.02 for the RBF-based ANNs, 0.00047 for the BN and 0.121 for the Log-MVND. The unsupervised tree ensembles generated virtual distributions with the highest similarity and reduced divergence with the real data.

Table 35. Performance evaluation of the virtually generated data for the unsupervised tree ensembles, the supervised tree ensembles, and the supervised RBF-based neural networks.

Feature	Real	Unsupervised tree ensembles			Supervised tree ensembles			Supervised RBF		
	Mean	Mean	gof	KL	Mean	gof	KL	Mean	gof	KL
Age	46.32	45.16	0.05	5.03 E-06	46.47	0.03	3.98E -06	47.69	0.09	1.77E -05
Ech_Echo_LA	44.03	43.95	0.03	5.60 E-05	43.75	0.03	5.19E -06	43.68	0.25	3.47E -03
Ech_Echo_IVS	19.13	19.38	0.03	7.04 E-07	19.21	0.02	6.29E -07	20.09	0.21	4.64E -06
Ech_Echo_Max_LVT	20.85	21.15	0.03	9.44 E-06	21.21	0.04	7.44E -06	21.33	0.18	1.96E -05
Ech_Echo_LVIDs	28.68	28.75	0.03	1.39 E-03	28.54	0.03	1.40E -03	29.36	0.31	3.85E -05
Past_Abnormal_Holter	0.07	0.07	0.01	1.68 E-09	0.06	0.02	6.02E -09	0.11	0.03	1.45E -08
Ech_ Aortic_Root	30.91	31.11	0.02	1.39 E-04	31	0.03	7.19E -05	31.27	0.27	2.34E -04
NYHA	1.72	1.77	0.06	1.12 E-08	1.76	0.03	7.47E -09	1.95	0.11	3.70E -07
Past_Arrthythmia_NSVT	0.08	0.06	0.01	4.55 E-09	0.08	0.00	3.82E -10	0.11	0.03	1.56E -08
Ech_Echo_LVIDd	46.53	46.62	0.03	1.49 E-03	46.06	0.05	7.61E -05	45.55	0.24	2.12E -03
Ech_Echo_PW	10.57	10.71	0.03	1.21 E-06	10.58	0.04	5.28E -07	11.72	0.38	1.72E -03
Ech_Echo_LA_Vol	88.37	87.84	0.02	1.01 E-01	87.51	0.04	2.17E -01	92.56	0.35	6.90E -02
Ech_Echo_LVEF	65.01	65.14	0.02	3.24 E-04	64.88	0.02	9.18E -03	62.71	0.31	5.41E -03
Mitral_Valve_E_DT	202.23	198.32	0.04	3.32 E-01	198.76	0.05	2.76E -01	215.40	0.36	1.89E -01
BP_Systolic	123.38	123.47	0.04	1.75 E-02	123.06	0.04	3.61E -02	128.99	0.32	5.44E -02
LVOTO_Rest	14.33	13.86	0.04	8.27 E-02	13.30	0.06	8.29E -02	33.61	0.58	4.57E -02
BMI	25.75	25.92	0.05	2.80 E-04	25.75	0.02	3.50E -04	25.51	0.20	2.18E -04
BSA	1.86	1.87	0.02	7.31 E-10	1.87	0.02	1.12E -08	1.86	0.21	6.23E -08
Height	169.34	169.44	0.02	3.33 E-06	169.17	0.03	3.02E -06	168.57	0.19	1.31E -02
LABEL__High_Risk	0.39	0.38	0.01	4.33 E-11	0.36	0.03	1.09E -09	0.39	0.00	2.22E -13

Table 36. Performance evaluation of the virtual data from the Bayesian networks and the Log-MVND.

Feature	Real	Bayesian Networks			Log MVND		
	Mean	Mean	gof	KL	Mean	gof	KL
Age	46.32	45.75	0.03	3.16E-06	45.66	0.07	0.07
Ech_Echo_LA	44.03	42.70	0.25	3.30E-05	43.86	0.22	0.01
Ech_Echo_IVS	19.13	23.25	0.32	1.47E-05	19.09	0.20	0.00
Ech_Echo_Max_LVT	20.85	23.41	0.26	1.25E-05	20.82	0.15	0.00
Ech_Echo_LVIDs	28.68	31.79	0.42	2.74E-05	28.58	0.24	0.01
Past_Abnormal_Holter	0.07	0.49	0.42	5.02E-07	0.07	0.01	0.00
Ech_ Aortic_Root	30.91	33.15	0.36	1.48E-05	30.94	0.25	0.00
NYHA	1.72	2.47	0.37	2.63E-06	1.74	0.03	0.00
Past_Arrthythmia_NSVT	0.08	0.50	0.43	5.05E-07	0.08	0.01	0.00
Ech_Echo_LVIDd	46.53	42.02	0.28	3.93E-05	46.69	0.20	0.01
Ech_Echo_PW	10.57	15.53	0.46	1.58E-05	10.65	0.28	0.00
Ech_Echo_LA_Vol	88.37	96.44	0.37	5.67E-03	86.55	0.35	1.41
Ech_Echo_LVEF	65.01	64.29	0.27	8.10E-05	65.09	0.26	0.05
Mitral_Valve_E_DT	202.23	216.84	0.39	2.12E-03	202.34	0.35	0.44
BP_Systolic	123.38	130.55	0.32	4.11E-04	123.26	0.27	0.08
LVOTO_Rest	14.33	50.58	0.63	9.74E-04	20.65	0.43	0.32
BMI	25.75	26.10	0.25	3.91E-06	25.80	0.23	0.00
BSA	1.86	1.87	0.22	1.50E-07	1.86	0.20	0.00
Height	169.34	168.69	0.20	2.72E-05	169.21	0.17	0.02
LABEL__High_Risk	0.39	0.50	0.11	1.47E-08	0.43	0.04	0.00

Table 37. Summary of the average performance evaluation measures for assessing the quality of the virtual data generated by each virtual population generation method for the HCM domain.

Virtual population generation method	Quality of the virtual data		
	GOF	KL-divergence	Correlation coefficient
Unsupervised tree ensembles	0.029	0.027	0.041 ± 0.033
Supervised tree ensembles	0.031	0.031	0.064 ± 0.076
Supervised RBF-based ANNs	0.23	0.02	0.078 ± 0.085
Bayesian networks	0.32	0.0047	0.117 ± 0.127
Log-MVND	0.198	0.121	0.031 ± 0.03

The absolute correlation difference between the real and virtual data that were generated by the unsupervised tree ensembles is depicted in Figure 73, where the average difference was 0.041 ± 0.033. Regarding the rest of the algorithms, the average correlation difference was 0.064 ± 0.076 for the supervised tree ensembles,

0.078±0.085 for the RBF-based ANNs, 0.117±0.127 for the Bayesian networks, and 0.031±0.03 for the Log-MVND.

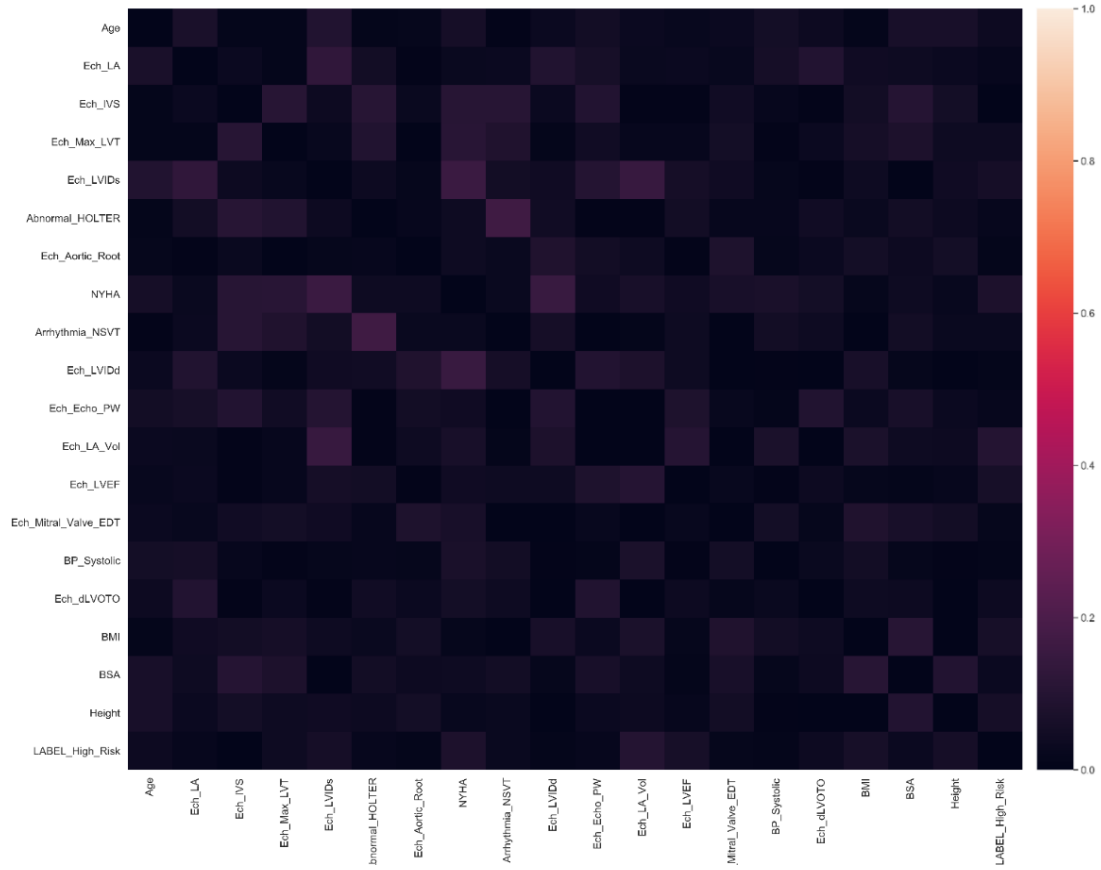


Figure 73. The absolute difference between the real and virtual correlation matrices for the HCM dataset, in the case of the unsupervised tree ensembles generator. The features are ordered based on their appearance in Supplementary Table **1**.

Although the Log-MVND schema achieved the smallest inter-correlation difference from the virtual population generators, it yielded significantly higher GOF and KL values than the unsupervised tree ensembles. The dark color pattern in Figure 73 denotes the absence of significant correlation differences between the real and the virtual data which suggests that in this case the unsupervised tree ensembles schema was able to generate virtual distributions with increased similarity (i.e., with highly similar correlation patterns) with the real distributions.

In this case, class imbalance handling is not required since the ratio of the patients with low and high risk for HCM (Table 38) is adequate. The application of the XGBoost on the real data using a 10-fold cross validation process yielded accuracy 0.597, sensitivity 0.564, specificity 0.708, and AUC 0.628. According to Table 39, the average

performance of the XGBoost on the aggregated real and virtual data from the unsupervised tree ensembles achieved better classification performance, yielding accuracy 0.758, sensitivity 0.733, specificity 0.845, and AUC 0.829. The performance of the XGBoost on the augmented data from the supervised tree ensembles comes next along with the RBF-based ANNs, the Bayesian networks and the Log-MVND which achieved slightly better performance than before but with less than 0.6 sensitivity and thus are excluded from Table 39.

Table 38. A summary of the data quality report.

Metadata	**UoA cohort**	**UNIFI cohort**
Number of features	162	123
Number of records (instances)	449	2454
Number of discrete features	58	37
Number of continuous features	73	86
Number of unknown features	31	0
Number of features with outliers	16	0
Number of features with inconsistencies	19	0
Number of bad quality features	77	36
Number of fair quality features	57	42
Number of good quality features	26	45
Class imbalance ratio	1:6.4	1:1.58
Final number of patients after class imbalance handling	210	776
Final number of acceptable features	65	20

The performance of the HCM risk stratification models from the AdaBoost and Random Forests using the augmented data from the tree ensembles was also higher than in the case of the real data. The application of the AdaBoost on the real data yielded accuracy 0.61, sensitivity 0.569, specificity 0.748, and AUC 0.611. According to Table 39, the average performance of the AdaBoost on the aggregated data from the unsupervised tree ensembles achieved better classification performance, yielding accuracy 0.665, sensitivity 0.622, specificity 0.811, and AUC 0.712.

As for the Random Forests, their application on the real data yielded accuracy 0.629, sensitivity 0.563, specificity 0.853, AUC 0.641, whereas the average performance on the aggregated real and virtual data from the unsupervised tree ensembles achieved better performance, yielding accuracy 0.723, sensitivity 0.664, specificity 0.925, and AUC 0.763.

Table 39. A summary of the HCM risk stratification results from the XGBoost, AdaBoost and Random Forests before and after data augmentation using the virtual data from each virtual population generator.

Virtual population generation method for data augmentation	HCM risk stratification performance			
	accuracy	sensitivity	specificity	AUC
XGBoost				
Before data augmentation	0.597	0.564	0.708	0.628
Unsupervised tree ensembles	0.758	0.733	0.845	0.829
Supervised tree ensembles	0.705	0.672	0.817	0.753
AdaBoost				
Before data augmentation	0.61	0.569	0.748	0.611
Unsupervised tree ensembles	0.665	0.622	0.811	0.712
Supervised tree ensembles	0.653	0.606	0.816	0.672
Random Forests				
Before data augmentation	0.629	0.563	0.853	0.641
Unsupervised tree ensembles	0.723	0.664	0.925	0.763
Supervised tree ensembles	0.686	0.621	0.908	0.705

The ROC curves are summarized in Figure 74, highlighting the classification performance of the unsupervised tree ensembles which yielded an increase by 16.1% in the accuracy, 16.9% in sensitivity, 13.7% in specificity, and 20.1% in AUC compared with the XGBoost trained on the real data.

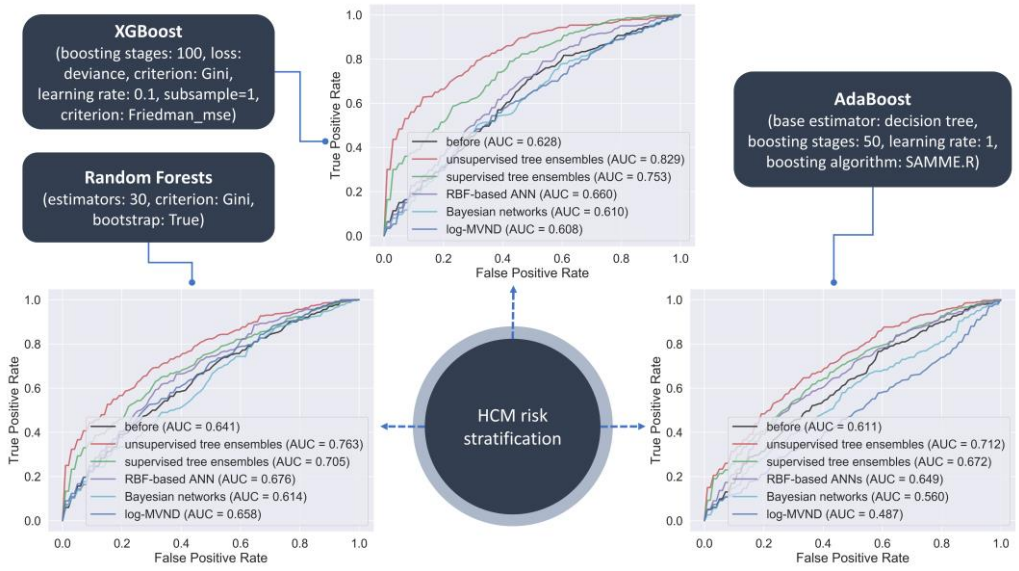


Figure 74. ROC curves depicting the classification performance of the XGBoost, the AdaBoost and the Random Forests for HCM risk stratification with and without data augmentation.

A similar increase is also observed in the case of the AdaBoost (5.5% in accuracy, 5.3% in sensitivity, 6.3% in specificity, and 10.1% in AUC), as well as, in the Random Forests (9.4% in accuracy, 10.1% in sensitivity, 7.2% in specificity, and 12.2% in AUC). Although the classification performance is significantly smaller than in pSS, data augmentation was still able to enhance the performance of the HCM risk stratification models. All in all, we examined the effectiveness of data augmentation in terms of enhancing the real clinical research databases with high-quality virtual data to enhance the performance of the disease classification and risk stratification models in hypertrophic cardiomyopathy. To do so, a computational pipeline was developed, where high-quality virtual data are aggregated with the real data to yield robust HCM risk stratification models, where the performance of each model was evaluated on testing instances of the real data to avoid any biases. The proposed pipeline was able to generate virtual distributions with increased similarity, correlation, and reduced divergence with the real distributions.

The aggregation of the real with the virtual patient data in both clinical domains yielded a notable increase in the classification accuracy, sensitivity, specificity, and area under the curve scores of the supervised machine learning models which were trained on the augmented clinical data compared to those trained on real data instances. Moreover, the performance of the HCM risk stratification model showed an increase by accuracy, 16.9% in sensitivity, 13.7% in specificity, and 20.1% in area under the curve against the one trained on the real HCM data (Table 39, Figure 74). A similar increase is also observed in the case of the AdaBoost (5.5% in accuracy, 5.3% in sensitivity, 6.3% in specificity, and 10.1% in AUC), as well as, in the case of the Random Forests (9.4% in accuracy, 10.1% in sensitivity, 7.2% in specificity, and 12.2% in AUC). The aggregation of the virtual data from the supervised tree ensembles with the real patient data yielded enhanced classification models at a similar extent (see Table 39, Figure 74 for HCM risk stratification).

The aggregation of the virtual data from the supervised RBF-based ANNs, the Bayesian networks and the Log-MVND with the real one yielded supervised machine learning models with partially enhanced performance while maintaining the increased performance than in the case of training on the real data only. All in all, our results validate the scientific and technical impact of data augmentation in the classification

accuracy, sensitivity, and specificity for HCM risk stratification models. To our knowledge, this is the first study that builds a computational pipeline which uses the high-quality semi-artificial patient data which are generated by machine learning-based approaches to enhance the performance of risk stratification models in HCM.

7.3. Cardiovascular diseases

This case study involves the application of the beyond the state of the art methods that were developed for data harmonization (CHAPTER 4) to address open issues and unmet needs (Section 1.4) in the domain of the cardiovascular diseases (Section 2.3.5).

7.3.1. Data harmonization

7.3.1.1. Case Study 1 – A reference model for CVD

The scope of this case study is to develop a "gold" reference ontology for the CVD domain. Metadata were obtained from 3 clinical centers on CVD, namely the TAUH (Tampere University Hospital), UMCU (University Medical Center Utrecht), and LURIC (LUdwigshafen RIsk and Cardiovascular Health). The number of terminologies was 6408 in TAUH, 1545 in LURIC, and 137 in UMC. The large number of terminologies in TAUH and LURIC included metanalysis results and data from multiple timepoints. The entity graph for the CVD ontology is depicted in Figure 75 and was organized according to the available information in datasets TAUH (6408 terminologies), UHEI_LURIC (1545 terminologies) and UMC (137 terminologies).

According to the entity graph, the hierarchy of the ontology is designed as follows:

- The main class is the class "Patient" who "has" the following subclasses:
    - The subclass "Demographics" which "includes":
        - The subclass "Basic information" describing urine related information.
        - The subclass "Alcohol consumption" describing blood test repeated information.
        - The subclass "Smoking status" describing MD related information.
    - The subclass "Laboratory tests" which describes various laboratory measures.

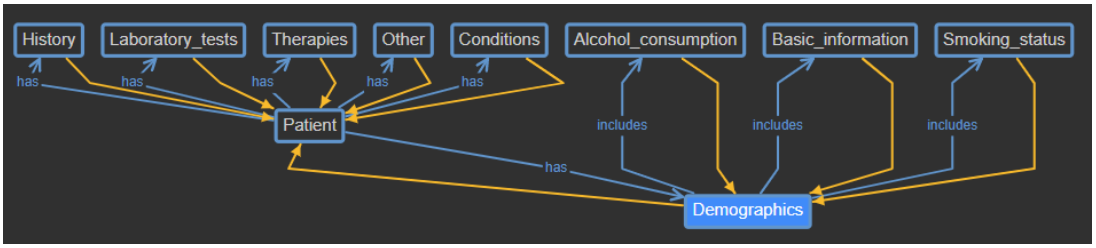


Figure 75. An instance of the entity graph for the CVD ontology from WebProtégé [403].

- The subclass "Conditions" which describes various medical conditions.
- The subclass "Therapies" which describes various medications.
- The subclass "History" which describes family history.
- The subclass "Other" which describes any other related information.

Another illustration of the CVD's entity graph is presented in Figure 76 using WebVOWL [404]. All in all, this is a first implementation of an FHIR-compliant, reference ontology for the domain of cardiovascular diseases based on a large number of terminologies from the TAUH center. The ontology is open and can be found in the following link: [GitHub - vpz4/TO_AITION: [TO_AITION] Preliminary versions of the CVD and Mental Disorders' ontologies](GitHub - vpz4/TO_AITION: [TO_AITION] Preliminary versions of the CVD and Mental Disorders' ontologies).

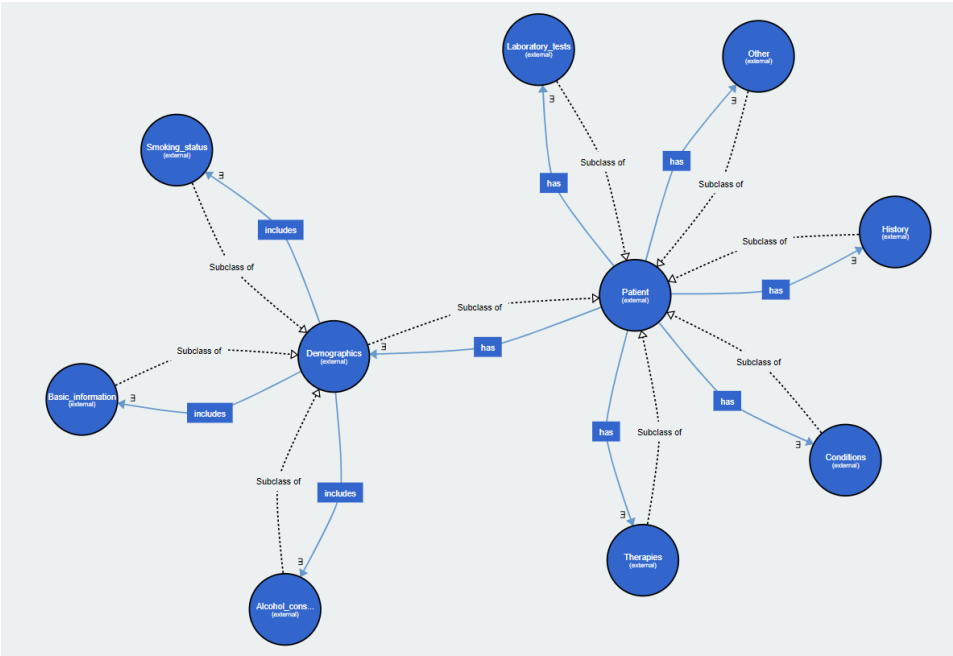


Figure 76. The first-degree hierarchy in the CVD reference ontology from WebVOWL [404].

7.3.1.2. Case Study 2 – Harmonization across 3 European centers

The scope of this case study is to harmonize three European cohorts in the domain of CVD using the proposed hybrid data harmonizer (Section 4.5) using the reference model for the CVD which has been developed in the previous section. Metadata were obtained from 3 clinical centers on CVD, namely the TAUH (Tampere University Hospital), UMCU (University Medical Center Utrecht), and LURIC (LUdwigshafen RIsk and Cardiovascular Health). The number of terminologies was 6408 in TAUH, 1545 in LURIC, and 137 in UMC. The large number of terminologies in TAUH and LURIC, included metanalysis results and data from multiple timepoints.

The proposed method was able to identify 85% (in average) of the relevant terminologies within the reference ontologies, on each domain, against simple lexical analysis which yielded 10% less precision. Regarding the individual analysis (Table 40), the hybrid approach identified 435 matches in LURIC, and 71 in UMCU yielding 15% more terminologies (in average) than lexical analysis only. Any identified prominent match in the range [0.6, 0.8) is considered as partially matched, whereas a match in the range [0.8, 1) is considered highly similar (Table 40).

Table 40. A summary of the individual and cross-domain analysis results using the proposed method for data harmonization.

Pair	**Matched terminologies**	**Exact**	**Partial**	**Highly-similar**
(OCVD, LURIC)*	435	6	332	103
(OCVD, UMCU)*	71	4	25	46
* CVD domain, OCVD = reference Ontology for CVD.				

The results of the lexical matching algorithm are summarized in Supplementary Table 2, Table 41, and Table 42 for the pairs (OCVD, UHEI_LURIC), (OCVD, UMC), and (UHEI_LURIC, UMC). The number of terminologies in TAUH was 6408 whereas in LURIC_UHEI the number of terminologies was 1545, including metanalysis results and information across multiple time points. Regarding UMC, the number of terminologies was 137. Since the CVD ontology was developed according to the structure of TAUH, the intersection of the pairwise lexical matches between (OCVD, UHEI_LURIC) and (TAUH, UMC) are the final matched terminologies.

According to Supplementary Table 2, our approach was able to identify 435 potential lexical matches between TAUH and UHEI_LURIC with more than 75% coherence. More specifically, the lexical matcher was able to identify 6 terminologies with exact similarity (i.e., matching score 1), as well as, 28 terminologies with matching score larger than (or equal to) 0.9, and 103 terminologies with matching score larger than (or equal to) 0.8. The 332 terminologies with scores in the range [0.6, 0.8) are considered as partially matched whereas the remaining 103 terminologies with scores in the range [0.8, 1) are considered as highly similar.

According to Table 41, the proposed approach was able to identify 71 potential lexical matches between OCVD and UMC with more than 75% coherence (overlap). More specifically, the lexical matcher was able to identify 4 terminologies with exact similarity (i.e., matching score 1), as well as, 25 terminologies with matching score larger than (or equal to) 0.9, and 46 terminologies with matching score larger than (or equal to) 0.8. The 25 terminologies with scores in the range [0.6, 0.8) are considered as partially matched whereas the remaining 46 terminologies with scores in the range [0.8,1) are considered as highly similar.

Table 41. A summary of the potentially matched terminologies between OCVD and UMC.

Terminologies from OCVD	**Terminologies from UMC**	**Score**
Interleukin-6	Interleukin 6	0.949
Father: myocardial infarction, I1, 191, 1 = no, 2 = yes	Myocardial infarction	0.647
Uncle or aunt diagnosed with other cardiovascular disease	Vascular endothelial growth factor A	0.646
Body Mass Index	body mass index	0.867
body-mass index	body mass index	0.896
Interleukin-9	Interleukin 9	0.949
Apo	Apolipoprotein B	0.729
Myocardial infarction, 0 = no, 1 = yes	Myocardial infarction	0.851
Apolipoprotein B; g/l	Apolipoprotein B	0.921
triglyceride	Triglycerides [mg/dL]	0.813
Stroke, including TIA, excl. all haemorrhages	Stroke	0.711
Monokine induced by interferon-gamma	Monokine induced by interferon-gamma	1.000
Monocyte chemotactic protein-1	Monocyte chemotactic protein 1	0.978

Terminologies from OCVD	Terminologies from UMC	Score
Body mass indexi	body mass index	0.936
Body mass index; weight/	body mass index	0.839
Interleukin-8	Interleukin 8	0.949
Stroke, including TIA	Stroke	0.762
Mother: myocardial infarction, 1 = no, 2 = yes	Myocardial infarction	0.668
Vascular endothelial growth factor	Vascular endothelial growth factor A	0.981
Pulse pressure	pulse pressure	0.875
LDL-cholesterol	LDL-cholesterol [mg/dL]	0.884
Father: myocardial infarction, I1,, 1 = no, 2 = yes	Myocardial infarction	0.657
Apolipoprotein B	Apolipoprotein B	1.000
Myocardial infarction - Year when first time diagnosed	Myocardial infarction	0.796
Interferon gamma-induced protein 10	Interferon gamma	0.819
Interleukin-5	Interleukin 5	0.949
Stroke, including SAH	Stroke	0.762
Glucose; mmol/l	Glucose [mmol/L]	0.893
P –Osteopontin; ng/ml. Limit of detection is 8. Values below this were measured by diluting the sample.	Osteopontin	0.702
Father: myocardial infarction, 1 = no, 2 = yes	Myocardial infarction	0.671
Apolipoprotein A-1	Apolipoprotein B	0.924
Creatinine value upon hospital admission	Composite cardiovascular endpoints	0.669
Body mass index at final measurements	body mass index	0.771
Gender, 1 = female, 2 = male	gender	0.671
Hematocrit; Osuus	Hematocrit	0.863
Number of myocardial infarctions	Myocardial infarction	0.701
Myocardial infarction	Myocardial infarction	1.000
Body mass index at initial measurements	body mass index	0.764
Homocysteine; µmol/l	Homocysteine [umol/L]	0.887
Triglyceride; mmol/l	Triglycerides [mmol/L]	0.906
Coronary artery disease - Year when first time diagnosed	Coronary artery disease	0.804
Interleukin-16	Interleukin 6	0.927
Major Adverse Cardiovascular Event	Major cardiovascular events	0.781
Stroke, including SAH and TIA	Stroke	0.736
Interleukin-4	Interleukin 4	0.949

Terminologies from OCVD	Terminologies from UMC	Score
Lipoprotein	Apolipoprotein B	0.761
Apolipoprotein E	Apolipoprotein B	0.958
Myocardial infarction, without ST-evelation	Myocardial infarction	0.829
Glucose	Glucose [mmol/L]	0.813
Interleukin-10	Interleukin 10	0.952
Body mass index weight/	body mass index	0.847
Interleukin-18	Interleukin 8	0.927
Interleukin-3	Interleukin 13	0.927
Myocardial infarction, strict	Myocardial infarction	0.908
Body mass index; kg/m2	body mass index	0.857
Total cholesterol	Total cholesterol [mg/dL]	0.893
Interleukin-2	Interleukin 2	0.949
Interleukin-13	Interleukin 13	0.952
Mother: myocardial infarction, I1,, 1 = no, 2 = yes	Myocardial infarction	0.653
Triglyceride	Triglycerides [mg/dL]	0.857
Interleukin-15	Interleukin 5	0.927
Stroke, excluding SAH	Stroke	0.762
Myocardial infarction, unclassifiable	Myocardial infarction	0.856
Diabetes type 2	type 2 diabetes status	0.608
HDL-cholesterol	HDL-cholesterol [mg/dL]	0.884
Myocardial infarction, with ST-evelation	Myocardial infarction	0.842
Mother: myocardial infarction, I1, 190, 1 = no, 2 = yes	Myocardial infarction	0.643
stroke volume	Stroke	0.739
Interleukin-17	Interleukin 10	0.905
Macrophage migration inhibitory factor	Macrophage migration inhibitory factor	1.000
Stroke - Year when first time diagnosed	Stroke	0.718

According to Table 42, the proposed approach was able to identify 46 potential lexical matches between UHEI_LURIC and UMC with more than 75% coherence (overlap). More specifically, the lexical matcher was able to identify 1 terminology with exact similarity (i.e., matching score 1), as well as, 18 terminologies with matching score larger than (or equal to) 0.9, and 32 terminologies with matching score larger than (or equal to) 0.8. The 14 terminologies with scores in the range [0.6, 0.8) are considered as partially matched whereas the remaining 32 terminologies with scores in the range [0.8,1) are considered as highly similar.

Table 42. A summary of the potentially matched terminologies between UHEI and UMC.

Terminologies from UHEI_LURIC	Terminologies from UMC	Score
apolipoprotein A-I	Apolipoprotein B	0.884
fasting glucose	age at inclusion	0.655
interleukin 1ß	Interleukin 10	0.905
Creatin [µmol/L]	Creatinin [umol/L]	0.924
interleukin 1	Interleukin 10	0.927
interleukin 6	Interleukin 6	0.949
total cholesterol	Total cholesterol [mg/dL]	0.840
myeloperoxidase	Myeloperoxidase in citrate plasma [ng/ml]	0.758
LDL apolipoprotein B	Apolipoprotein B	0.896
cholesterol	LDL-cholesterol [mg/dL]	0.826
enddiastolic lv pressure	Composite cardiovascular endpoints	0.634
interleukin 2	Interleukin 2	0.949
interleukin 12	Interleukin 12	0.952
body mass index	body mass index	1.000
vitamin B12	Vitamin B12 [pmol/L]	0.803
Creatinin [µmol/L]	Creatinin [umol/L]	0.963
Stroke/PRIND/TIA	Stroke	0.792
aldosterone	Aldosteron	0.906
interleukin 8	Interleukin 8	0.949
homocysteine	Homocysteine [umol/L]	0.813
cystatin C clearance	Cystatin C	0.783
interleukin 10	Interleukin 10	0.952
lipoprotein	Apolipoprotein B	0.805
creatinin	Creatinin [umol/L]	0.778
cysteine	Homocysteine [umol/L]	0.794
Cystatin clearance 100/cystatin C	Cystatin C	0.724
stenosis of carotid artery	Coronary artery disease	0.631
age	age at inclusion	0.729
interleukin 9	Interleukin 9	0.949
renin	Renine	0.822
vitamin B1	Vitamin B12 [pmol/L]	0.783
interleukin 4	Interleukin 4	0.949
apolipoprotein E	Apolipoprotein B	0.917
apolipoprotein A-II	Apolipoprotein B	0.871
hematocrit	Hematocrit	0.933
HDL-cholesterol	HDL-cholesterol [mg/dL]	0.884

Terminologies from UHEI_LURIC	Terminologies from UMC	Score
cystatin C	Cystatin C	0.933
triglycerides	Triglycerides [mg/dL]	0.832
total protein	Apolipoprotein B	0.622
hemoglobin	Hemoglobin [mmol/L]	0.791
VLDL apolipoprotein B	Apolipoprotein B	0.884
LDL-cholesterol	LDL-cholesterol [mg/dL]	0.884
myocardial infarction	Myocardial infarction	0.968
apolipoprotein C-II	Apolipoprotein B	0.871
apolipoprotein B	Apolipoprotein B	0.958
Platelets	anti-platelets use	0.685

The terminology in OCVD having the highest frequency in UHEI_LURIC is related to "cholesterol" with 47 occurrences, in total. The terminology "triglycerides" comes next with 35 occurrences, along with the terminologies "cholesterol ester", "free cholesterol" (with 28 occurrences) and myocardial infarction (16 occurrences). Other terminologies with an adequate number of occurrences, include the "age", "heart rate", and "phospholipid", among others. The terminology in OCVD having the highest frequency in UHEI_LURIC is related to "Myocardial infarction" with 14 occurrences, in total. The terminology "body mass index" comes next with 8 occurrences, along with the terminologies "Stroke" (7 occurrences) and "Apolipoprotein B" (6 occurrences). Other terminologies with an adequate number of occurrences, include the "Interleukin 5", "Interleukin 10", and "Interleukin 13", among others. The terminology in UHEI_LURIC having the highest frequency in UMC is related to "Apolipoprotein B" with 9 occurrences, in total. The terminologies "Interleukin 5", "Creatinin [umol/L]", and "Cystatin C" come next with 3 occurrences, along with the terminologies "LDL-cholesterol [mg/dL]", "Vitamin B12 [pmol/L]" and "Homocysteine [umol/L]" with 2 occurrences each.

We proposed an automated data harmonization workflow which adopts a hybrid strategy that combines lexical analysis with semantic models (ontologies) to identify terminologies with lexical and conceptual overlap. The proposed method was used to match terminologies in cardiovascular disease (CVD), yielding matched terminologies with 85% overlap and 10% higher performance than conventional lexical analysis, as well as, in favorable execution time against manual terminology mapping. The hybrid data harmonization workflow combines lexical analysis with relational modeling to

overcome structural heterogeneities that obscure the interlinking of retrospective data from multiple clinical databases without standardized data collection protocols. The proposed method uses a hybrid approach which utilizes lexical and semantic analysis to identify terminologies with common conceptual and lexical basis. Our method was compared against the application of conventional lexical analysis and manual mapping across 8090 terminologies from the CVD domain. The computational complexity of the method was proportional to the number of input terminologies. In some cases, the number of matched terminologies was high (e.g., 435 in LURIC) enough. In addition, the proposed method for data harmonization can be applied to any clinical domain given a reference ontology as input. The overall value of the proposed method lies on the fact that it can be used to deal with open issues and unmet needs in various clinical domains which enhances its scientific and clinical impact. As a future work, we plan to apply the proposed method across multiple datasets in other domains, as well as, evaluate the consistency of the harmonized data after the execution of the data harmonizer.

7.4. Mental disorders

This case study involves the application of the beyond the state of the art methods that were developed for data harmonization (CHAPTER 4) to address open issues and clinical unmet needs (Section 1.4) in the domain of the mental disorders (Section 2.3.6).

7.4.1. Data harmonization

7.4.1.1. Case Study 1 – A reference model for mental disorders

The scope of this case study is to develop a "gold" reference model for mental disorders. Metadata were obtained from 3 clinical centers on mental disorders, namely the LODZ (Medical University of Łódź), NESDA (Netherlands Study of Depression and Anxiety), and UVA (Universiteit van Amsterdam). The number of terminologies was 43 in LODZ, 49 in NESDA, and 906 in UVA. The large number of terminologies in UVA included metanalysis results and data from multiple timepoints. The entity graph for the mental disorders' ontology is depicted in Figure 77 and was organized according to the available information in datasets LODZ (43 terminologies), NESDA (49 terminologies) and UVA (906 terminologies. According to the entity graph, the hierarchy of the ontology has been designed as follows:

- The main class is the class “Patient” who “has” the following subclasses:
  - The subclass “Demographics” describing demographic related information.
  - The subclass “Laboratory tests” which “includes”:
    - The subclass “Urine tests” describing urine related information.
    - The subclass “Blood tests” describing blood test repeated information.
    - The subclass “Mental tests” describing MD related information.
  - The subclass “Medical Conditions” describing various medical conditions.
  - The subclass “Medications” which describes various medications.
  - The subclass “Other” which describes any other related information.

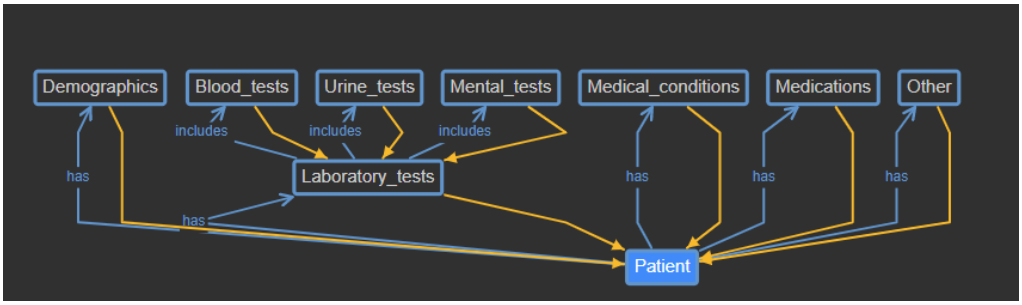


Figure 77. An instance of the entity graph for the MD ontology from WebProtégé [403].

An illustration of the MD entity graph is depicted in Figure 78 using WebVOWL [404].

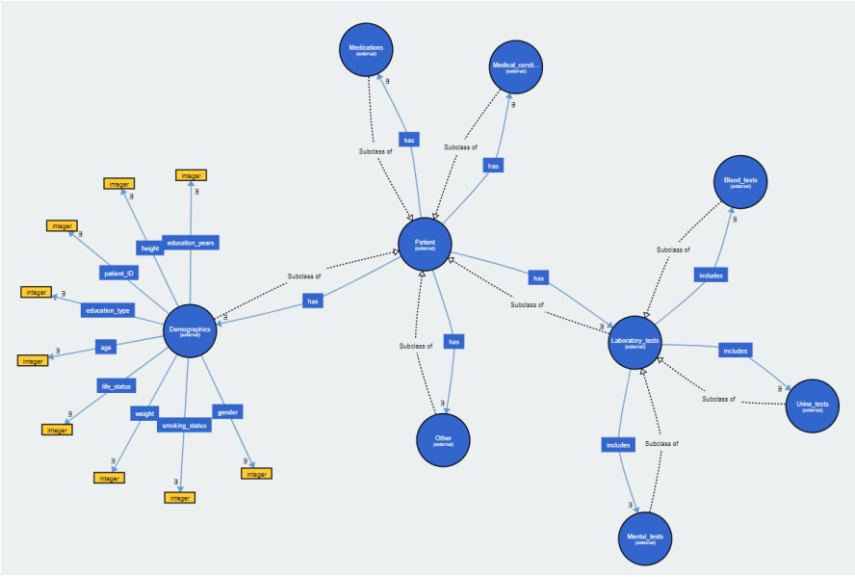


Figure 78. The first-degree hierarchy in the MD ontology from WebVOWL [404].

This is a first implementation of an ontology for mental disorders. The ontology can be found in the following link: GitHub - vpz4/TO_AITION: [TO_AITION] Preliminary versions of the CVD and Mental Disorders' ontologies.

7.4.1.2. Case Study 2 – Individual and cross-domain data harmonization

The scope of this case study is to harmonize 3 European cohorts in MD using the proposed hybrid data harmonizer. Metadata were obtained from 3 clinical centers on mental disorders, namely the LODZ (Medical University of Łódź), NESDA (Netherlands Study of Depression and Anxiety), and UVA (Universiteit van Amsterdam). The number of terminologies was 43 in LODZ, 49 in NESDA, and 906 in UVA. The large number of terminologies in UVA included metanalysis results and data from multiple timepoints. The proposed method was able to identify 85% (in average) of the relevant terminologies within the reference ontologies, on each domain, against simple lexical analysis which yielded 10% less precision.

Regarding the individual analysis (Table 43), the hybrid approach identified 435 matches in LURIC, and 71 in UMCU yielding 15% more terminologies (in average) than lexical analysis only. In the mental disorders' domain, the number of matches was 25 for NESDA and 14 for LODZ, where the small number of matches reflects the presence of CVD-oriented information in the mental disorders' centers. To deal with this issue, the hybrid approach was applied in a cross-domain manner, yielding 81 matches in LODZ (4 exact, 16 partially matched, 16 highly similar), 139 in NESDA (4 exact, 90 partially matched, 49 highly similar), and 288 in UVA (6 exact, 222 partially matched, 66 highly similar). Any identified prominent match in the range [0.6, 0.8) is considered as partially matched, whereas a match in the range [0.8, 1) is considered highly similar (Table 43).

Table 43. A summary of the individual and cross-domain analysis results.

Pair	Matched terms	Exact	Partial	Highly-similar
(OMD, NESDA)**	25	2	9	7
(OMD, LODZ)**	14	1	2	9
(OCVD, LODZ)***	81	4	65	16
(OCVD, NESDA)***	139	4	90	49
(OCVD, OMD)***	288	6	222	66
** mental disorders domain, *** cross-domain, OCVD = Ontology for CVD, OMD = Ontology for the mental disorders' domain.				

The results are summarized in Table 44, Table 45 and Table 46 for the pairs (LODZ, OMD), (NESDA, LODZ), and (NESDA, OMD). The number of terminologies in LODZ was 43 whereas in NESDA the number of terminologies was 49. Regarding UVA, the number of terminologies was 906, including metanalysis results and information across multiple time points. Since the mental disorders' ontology (7.4.1.1) was developed according to the structure of LODZ, the intersection of the pairwise lexical matches between (LODZ, OMD) and (NESDA, LODZ) are the final matched terminologies.

According to Table 44, the proposed approach was able to identify 14 potential lexical matches between LODZ and OMD with more than 75% coherence (overlap). More specifically, the lexical matcher was able to identify 1 terminology with exact similarity (i.e., matching score 1), as well as, 3 terminologies with matching score larger than (or equal to) 0.9, and 9 terminologies with matching score larger than (or equal to) 0.8. The 3 terminologies with scores less than 0.6 are considered as lexically non-similar, whereas the 2 terminologies with scores in the range [0.6, 0.8) are considered as partially matched whereas the remaining 9 terminologies with scores in the range [0.8,1) are considered as highly similar.

Table 44. A summary of the potentially matched terminologies between LODZ and UVA.

Terminologies from LODZ	**Terminologies from OMD**	**Score**
Creatinine	creatinine	0.864
NYHA	Heart failure, NYHA stages	0.428
sRAGE	Age	0.511
Hypertension	hypertension	0.883
The degree of severity of ischemic heart disease and heart failure	Diagnostics / medical history of cardiovascular risk	0.633
Age	Age	1.000
Gender	gender	0.889
Urea	urea	0.833
Education	Participation in patient education	0.503
Pacemaker	Pacemaker	0.917
Height	height	0.889
Diabetes	diabetes	0.837
Weight	Weight	0.952
Atrial fibrillation	atrial fibrillation	0.792

According to Table 45, the proposed approach was able to identify 7 potential lexical matches between NESDA and LODZ with more than 75% coherence (overlap). More specifically, the lexical matcher was able to identify 2 terminologies with exact similarity (i.e., matching score 1), as well as, 4 terminologies with matching score larger than (or equal to) 0.9, and 5 terminologies with matching score larger than (or equal to) 0.8. The terminology "Sex" in NESDA has a score 0.5 with the terminology "Gender" in LODZ which validates the lexical dissimilarity but they have been correctly identified as matched by the lexical matcher since they are synonyms. The 1 terminology ("smoking") with score in the range [0.6, 0.8) is considered as partially matched whereas the remaining 5 terminologies with scores in the range [0.8,1) are considered as highly similar.

Table 45. A summary of the potentially matched terminologies between NESDA and LODZ using the proposed approach.

Terminologies from NESDA	**Terminologies from LODZ**	**Score**
education	Education	0.926
Height	Height	1.000
Weight	Weight	0.952
Sex	Gender	0.500
Diabetes Type 2	Diabetes	0.844
Hypertension	Hypertension	1.000
smoking	Smoking tobacco	0.752

According to Table 46, the proposed approach was able to identify 25 potential lexical matches between NESDA and OMD with more than 75% coherence (overlap). More specifically, the lexical matcher was able to identify 2 terminologies with exact similarity (i.e., matching score 1), as well as, 4 terminologies with matching score larger than (or equal to) 0.9, and 7 terminologies with matching score larger than (or equal to) 0.8. The 9 terminologies with scores less than 0.6 are considered as lexically non-similar apart from the terminologies "Sex" and "gender" which might be lexically non-similar, but they have been correctly identified by the lexical matcher as synonyms. Finally, the 9 terminologies with lexical matching scores in the range [0.6, 0.8) are considered as partially matched whereas the remaining 7 terminologies with scores in the range [0.8,1) are considered as highly similar.

Table 46. A summary of the potentially matched terminologies between NESDA and OMD.

Terminologies from NESDA	Terminologies from OMD	Score
Heart rate variability	heart rate	0.770
Cardiovascular disease	last RR value diastolic	0.530
education	Participation in patient education	0.503
Heart rate	heart rate	0.933
Systolic blood pressure	Target systolic blood pressure	0.708
Total cholesterol levels	Total cholesterol	0.903
Triglyceride levels	triglycerides	0.794
Antidepressant use	antidepressants	0.721
sleep	sleep disorders	0.759
Use of oral anti-diabetic medication or insulin	medication	0.412
physical activity	physical activity	1.000
Use of antihypertensive medication	medication	0.453
Height	height	0.889
Weight	Weight	1.000
Sex	gender	0.500
Diabetes Type 2	diabetes type	0.683
Metabolic Syndrome	metabolic syndrome	0.790
Diastolic blood pressure	Target diastolic blood pressure	0.719
Other psychotropic medication use	medication	0.382
Anti-inflammatory medication use	medication	0.385
Angina pectoris	stable angina pectoris	0.439
Use of cholesterol-lowering medication	medication	0.349
Hypertension	hypertension	0.883
Glucose levels	glucose	0.762
smoking	smoking status	0.833

The terminology in LODZ having the highest frequency in OMD is related to “Age” with 2 occurrences. The terminologies “creatinine”, “Heart failure, NYHA stages”, “hypertension”, “diabetes”, and “atrial fibrillation” come next having the same frequency with the rest of the terminologies. The terminologies in NESDA having the same frequency in LODZ are related to the terminologies “Height”, “Weight”, “Gender”, “Diabetes”, “Hypertension”, and “Smoking tobacco” in NESDA. The terminology in NESDA having the highest frequency in OMD is related to “medication” with 5 occurrences. The terminology “heart rate” comes next (with 2 occurrences). The rest of the terminologies had the same frequency, including the “last

RR value diastolic", "Participation in patient education", "Target systolic blood pressure", "Total cholesterol", "triglycerides", "hypertension", and "antidepressants", among others. Most of the terminologies are related to CVD oriented terminologies, a fact which has been further investigated in the next Section through the application of cross domain matches.

By taking into consideration the fact that the number of occurrences was low in the pairs of datasets regarding the mental disorders' domain, as well as, the fact that the majority of the terminologies in UVA and NESDA include multiple CVD-related terminologies (e.g., "systolic blood pressure", "cholesterol", "triglycerides", "hypertension"), cross domain matches were also applied. The results of the potentially matched terminologies between the pairs of cross domain datasets (TAUH, LODZ), (TAUH, NESDA), and (TAUH, UVA) are summarized in Table 47, Table 48 and Supplementary Table 3, respectively.

According to Table 47, the proposed approach was able to identify 81 potential lexical matches between TAUH and LODZ with more than 75% coherence (overlap). More specifically, the lexical matcher was able to identify 4 terminologies with exact similarity, as well as, 16 terminologies with matching score larger than (or equal to) 0.8. The 65 terminologies with scores in the range [0.6, 0.8) are considered as partially matched whereas the remaining 16 terminologies with scores in the range [0.8,1) are considered as highly similar.

Table 47. A summary of the potentially matched terminologies between TAUH and LODZ using the proposed approach.

Terminologies from TAUH	**Terminologies from LODZ**	**Score**
Age of youngest child in the family, I2, 93	Age	0.690
Age-adjusted expected maximum HR	Age	0.698
Age of starting habitual smoking	Age	0.698
Diabetes - Year when first time diagnosed	Diabetes	0.730
Participates in some instructed physical activity	Chronic obstructive pulmonary disease	0.631
Age of starting regular sports training	Age	0.692
Diabetes; father, I1, 218, 1 = no, 2 = yes, 9 = mother is missing from family	Diabetes	0.701
Age of youngest child in the family, I2, 49	Age	0.690

Terminologies from TAUH	Terminologies from LODZ	Score
Chronic obstructory pulmonary disease	Chronic obstructive pulmonary disease	0.868
Age when first trying alcohol, I2, 810, 99 = has never tried	Age	0.683
Indication for the test, evaluation of arrhythmia	Other diseases of the central nervous system	0.651
Age when first trying smoking, I2, 780, 99 = has never smoked	Age	0.683
Asthma; father, I1, 214	Asthma	0.754
Age of oldest child in the family, I2, 51	Age	0.691
Asthma, I1, 77, 1 = no, 2 = yes	Asthma	0.731
Age at death of uncle or aunt died of myocardial infarction	Age	0.684
Age at time of action	Age	0.714
Diabetes type 2	Diabetes	0.844
Diabetes, type 1	Diabetes	0.833
Education years of parents, Education years have been created by using information of years 1980, 1983 and 1986 measurements. Variable vkoulv is the maximum of parent's school years from those years.	Education	0.682
Creatinine in different unit.	Creatinine	0.782
Hypertension - Year when first time diagnosed	Hypertension	0.756
? Age when celiac disease observed	Age	0.696
Height	Height [cm]	0.848
Age of starting habitual use of alcohol, I2, 812, 98 = does not use alcohol habitually, 99 = has never tried	Age	0.676
Diabetes; mother, I1, 217, 1 = no, 2 = yes, 9 = mother is missing from family	Diabetes	0.701
age in 2007	Age	0.616
Mothers daily consumption of bread	Chronic obstructive pulmonary disease	0.711
Fathers daily consumption of bread	Chronic obstructive pulmonary disease	0.693
Age of starting habitual use of alcohol	Age	0.692

Terminologies from TAUH	Terminologies from LODZ	Score
Any other significant heart disease that may have caused sudden death eg aortic stenosis, HCM, DCM	The degree of severity of ischemic heart disease and heart failure	0.654
TIMI classification of the second PCI session in the target vessel of the first PCI after the procedure	The degree of severity of ischemic heart disease and heart failure	0.604
TIMI classification in the target vessel of the second PCI after the procedure	The degree of severity of ischemic heart disease and heart failure	0.701
Atrial fibrillation	Atrial fibrillation	1.000
Creatinine. Same procedure as in 2001; µmol/l	Creatinine	0.741
Age of starting regular sports training, I2,1021	Age	0.688
Age when first trying alcohol	Age	0.701
Hypertension/Hypertonia	Hypertension	0.841
People have recommended that I should tell more about my feelings	The degree of severity of ischemic heart disease and heart failure	0.649
diabetes diet , 0 = no, 1 = yes	Diabetes	0.700
Age at DEATH	Age	0.750
Lives abroad, so the accuracy of death data may be incomplete	Lives	0.694
TIMI classification in the third PCI target vessel after the procedure	The degree of severity of ischemic heart disease and heart failure	0.703
Hypertension, 0 = no, 1 = yes	Hypertension	0.805
Mother: age of diagnosis of other cardiovascular disease	The degree of severity of ischemic heart disease and heart failure	0.695
age in 2011	Age	0.616
Asthma diagnosed by doctor, I1, 142, 1 = no, 2 = yes	Asthma	0.705
Diabetes, type 2	Diabetes	0.833
Creatinine value upon hospital admission	Creatinine	0.750
Gender, 1 = female, 2 = male	Gender	0.738
Other	Other diseases of the central nervous system	0.705
height	Height [cm]	0.763
Creatinine; µmol/l	Creatinine	0.852
Age	Age	1.000
Age 1986	Age	0.792
Age at baseline	Age	0.733
Weight	Weight [kg]	0.848

Terminologies from TAUH	Terminologies from LODZ	Score
Asthma	Asthma	1.000
Asthma; mother, I1, 213	Asthma	0.754
Creatinine; mg/dl	Creatinine	0.863
Carotid diameter in the end of systole, average measure; mm	The degree of severity of ischemic heart disease and heart failure	0.673
age in 1980	Age	0.616
It's hard for me to reveal my innermost feelings even to close friends	The degree of severity of ischemic heart disease and heart failure	0.680
Diabetes treated with insulin, 0 = no, 1 = yes	Diabetes	0.725
Diabetes diagnosed by doctor, I1, 134, 1 = no, 2 = yes	Diabetes	0.716
The pleasure disappears from movies or plays if you try to find deep meanings from them	The degree of severity of ischemic heart disease and heart failure	0.705
Hypotensive or hemodynamically unstable	Chronic obstructive pulmonary disease	0.682
weight	Weight [kg]	0.763
Age when first trying smoking, I2, 647, 99 = has never smoked	Age	0.683
Diabetes mellitus, tyype 2	Diabetes	0.769
Diabetes type 1	Diabetes	0.844
RVP A; signal detection measure of sensitivity to the target	The degree of severity of ischemic heart disease and heart failure	0.637
other	Other diseases of the central nervous system	0.630
Age when ovariectomy for both ovary?	Age	0.694
Education level of parents in 1983	Education	0.755
Diabetes	Diabetes	1.000
Creatinine upn arrival or mean value from hospitalization or before hospitalization	Creatinine	0.707
Atrial fibrillation, 0 = no, 1 = yes	Atrial fibrillation	0.843
Indication for the test, evaluation after myocardial infraction	Vascular diseases of the central nervous system	0.634
Diabetes mellitus, tyype 1	Diabetes	0.769
TIMI classification in the first PCI target vessel after the procedure	The degree of severity of ischemic heart disease and heart failure	0.696

According to Table 48, the proposed approach was able to identify 139 potential lexical matches between TAUH and NESDA with more than 75% coherence (overlap). More

specifically, the lexical matcher was able to identify 4 terminologies with exact similarity, 13 terminologies with matching score larger than (or equal to) 0.9, as well as, 49 terminologies with matching score larger than (or equal to) 0.8. The 90 terminologies with scores in the range [0.6, 0.8) are considered as partially matched whereas the remaining 49 terminologies with scores in the range [0.8,1) are considered as highly similar.

Table 48. A summary of the potentially matched terminologies between TAUH and NESDA using the proposed approach.

Terminologies from TAUH	**Terminologies from NESDA**	**Score**
Other chronic disease	Coronary heart disease	0.676
Systolic blood pressure average, F6.2, 970	Systolic blood pressure	0.849
Metabolic syndrome according IDF criterion , 0 = no, 1 = yes ,	Metabolic Syndrome	0.740
Weight of father in 1986; kg	Weight	0.738
Carotid intima-media thickness, 3rd measure point; mm	Intima-media thickness	0.688
Systolic blood pressure 1st measurement	Systolic blood pressure	0.863
Waist circumference; cm	Waist circumference	0.942
Metabolic syndrome according to the 2001 NCEP definition	Metabolic Syndrome	0.749
Weight of mother in 1980; kg	Weight	0.738
Carotid intima-media thickness, 2nd measure point; mm	Intima-media thickness	0.688
The metabolic syndrome by the Harmonizing definition, 0 = ei ole, 1 = on	Metabolic Syndrome	0.620
Heart rate	Heart rate	1.000
Diastolic blood pressure average when sitting, from the status form; mmHg	Diastolic blood pressure	0.776
Height of father; cm. Average from values collected between 1980 and 1989.	Height	0.694
Sleep time in hours/day in 1986	sleep	0.643
smoking score, 0 = currently smoking, 1 = never smoked or has quit or is on a break	smoking	0.695
Systolic Blood Pressure	Systolic blood pressure	0.942
What other cardiovascular disease	Cardiovascular disease	0.721
Index of physical activity 1986.	physical activity	0.756

Terminologies from TAUH	Terminologies from NESDA	Score
Metabolic syndrome according to the 2005 NCEP definition. , 0 = no, 1 = yes ,	Metabolic Syndrome	0.722
smoking score, 0 = adult currently smoking /, child smoked at least one cigarette, 1 = adult never smoked or has quit or is on a break / , child never smoked	smoking	0.682
Smoking at age 18-24. Has the participant smoked daily at least in some point of his/her youth age	smoking	0.639
Systolic blood pressure, 3rd measurement	Systolic blood pressure	0.858
Weight of father in 1983; kg	Weight	0.738
Diabetes type 2	Diabetes Type 2	0.956
Diabetes, type 1	Diabetes Type 2	0.893
Father: coronary heart disease, 1 = no, 2 = yes	Coronary heart disease	0.658
Metabolic syndrome by harmonizing definition, 0 = no, 1 = yes	Metabolic Syndrome	0.741
Metabolic syndrome according to year 2001 NCEP criterion., 0 = no, 1 = yes	Metabolic Syndrome	0.725
Heart rate on arrival, GRACEdataset patient only	Heart rate	0.736
Education years of parents, Education years have been created by using information of years 1980, 1983 and 1986 measurements. Variable vkoulv is the maximum of parent's school years from those years.	education	0.643
Smoking during the past week or during the week before attempting to quit, I3, 803,	smoking	0.643
Father: coronary heart disease, I1,, 1 = no, 2 = yes	Coronary heart disease	0.643
Hypertension - Year when first time diagnosed	Hypertension	0.756
Systolic blood pressure 3rd measurement, F3.0, 776	Systolic blood pressure	0.820
Metabolic syndrome according to the IDF definition	Metabolic Syndrome	0.761
Index of physical activity 2011.	physical activity	0.756
Heart rate/min 3rd measurement	Heart rate	0.778
Height	Height	1.000
Weight of father in 1989; kg	Weight	0.738
triglyceride	Triglyceride levels	0.832
Sex, I2, 686, 1 = girl, 2 = boy	Sex	0.699
Diastolic blood pressure, 2nd measure; mmHg	Diastolic blood pressure	0.853
Cardiovascular diseases	Cardiovascular disease	0.986
Weight of mother in 1983; kg	Weight	0.738
Systolic blood pressure, 1st measure; mmHg	Systolic blood pressure	0.849
Waist-hip ratio waist07/hip07	Waist-hip ratio	0.839

Terminologies from TAUH	Terminologies from NESDA	Score
Systolic blood pressure 3rd measurement	Systolic blood pressure	0.863
Systolic blood pressure average, F6.2, 779	Systolic blood pressure	0.849
Heart rate/min 2nd measurement	Heart rate	0.778
Systolic blood pressure, 1st measurement	Systolic blood pressure	0.858
Alcohol	alcohol intake	0.762
Metabolic syndrome according to the IDF definition., 0 = no, 1 = yes ,	Metabolic Syndrome	0.729
Actual systolic blood pressure, 3rd measurement	Systolic blood pressure	0.672
Heart rate/min 1st measurement	Heart rate	0.778
Weight of the participant; kg	Weight	0.736
Systolic blood pressure 2nd measurement	Systolic blood pressure	0.863
Metabolic syndrome according to year 2005 NCEP criterion. , 0 = no, 1 = yes ,	Metabolic Syndrome	0.722
Systolic blood pressure 1st measurement, F3.0, 961	Systolic blood pressure	0.820
smoking score, 0 = current smoking, 1 = never smoked or has quit or is on a break	smoking	0.695
Mother: coronary heart disease, 1 = no, 2 = yes	Coronary heart disease	0.642
Hypertension/Hypertonia	Hypertension	0.841
Smoking at age 12-18. Has the participant smoked daily at least in some point of his/her youth age	smoking	0.639
Index of physical activity 1980.	physical activity	0.756
Index of physical activity 1989.	physical activity	0.756
Diastolic blood pressure, 1st measure; mmHg	Diastolic blood pressure	0.853
Triglyceride	Triglyceride levels	0.877
Actual systolic blood pressure, 2nd measurement	Systolic blood pressure	0.672
Mother: coronary heart disease, I1,, 1 = no, 2 = yes	Coronary heart disease	0.627
Hypertension, 0 = no, 1 = yes	Hypertension	0.805
Systolic blood pressure upon arrival	Systolic blood pressure	0.880
Index of physical activity 1983.	physical activity	0.756
Major coronary heart disease event excluding revascularizations	Coronary heart disease	0.620
I am fairly self-confident	Anti-inflammatory medication use	0.626
Heart rate mean value, from the status form	Heart rate	0.744
Weight at birth; g	Weight	0.778
Metabolic syndrome according to the EGIR definition., 0 = no, 1 = yes ,	Metabolic Syndrome	0.728
Diabetes, type 2	Diabetes Type 2	0.936

Terminologies from TAUH	Terminologies from NESDA	Score
Father: other cardiovascular disease	Cardiovascular disease	0.611
Height of the participant; cm	Height	0.736
Sex: 1 = male, 2 = female	Sex	0.707
Metabolic syndrome according to EGIR criterion, 0 = no, 1 = yes ,	Metabolic Syndrome	0.735
Weight of mother in 1989; kg	Weight	0.738
Systolic blood pressure, 2nd measure; mmHg	Systolic blood pressure	0.849
Heart rate upon arrival	Heart rate	0.812
Cardiovascular disease, including CHD, STR, SAH & TIA	Cardiovascular disease	0.805
Other	Other inflammatory markers	0.731
height	Height	0.889
Coronary heart disease, 0=no, 1=yes, FC3 needs confirmation!	Coronary heart disease	0.789
Metabolic syndrome according to the 2001 NCEP definition., 0 = no, 1 = yes	Metabolic Syndrome	0.725
Weight	Weight	1.000
Systolic blood pressure average, from the status form; mmHg	Systolic blood pressure	0.797
Waist-hip ratio waist11/hip11	Waist-hip ratio	0.839
Glucose	Glucose levels	0.833
Systolic blood pressure 2nd measurement, F3.0, 773	Systolic blood pressure	0.820
Total cholestrerol	Total cholesterol levels	0.865
Heart Rate	Heart rate	0.933
Index of physical activity. The higher the value, the more active the participant is.	physical activity	0.645
Total cholesterol	Total cholesterol levels	0.903
Diastolic blood pressure on entry, GRACEdataset patient only	Diastolic blood pressure	0.800
Systolic blood pressure 1st measurement, F3.0, 770	Systolic blood pressure	0.820
Heart rate on arrival, only for MI dataset patient	Heart rate	0.733
Systolic blood pressure average Circmon, from the status form; mmHg	Systolic blood pressure	0.781
Index of physical activity 1992.	physical activity	0.756
Metabolic syndrome according to the 2005 NCEP definition	Metabolic Syndrome	0.749
Waist circumference	Waist circumference	1.000
Major coronary heart disease event	Coronary heart disease	0.715

Terminologies from TAUH	Terminologies from NESDA	Score
weight	Weight	0.889
Waist circumference average measurement; cm	Waist circumference	0.814
Coronary artery induced chest pain	Other psychotropic medication use	0.658
Sex, I2, 863	Sex	0.750
Diabetes type 1	Diabetes Type 2	0.911
Diastolic blood pressure average Circmon, from the status form; mmHg	Diastolic blood pressure	0.784
Index of physical activity 2001.	physical activity	0.756
other	Other inflammatory markers	0.651
Insulin	Insulin levels	0.833
Weight of mother in 1986; kg	Weight	0.738
Heart rate mean value when sitting, from the status form; Heart rate/min	Heart rate	0.713
Systolic blood pressure average when sitting, from the status form; mmHg	Systolic blood pressure	0.773
Index of physical activity 2007.	physical activity	0.756
Actual systolic blood pressure, 1st measurement	Systolic blood pressure	0.672
Carotid intima-media thickness, 4th measure point; mm	Intima-media thickness	0.688
Education level of parents in 1983	education	0.708
Carotid intima-media thickness, 1st measure point; mm	Intima-media thickness	0.688
Diabetes	Diabetes Type 2	0.844
recorded baseline mortality	Metabolic Syndrome	0.611
Systolic blood pressure, 2nd measurement	Systolic blood pressure	0.858
Metabolic syndrome according to harmonized definition , 0 = no, 1 = yes	Metabolic Syndrome	0.728
Metabolic syndrome according to the EGIR definition	Metabolic Syndrome	0.759
Systolic blood pressure 2nd measurement, F3.0, 964	Systolic blood pressure	0.820
Systolic blood pressure 3rd measurement, F3.0, 967	Systolic blood pressure	0.820
Systolic blood pressure on entry, GRACEdataset patient only	Systolic blood pressure	0.797
Smoking during the past week or during the week before attempting to quit, I3, 666,	smoking	0.643
Smoking, 0=no, 1=yes. This variable should be used for contemporary smoking.	smoking	0.645

Terminologies from TAUH	Terminologies from NESDA	Score
How often the participant consumes mild-strength beer	Use of cholesterol-lowering medication	0.619
Diastolic Blood Pressure	Diastolic blood pressure	0.944
Height at birth; cm,	Height	0.767
Weight of father in 1980; kg	Weight	0.738
Height of mother; cm. Average from values collected between 1980 and 1989.	Height	0.694

According to Supplementary Table 3, the proposed approach was able to identify 288 potential lexical matches between TAUH and UVA with more than 75% coherence (overlap). More specifically, the lexical matcher was able to identify 6 terminologies with exact similarity, 25 terminologies with matching score larger than (or equal to) 0.9, as well as, 66 terminologies with matching score larger than (or equal to) 0.8. The 222 terminologies with scores in the range [0.6, 0.8) are considered as partially matched whereas the remaining 66 terminologies with scores in the range [0.8,1) are considered as highly similar. The frequencies of the uniquely matched terminologies from TAUH in LODZ, as well as, from TAUH in NESDA and from TAUH in UVA are summarized Table 49, Table 50, and Table 51, respectively. Recall that since the number of occurrences was low in the pairs of datasets regarding the mental disorders' domain, cross domain matches were applied based on TAUH and the number of occurrences are described in the current section. According to Table 49, the terminology in TAUH having the highest frequency in LODZ is related to "Age" with 24 occurrences. The terminologies "Diabetes", "The degree of severity of ischemic heart disease and heart failure", come next with 13 and 11 occurrences, respectively. Terminologies with a relatively high number of occurrences include the "Creatinine", "Asthma", and "Chronic obstructive pulmonary disease", among others. The rest of the terminologies include the "Other diseases of the central nervous system", "Atrial fibrillation", "Weight [kg]" and "Hypertension" with a frequency 2, as well as, the terminologies "Gender", and "Vascular diseases of the central nervous system", with a frequency 1.

Table 49. The frequencies of the uniquely matched terminologies from LODZ in TAUH.

Matched terminology from LODZ	Frequency (number of occurrences) in TAUH
Age	24

Matched terminology from LODZ	Frequency (number of occurrences) in TAUH
Diabetes	13
Chronic obstructive pulmonary disease	5
Other diseases of the central nervous system	3
Asthma	5
Education	2
Creatinine	6
Hypertension	3
Height [cm]	2
The degree of severity of ischemic heart disease and heart failure	11
Atrial fibrillation	2
Lives	1
Gender	1
Weight [kg]	2
Vascular diseases of the central nervous system	1

According to Table 50, the terminology in TAUH having the highest frequency in NESDA is related to “Systolic blood pressure” with 25 occurrences. The terminologies “Metabolic Syndrome”, “Weight”, and “Heart rate” come next with 16, 12 and 10 occurrences, respectively. Terminologies with a relatively high number of occurrences include the “physical activity”, “Coronary heart disease”, “smoking”, “Height”, and “Diastolic blood pressure” among others. The rest of the terminologies include the “Intima-media thickness”, “Waist-hip ratio”, “Total cholesterol levels”, and “Insulin levels”, either with frequency 2 or 1, among others.

Table 50. The frequencies of the uniquely matched terminologies from NESDA in TAUH.

Matched terminology from NESDA	Frequency (number of occurrences) in TAUH
Coronary heart disease	8
Systolic blood pressure	25
Metabolic Syndrome	16
Weight	12
Intima-media thickness	4
Waist circumference	3
Heart rate	10
Diastolic blood pressure	6
Height	6

Matched terminology from NESDA	Frequency (number of occurrences) in TAUH
sleep	1
smoking	8
Cardiovascular disease	4
physical activity	9
Diabetes Type 2	5
education	2
Hypertension	3
Triglyceride levels	2
Sex	3
Waist-hip ratio	2
alcohol intake	1
Anti-inflammatory medication use	1
Other inflammatory markers	2
Glucose levels	1
Total cholesterol levels	2
Other psychotropic medication use	1
Insulin levels	1
Use of cholesterol-lowering medication	1

According to Table 51, the terminology in TAUH having the highest frequency in UVA is related to “Total cholesterol” with 38 occurrences, as well as, to “triglycerides” with 35 occurrences. The terminologies “Age”, “Lipoprotein”, and “Weight” come next with 24, 13 and 12 occurrences, respectively. Terminologies with a relatively high number of occurrences include the “physical activity”, “HDL cholesterol”, “glucose”, “height”, “C-reactive protein”, “LDL cholesterol”, and “Total cholesterol in mg/dl”, among others. The rest of the terminologies include the “Depression found”, “medication”, “diabetes type”, “Type I diabetes”, “hematocrit”, “Myocardial infarction, first time”, and “leukocytes”, with a smaller frequency, among others.

Table 51. The frequencies of the uniquely matched terminologies from UVA in TAUH.

Matched terminology from UVA	Frequency (number of occurrences) in TAUH
chronic disease	1
Pain disappear, remain when standing or walking slowly	3
HDL cholesterol	7
triglycerides	35

Matched terminology from UVA	**Frequency (number of occurrences) in TAUH**
Depression found	1
medication	3
Mean cellular hemoglobin concentration of erythrocytes	2
Total cholesterol	38
heart failure	1
Weight	12
Lipoprotein	13
Age	24
C-reactive protein	5
marital status	7
date	8
glucose	7
When complaints for the first time occurred in the life	10
almost never, things develop according to my ideas	4
size	12
myocardial infarction	1
heart rate	10
height	6
Target systolic blood pressure	1
physical activity	9
I rarely count that happens to me something good	2
Total cholesterol in mg/dl	4
leukocytes	1
Myocardial infarction, first time	1
Long nitrates	1
diabetes type	2
when in a hurry or during physical exertion	2
Oral glucose tolerance test	1
among my friends I feel comfortable	3
hyperlipidemia	1
Difficulty concentrating / decision problems	3
LDL cholesterol	4
with acute coronary syndrome	1
Alcohol consumption in last 12 months	1
before how many months last determined by the doctor	3
atrial fibrillation	1
Total cholesterol in mmol/l	7

Matched terminology from UVA	Frequency (number of occurrences) in TAUH
erythrocytes	1
Date of birth	2
Myocardial infarction, Number	1
Type I diabetes	1
current setting and compliance	3
testosterone	1
more diagnoses	1
master stressful situations	1
more alcohol drinking justifiable as health	1
gender	1
Other	1
homocysteine	1
Number of almost daily Drinks	1
Mean cellular hemoglobin content of erythrocytes	1
Diagnostic status hypertension	1
against elevated blood lipids	1
other	1
insulin	1
uric acid	1
hematocrit	1
diabetes	1
Number of days	1
metabolic syndrome	1
Carotid stenosis	1
Target diastolic blood pressure	1
morning	1
Fasting blood sugar	1

Our method was compared against the application of conventional lexical analysis and manual mapping across 998 terminologies from the mental disorders' domain, yielding a set of individual and cross domain matched terminologies with 85% precision (in average) and 10% higher performance than conventional lexical analysis. The proposed method yielded an increased number of cross-domain matched terminologies in less execution time and with higher overlap than conventional lexical analysis and manual mapping of terminologies which is extremely time consuming. The computational complexity of the method was proportional to the number of input terminologies. In

some cases, the number of matched terminologies was high (e.g., 435 in LURIC) enough, whereas in other cases, like in the mental disorders', the number was adequate. Considering the small number of matched terminologies in the mental disorders' domain and the fact that most of these terminologies were CVD-oriented, such as, "cholesterol", and "hypertension", cross-domain matches were investigated for the first time in the literature.

7.5. Systemic autoinflammatory diseases

This case study involves the application of the beyond the state of the art methods that were developed for data curation (CHAPTER 3) and a local training and testing scenario under federated learning (CHAPTER 6) to address open issues and clinical unmet needs (Section 1.4) in the domain of the systemic autoinflammatory diseases (Section 2.3.3) and particularly in patients who have been diagnosed with Kawasaki disease (KD).

7.5.1. Data curation

7.5.1.1. Case Study 1 – Curation across 8 open-source datasets from GEO

The scope of this case study is to enhance the quality of 8 gene expression datasets from GEO in SAIDs. To this end, microarray data were collected from the Gene Expression Omnibus (GEO) public functional genomics data repository [405] for: (i) common platform analysis, where diagnostic biomarkers for KD are extracted from time-series gene expression data across three different KD phases followed by a validation of the extracted biomarkers against the known ones in the literature, and (ii) cross-platform analysis, where the proposed diagnostic biomarkers are further compared against the known KD genes through the integration of six more datasets. The genetic samples were tested for joint variabilities by calculating the covariance matrix and discarding genes with significantly high covariance. Any missing genetic samples were replaced with zero. Any incompatible fields and outliers were removed from the computational workflow to prior to the imputation process to avoid data contamination yielding high-quality genetic data.

Due to the variation of the range of values across the microarray data which were obtained from the two datasets in the GPL6271 (Table 52), as well as, from the six

datasets across the GPL570 and GPL10558 platforms (Table 53), a meta-analysis procedure was performed on each individual dataset based on the quantile normalization approach [406]. Specifically, the average of each quantile across the proposed KD genes was used as the reference to transform (adjust) their distributions. The same process was applied on the known KD genes. Since one gene might have more than one probes, the median of the probes was extracted per gene, prior to the quantile normalization process.

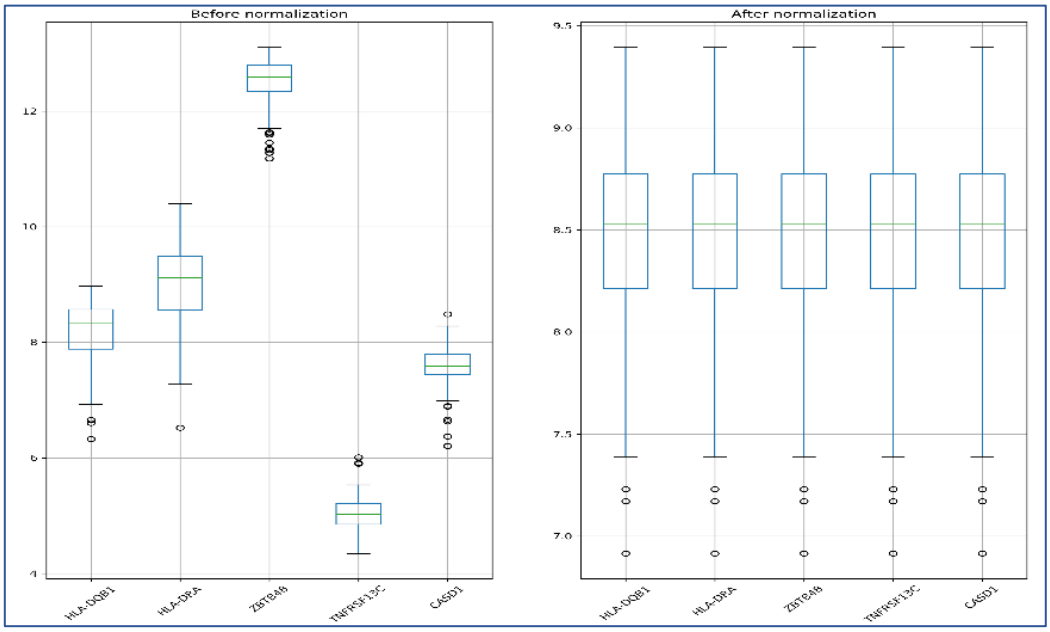


Figure 79. Quantile normalization is used to align the distributions of the proposed and known genes for KD diagnosis.

Table 52. A summary of the datasets which participated in the common platform analysis.

Platform	Dataset	Disorder	Values	Patient samples
GPL6271	GSE9863 [407]	Kawasaki	Log2 median ratio	20 KD (at three phases)
GPL6271	GSE47683 [408]	Renal transplantation	Normalized log ratio	67 Non-KD (8 healthy subjects)

Table 53. A summary of the datasets which participated in the cross-platform analysis.

Platform	Dataset	Disease	Values	Patient samples
GPL570	GSE80060 [409]	SJIA	Linear scale RMA normalized relative expression values	206 Non-KD (22 healthy)
	GSE61635	SLE	RMA signal intensity in log2 scale	129 Non-KD (30 healthy)

Platform	Dataset	Disease	Values	Patient samples
GPL10558	GSE73461 [410]	KD, other inflammatory, bacterial/viral infections	Illumina calculated signal intensity	78 KD, 381 Non-KD (55 healthy)
	GSE63881 [411]	KD	Z-score normalization	171 KD (10 healthy)
	GSE68004 [412]	KD, HAdV, GAS	Average normalization	76 KD, 73 Non-KD (37 heathy)
	GSE73463 [410]	KD	Illumina calculated signal intensity	233 KD

All in all, the quality of the Gene expression microarray data was significantly enhanced through the quantile standardization process.

7.5.2. Federated/distributed learning (local case)

7.5.2.1. Case Study 1 – A new set of biomarkers for Kawasaki disease

The scope of this case study is to identify a new set of genetic biomarkers for Kawasaki. To this end, we used the data presented in Section 7.5.1.1. The rectangular grid of the second stage (intra-phase) SOM is depicted in the left-hand side of Figure 80. The SOM consists of five clusters (prototypes), where, cluster 1 consists of six patients (KD3004, KD3014, KD3033, KD3037, KD3047, KD3054), cluster 3 consists of five patients (KD1502, KD1505, KD3016, KD3019, KD3038), cluster 7 consists of two patients (KD3027, KD3028), cluster 8 consists of one patient (KD3049), and cluster 9 consists of six patients (KD1506, KD3007, KD3046, KD3058, KD3059, KD3064).

Figure 80. An illustration of the second stage SOM along with the detected super-clusters.

It is interesting to note that the clusters 2, 4, 5, and 6 of the 3x3 SOM were empty since no samples were projected in those grid cells. To merge prototypes with similar patterns, the five clusters were aggregated into super-clusters by applying hierarchical clustering on the Euclidean distances between them yielding the four prototypes (super-clusters) which are depicted in the right-hand side of Figure 80. More specifically, the dendrogram which was generated by hierarchical clustering was partitioned into four super-clusters, where, super-cluster 1 consists of the patients in cluster 1, super-cluster 2 consists of the six patients in cluster 3, super-cluster 3 consists of the two patients in cluster 7 and super-cluster 4 is the union of clusters 8 and 9. The labels of the super-clusters were used subsequently to identify the proposed genes.

The FDR-based feature selection schema was able to identify the following gene reference IDs as significant ($p < 0.01$, Benjamini-Hochberg adjusted) across all three phases: 15658, 15660, 22055, 26049, and 35359. The proposed genes for KD diagnosis are presented in Table 54. It should be noted that in order to map the gene IDs to the available gene probes and since the employed KD dataset does not provide any information on the utilized genes (ID, name or description) we performed a BLAST (Basic Local Alignment Search Tool) search on GenBank [42], to detect the most homolog sequence and subsequently the corresponding gene. The gene CASD1 achieved the highest score in phase A (F-score = 19.93), the gene TNFRSF13C in phase SA (F-score = 15.74), and the gene CASD1 again for phase C (F-score = 12.12).

Table 54. The proposed set of genes for KD diagnosis.

ID_REF	Gene ID	ANOVA F-scores (with p<0.01, BH-adjusted)		
		A	SA	C
15658	HLA-DQB1	9.79	10.05	10.51
15660	HLA-DRA	11.12	9.95	9.03
22055	ZBTB48	13.74	11.25	11
26049	TNFRSF13C	12.22	15.74	11.59
35359	CASD1	19.93	9.37	12.12

The known KD genes from the literature are presented in Table 55 along with the corresponding ID_REF and a short description. Based on the work in [408], [413], [414] and the associated genes which are listed, we detected those that are also listed in [415] which, as already mentioned, uses the same experimental platform with the employed KD dataset but also provides the corresponding gene IDs. Probes with IDs

253, 29567 belong to the TLR6 (Toll-like receptor 6) family which is related with pathogen recognition and activation of innate immunity. Probes with IDs 9368, 34805 correspond to the COPB2 gene (COPI Coat Complex Subunit Beta 2) family which is part of the Golgi coatomer complex [414] that constitutes the coat of nonclathrin-coated vesicles and is essential for Golgi budding and vesicular trafficking. Probe ID 12792 corresponds to the FCGR2A (Fc Fragment Of IgG Receptor IIa) which belongs to the family of immunoglobulin Fc receptor genes that exist on the surface of many immune response cells. The probe with ID 26786 is the CD40 molecule which is essential for mediating a broad variety of immune and inflammatory responses [414]. Probe IDs 33880, 37136 belong to the BLK Proto-Oncogene family whose protein is involved in B-cell receptor signaling and development and finally the gene with ID 34697 is the Caspase 3 (CASP3) which is highly involved in the execution-phase of cell apoptosis [414].

Table 55. Known genes for KD diagnosis.

ID_REF	Gene ID	Description
253	TLR6	Toll-like receptor 6 as plays a fundamental role in pathogen recognition and activation of innate immunity.
29567		
9368	COPB2	COPI Coat Complex Subunit Beta 2 constitutes the coat of nonclathrin-coated vesicles and is essential for Golgi budding and vesicular trafficking.
34805		
12792	FCGR2A	Fc Fragment Of IgG Receptor IIa encodes a family member of immunoglobulin Fc receptor genes found on the surface of many immune response cells.
26186	CD40	The CD40 molecule belongs to the TNF-receptor superfamily and is a receptor on antigen-presenting cells of the immune system which is essential for mediating a broad variety of immune and inflammatory responses.
33880	BLK	BLK Proto-Oncogene is a protein which has a functional role in B-cell receptor signaling and B-cell development.
37136		
34697	CASP3	Caspase 3 is a gene whose encoded protein is a cysteine-aspartic acid protease that plays a central role in the execution-phase of cell apoptosis.

Each gene expression dataset from Table 52 and Table 53 was adjusted based on the quantile normalization process. No outliers or genes with joint variability were detected in the two datasets. The performance evaluation results of the XGBoost on both the proposed and the known genes are presented in Table 56. The procedure was repeated

using the AdaBoost algorithm as a second boosting classifier to further compare the classification outcomes among the two cases (Table 56). A repeated stratified 10-fold cross validation procedure was applied for the performance evaluation of both boosting schemas, where four measures were averaged across the folds, namely, the accuracy, sensitivity, specificity, and AUC. Through the stratified strategy, the number of KD patients is the same across each fold. The corresponding ROC curves of the XGBoost and the AdaBoost are depicted in Figure 81 for phases A, SA, and C and for each training case (case 1: on the dataset with the proposed genes and case 2: on the dataset with the known KD genes). In both boosting schemas, the proposed set of genes yielded a notable performance on the Acute and Subacute phases which is reflected by the high-performance evaluation results in Table 56.

Table 56. Performance evaluation results for the XGBoost and the AdaBoost across the three phases for both the known and the proposed set of genes.

XGBoost												
Set of genes	**Accuracy**			**Sensitivity**			**Specificity**			**AUC**		
	A	**SA**	**C**	**A**	**SA**	**C**	**A**	**SA**	**C**	**A**	**SA**	**C**
Known	0.956	0.989	0.989	0.918	0.975	0.975	0.986	1	1	0.981	0.988	0.995
Proposed	1	1	0.978	1	1	0.986	1	1	0.971	0.995	0.995	0.995
AdaBoost												
Set of genes	**Accuracy**			**Sensitivity**			**Specificity**			**AUC**		
	A	**SA**	**C**	**A**	**SA**	**C**	**A**	**SA**	**C**	**A**	**SA**	**C**
Known	0.944	0.911	0.954	0.929	0.925	0.970	0.957	0.9	0.940	0.950	0.947	0.967
Proposed	1	0.976	0.989	1	0.968	0.993	1	0.986	0.986	0.995	0.995	0.995

Regarding the XGBoost algorithm (Table 56), the classification outcomes using the known set of genes yielded accuracy 0.956 for phase A, 0.989 for phase SA, and 0.989 for phase C, and the AUC scores were 0.981, 0.988, and 0.995, respectively (Figure 81). On the other hand, the performance of the XGBoost on the proposed set of genes was higher in phases A and SA, yielding accuracy 1.0 for phase A and SA, and 0.978 for phase C, where the AUC scores were 0.995 across all phases (with a standard deviation $\pm$ 0.1). Although in phase C the sensitivity of the XGBoost on the proposed set of genes was 1.1% higher than the one on the known set of genes, the specificity was smaller thus yielding a slightly reduced performance.

As far as the AdaBoost algorithm is concerned, the increased performance of the proposed set of genes against the known ones is preserved, however, with an increased

performance across all three phases. According to Table 56, the known set of genes yielded accuracy 0.944 for phase A, 0.911 for phase SA, and 0.954 for phase C, where the AUC scores were 0.950, 0.947, and 0.967, respectively. On the other hand, the performance of the AdaBoost algorithm on the proposed set of genes was higher in all phases, yielding accuracy 1.0 for phase A, 0.976 for phase SA, and 0.989 for phase C. The sensitivity values were 1, 0.968, 0.993 and the specificity values were 1, 0.986, and 0.986, respectively, yielding increased AUC scores across all phases.

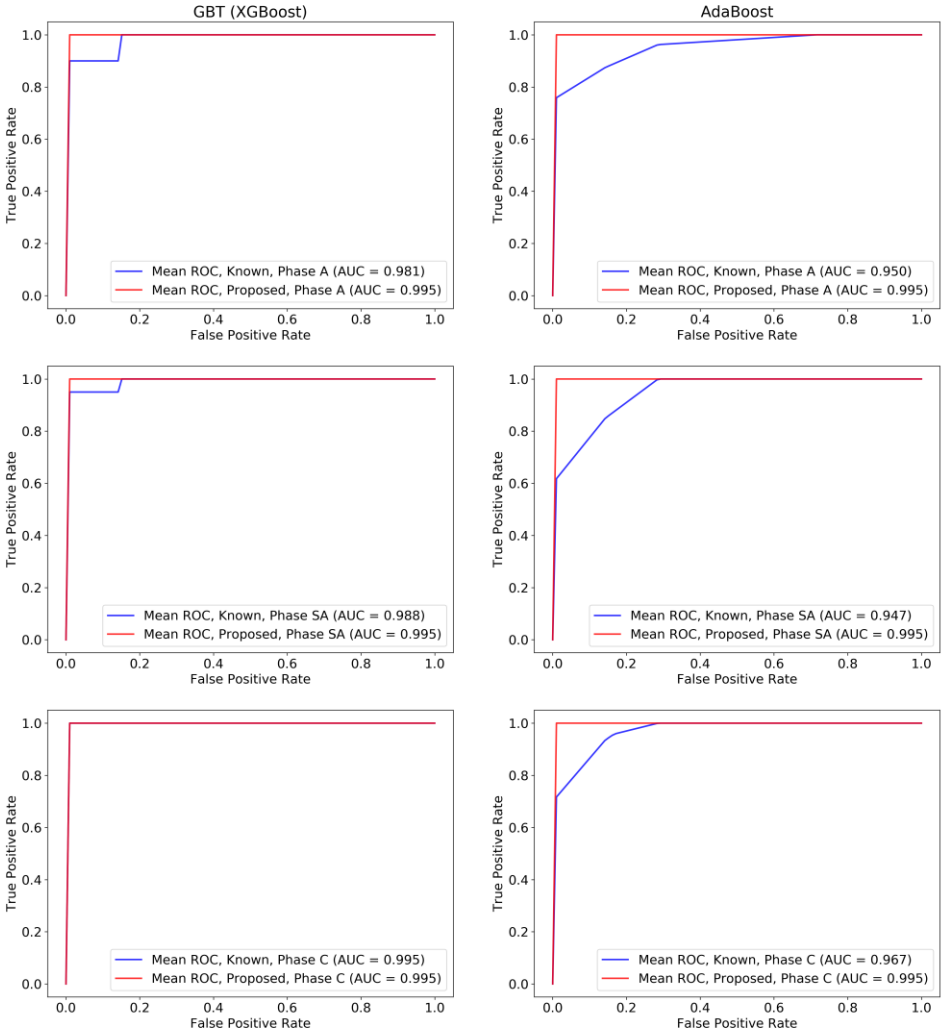


Figure 81. A comparison of the Receiver Operating Characteristic (ROC) curves (the true positive rate against the false positive rate) between the GBT (XGBoost) algorithm which was trained on the dataset with the proposed genes (red line) and the known KD genes (blue line), for phases A, SA, and C.

In total, the classifiers yielded an average increase by 4.40% in the accuracy, 5.52% in sensitivity, and 3.57% in specificity compared with the known set of genes in phases A

and SA. The contribution of the proposed set of genes appears to be significantly higher in phase A and SA, a fact which is also present in the AdaBoost schema. This implies that the high tendency of the proposed genes against the known genes is preserved in these two phases apart from the boosting schema. Regarding phase C, the high performance is maintained in the AdaBoost whereas in the GBT the reduced specificity results in a slightly smaller performance. Each gene expression dataset (GSE80060, GSE61635, GSE73461, GSE63881, GSE68004, GSE73463) from Table 53 was individually transformed (adjusted) using the quantile normalization process. No outliers or genes with joint variability were detected. The median of the probes was extracted in the case of genes with more than one probes. The transformed data were then integrated into two different data structures which included the proposed biomarkers and the known diagnostic biomarkers, respectively.

The non-KD patients (including patients who have been diagnosed with SLE, SJIA or other inflammatory diseases, bacterial or viral infections, HAdV and GAS) were annotated with a value 0 whereas the KD patients were annotated with a value 1 to solve a binary classification problem using the XGBoost and the AdaBoost classifiers. Regarding the XGBoost algorithm (Figure 81), the classification outcomes using the known set of genes yielded accuracy 0.847, sensitivity 0.845, specificity 0.894, and AUC 0.906, respectively. On the other hand, the performance of the XGBoost algorithm on the proposed set of genes was higher (Figure 81), yielding accuracy 0.872, sensitivity 0.869, specificity 0.939, and AUC 0.927.

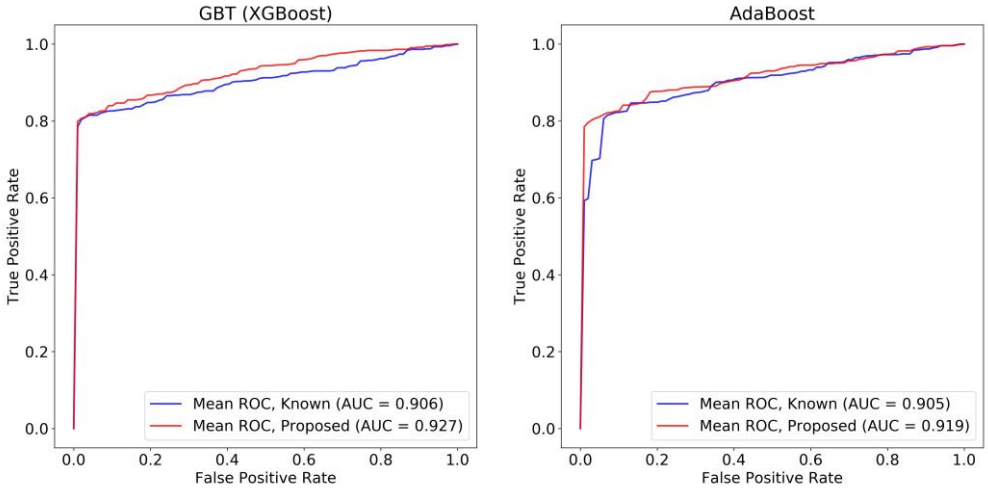


Figure 82. A comparison of the Receiver Operating Characteristic (ROC) curves (the true positive rate against the false positive rate) between the GBT (XGBoost) algorithm (on the left-

hand side) and the AdaBoost algorithm (on the right-hand side) which were trained on the proposed genes (red line) and the known KD genes (blue line) across the cross-platform data.

As for the AdaBoost algorithm (Figure 81), the increased performance of the proposed set of genes against the known ones is once more preserved, however, with a reduced performance than the XGBoost, like in the common platform analysis. The classification outcomes from the known set of genes yielded accuracy 0.848, sensitivity 0.846, specificity 0.892, and AUC 0.905. On the other hand, the performance of the AdaBoost algorithm on the proposed set of genes was higher (Figure 82), yielding accuracy 0.868, sensitivity 0.865, specificity 0.94, and AUC 0.919. In total, both classifiers yielded an average increase by 2.30% in the accuracy, 2.20% in sensitivity, 4.70% in specificity, and in 1.70% in AUC.

To address the need for KD diagnosis, we proposed a computational pipeline which clusters KD patients with similar gene expression profiles across the three different KD phases, namely, the Acute (A), Subacute (SA) and Convalescent (C), and uses the resulting clustermap to detect prominent genes as biomarkers for KD diagnosis. To do so, we construct Self-Organizing Maps (SOMs) to group patients with similar gene expressions into homogeneous clusters across the three phases. Then, we apply FDR-based feature selection to detect genes that significantly deviate across the clusters on each phase. As a last step, we extract the final set of proposed genes as those that are present across all phases and compare their performance against known KD genes in the literature by training two ML algorithms for KD classification.

According to the results, five prominent genes for KD diagnosis are proposed for the first time, namely the HLA-DQB1, HLA-DRA, ZBTB48, TNFRSF13C, and CASD1. These genes were used to develop a KD boosting classifier which yielded better performance against the one trained on the known KD genes in terms of increased accuracy, sensitivity, specificity, and AUC.

To our knowledge, this is the first ML-based computational workflow using intra-phase and inter-phase clustering for KD genomic data analysis towards the discovery of biomarkers for KD diagnosis. Further examination of the proposed genes in terms of functional analysis, as well as, clinical validation may unveil new insights concerning the pathogenesis of KD and the underlying genetic mechanisms.

7.6. COVID-19

This case study involves the application of the beyond the state-of-the-art methods that were developed for data curation (CHAPTER 3) and a local training and testing scenario under federated learning (CHAPTER 6) to address open issues and clinical unmet needs (Section 1.4) in the domain of COVID-19 (Section 2.3.4).

7.6.1. Data curation

7.6.1.1. Case Study 1 – Sotiria Hospital

The scope of this case study is to enhance the quality of a medium size clinical database across 3 timepoints with hospitalized COVID-19 patients in Greece. Anonymized patient data were collected from 324 hospitalized patients with average age 60.65 ($\pm$14.44) who were diagnosed with COVID-19 from the 21st Department of Pulmonary Medicine, National and Kapodistrian University of Athens, in "Sotiria" Hospital for the diseases of the chest, as described in [416]. According to Table 57, the data include demographic information, comorbidities, laboratory tests (e.g., C-reactive protein), therapies (corticosteroids and antiviral agents) as well as cytokines and interleukins measurements at four time intervals. Patient records having at least one missing value in the admission ICU date or in mortality were ignored from the analysis (110 patients).

Thus, the final population included 214 patients with average age 60.93 ($\pm$15.38). Patients were categorized into four groups based on their admission in the ICU and/or mortality, where Group A included those who survived without ICU admission (131 patients, average age 55.99 ($\pm$15.1)), Group B included patients who were not admitted to the ICU but died (4 patients, average age 81 ($\pm$6.52)), Group C included those who were admitted to the ICU and survived (43 patients, average age 63.79 ($\pm$11.62)), and Group D included patients who were admitted to the ICU and died (36 patients, average age 73.81 ($\pm$9.62)).

The initial dataset included 110 features with 324 instances. Out of 324 features, 36 were discrete, 57 were continuous and 17 features had unknown data type (i.e., mixed data types). The total number of missing values was 54.53%. After the end of the first stage of the data curation process (Stage I), the total number of features was 57 with 214 instances. Out of 57 features, 20 were discrete and 37 were continuous with a total

of 35.38% missing values (Stage II). In the final stage (Stage III), the k-NN approach was applied for data imputation only on features with an acceptable percentage of missing values (≤40%) to increase the completeness of the data before the application of the classification models. In addition, highly associated features with the target feature, such as, the days in the ICU and the hospitalization time were removed from the analysis.

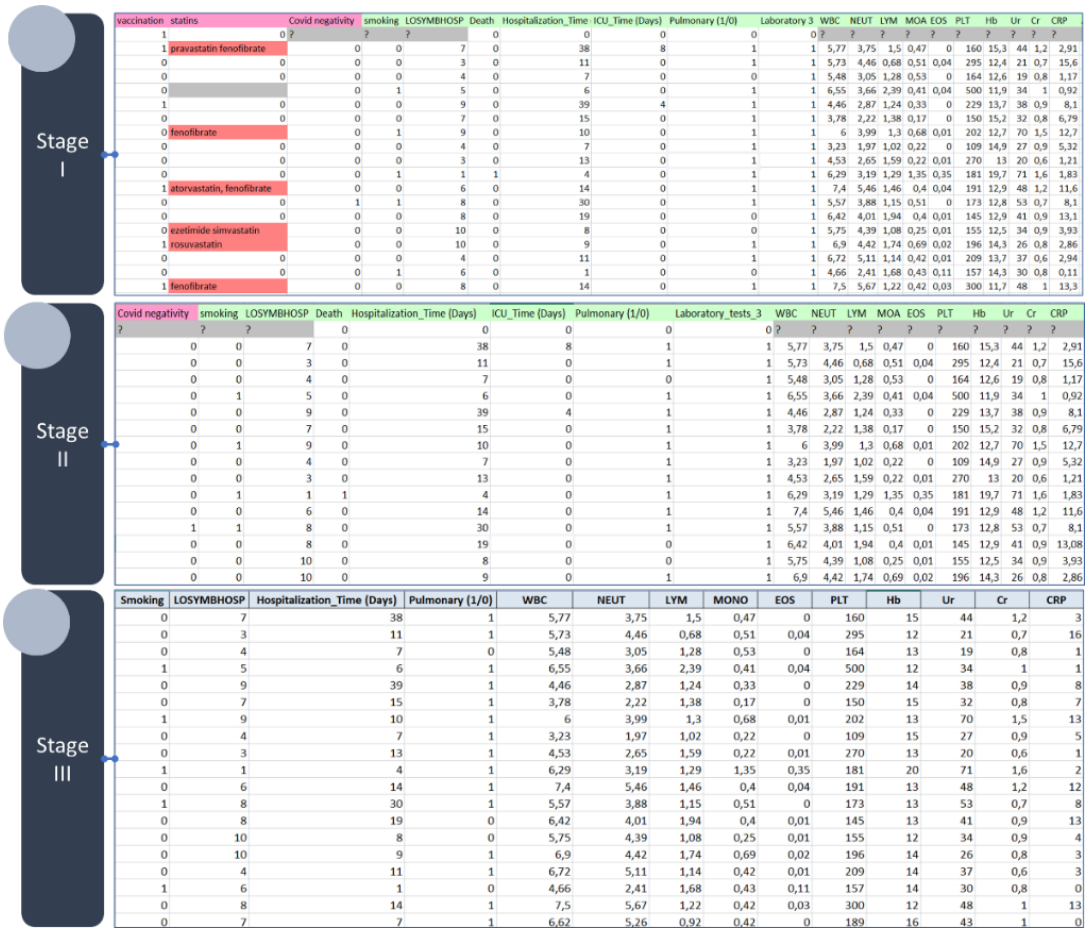


Figure 83. An indicative instance of the anonymized data before (on top) and after (on bottom) data curation. The acronyms of the features are described in Table 57.

Table 57. A summary of the features that participated in the analysis (after data curation).

Feature	Description	Value range
Age	-	[18, 91]
Gender	-	[0, 2]
Comorbidities (presence)	-	[0, 1]
Diabetes type I	-	[0, 1]
Diabetes type II	-	[0, 1]
Dyslipidemia	-	[0, 1]

Feature	Description	Value range
Hypertension	-	[0, 1]
Thyroidism	-	[0, 1]
COPD	Chronic obstructive pulmonary disease	[0, 1]
Atrial fibrillation	-	[0, 1]
Allergic rhinitis	-	[0, 1]
Asthma	-	[0, 1]
Others	Presence of any other comorbidities	[0, 1]
APACHE II	-	[0, 20]
Vaccination	-	[0, 1]
Smoking	-	[0, 2]
LOSYMBHOSP	Day interval from the first symptom until the admission to the hospital	[1, 29]
WBC	White blood cell count	[2.53, 20]
NEUT	Neutrophils	[0.97, 17.32]
LYM	Lymphocytes	[0.14, 3.09]
MONO	Monocytes	[0.04, 1.35]
EOS	Eosinophils	[0, 0.71]
PLT	Platelets	[52, 560]
Hb	Hemoglobin	[8, 73]
Cr	Creatinine	[0.5, 3.2]
CRP	C-reactive protein	[0.11, 29]
AST	Aspartate Aminotransferase	[15, 380]
ALT	Alanine Aminotransferase	[8, 223]
LDH	Lactate Dehydrogenase	[38, 1394]
Oxygen type	0: No oxygen, 1: Ventilator, 2: Oxygen (Mask, nasal Canula), 3: None of the above, 4: NonInvasive (CPAP, BIPAP), 5: High flow	[0, 5]
IL1b_days_0_2	Interleukin 1 beta in time interval 1 – INT1 (averaged across days 0 – 2)	[0.028, 1.82]
IL6_days_0_2	Interleukin 6 in time interval 1 – INT1 (averaged across days 0 – 2)	[0.137, 60.891]
IL8_days_0_2	Interleukin 8 in time interval 1 – INT1 (averaged across days 0 – 2)	[0.287, 90.426]
TNF_days_0_2 (or TNFa_days_0_2)	Tumor necrosis factor (alpha) in time interval 1 – INT1 (averaged across days 0 – 2)	[0.963, 17.23]
IL1b_days_3_5	Interleukin 1 beta in time interval 2 – INT2 (averaged across days 3 – 5)	[0.107, 1.58]
IL6_days_3_5	Interleukin 6 in time interval 2 – INT2 (averaged across days 3 – 5)	[0.06, 58.841]
IL8_days_3_5	Interleukin 8 in time interval 2 – INT2 (averaged across days 3 – 5)	[1.635, 111.816]
TNF_days_3_5 (or TNFa_days_3_5)	Tumor necrosis factor (alpha) in time interval 2 – INT2 (averaged across days 3 – 5)	[1.191, 21.086]
IL1b_days_6_8	Interleukin 1 beta in time interval 3 – INT3 (averaged across days 6 – 8)	[0.246, 0.968]
IL6_days_6_8	Interleukin 6 in time interval 3 – INT3 (averaged across days 6 – 8)	[0.248, 17.788]

Feature	Description	Value range
IL8_days_6_8	Interleukin 8 in time interval 3 – INT3 (averaged across days 6 – 8)	[2.519, 115.776]
TNF_days_6_8 (or TNFa_days_6_8)	Tumor necrosis factor (alpha) in time interval 3 – INT3 (averaged across days 6 – 8)	[1.297, 10.018]
IL1b_days_9_11	Interleukin 1 beta in time interval 4 – INT4 (averaged across days 9 – 11)	[0.055, 2.291]
IL6_days_9_11	Interleukin 6 in time interval 4 – INT4 (averaged across days 9 – 11)	[0.0002, 176.207]
IL8_days_9_11	Interleukin 8 in time interval 4 – INT4 (averaged across days 9 – 11)	[0.652, 56.1]
TNF_days_9_11 (or TNFa_days_9_11)	Tumor necrosis factor (alpha) in time interval 4 – INT4 (averaged across days 9 – 11)	[1.202, 9.408]
Group*	0: Group A, 1: Group B, 2: Group C, 3: Group D	[0, 3]
* Group A: patients who were not admitted to the ICU and survived, Group B: patients who were not admitted to the ICU but died, Group C: patients who were admitted to the ICU but survived, Group D: patients who were admitted to the ICU and died.		

All in all, the quality of the time-series data from the "Sotiria" hospital was significantly enhanced across the available timepoints.

7.6.1.2. Case Study 2 – University Hospital of Ioannina

The scope of this case study is to enhance the quality of a large size clinical database across 7 timepoints with hospitalized COVID-19 patients in Greece [417]. Anonymized baseline and follow up clinical data were acquired from the Dept. of Internal Medicine at the University Hospital of Ioannina. In total, 422 hospitalized COVID-19 patients were included in the analysis with an average age of 64.28 (±16.72) years. The time-series data consisted of 51 clinical features across 7 timepoints: 1, 3, 5, 7, 9, 11, and 15 days after hospitalization. Out of 422 patients, 25 patients (5.92%) were admitted in the ICU and 49 patients died (11.61%). Out of the 49 patients who died, 18 were admitted in the ICU. The classification tasks are formulated as follows: (i) in the first case, the target group consists of the patients who were admitted in the ICU (25 patients), and (ii) in the second case, the target group consists of the patients who died (49 patients). In each case, the remaining patients are assigned to the control group.

The number of features with either good or fair quality status was 70 in timepoint 1, 66 in timepoints 1-2, 55 in timepoints 1-3, 51 in timepoints 1-4, 48 in timepoints 1-5, 28 in timepoints 1-6, and 20 in timepoints 1-7, where the time-points refer to hospitalization days. Consequently, only the 51 features having either fair or good quality status in timepoints 1-4 were considered as eligible for the analysis since the

inclusion of information from additional timepoints would result in information loss due to the bad quality status. The quality status for each one of the 51 eligible features (32 continuous, 19 discrete) is summarized in Supplementary Table 4, where an overall description of the quality of the eligible features across the seven time-points is presented in Supplementary Table 4. Out of the 32 continuous features (Figure 84), 9.37% was good, 65.18% was fair and 25.45% was bad whereas out of 19 discrete features (Figure 84), 15.78% were good, 57.9% were fair and 26.32% were bad, on average, across the available time-points. Data imputation based on the kNN approach was only applied for the features with fair quality. The abbreviations for the input features are presented in Supplementary Table 4.

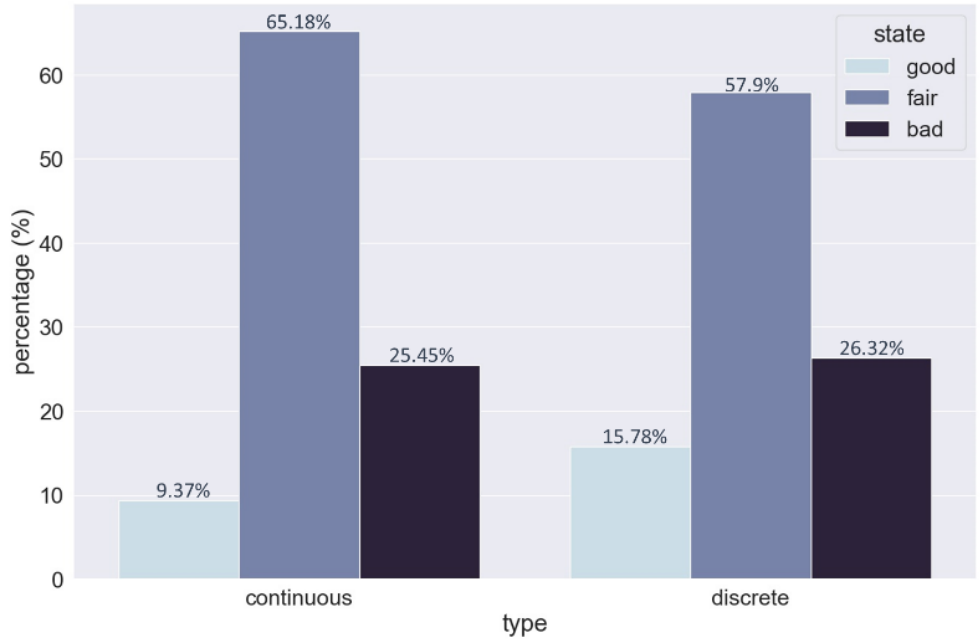


Figure 84. Quality status across the time-points for the continuous and the discrete features.

According to Table 58, the number of discrete features was 19 whereas the number of continuous features was 32. In both cases, the quality of the features is considered as adequate for the analysis until the 4th day of hospitalization.

Table 58. Quality of the features across the seven time-points upon hospitalization.

Time interval	**Continuous features = 32/51 (63%)**			**Discrete features = 19/51 (37%)**		
	Good (%)	**Fair (%)**	**Bad (%)**	**Good (%)**	**Fair (%)**	**Bad (%)**
day 1	3/32 (9.37%)	29/32 (90.63%)	0	3/19 (15.78%)	16/19 (84.22%)	0
day 3	3/32 (9.37%)	29/32 (90.63%)	0	3/19 (15.78%)	16/19 (84.22%)	0

Time interval	Continuous features = 32/51 (63%)			Discrete features = 19/51 (37%)		
	Good (%)	**Fair (%)**	**Bad (%)**	**Good (%)**	**Fair (%)**	**Bad (%)**
day 5	3/32 (9.37%)	29/32 (90.63%)	0	3/19 (15.78%)	16/19 (84.22%)	0
day 7	3/32 (9.37%)	29/32 (90.63%)	0	3/19 (15.78%)	16/19 (84.22%)	0
day 9	3/32 (9.37%)	26/32 (81.26%)	3/32 (9.37%)	3/19 (15.78%)	13/19 (68.44%)	3/19 (15.78%)
day 11	3/32 (9.37%)	4/32 (12.5%)	25/32 (78.13%)	3/19 (15.78%)	0	16/19 (84.22%)
day 15	3/32 (9.37%)	0	29/32 (90.63%)	3/19 (15.78%)	0	16/19 (84.22%)
Total	9.37%	65.18%	25.45%	15.78%	57.9%	26.32%

All in all, the quality of the time-series data from the "University Hospital of Ioannina" was significantly enhanced across the multiple timepoints.

7.6.2. Federated/distributed learning (local case)

7.6.2.1. Case Study 1 – ICU admission and mortality prediction across 3 timepoints

The scope of this case study is to evaluate a multimodal AI-based approach which combines explainable AI models with dynamic modeling methods to shed light into the clinical features of COVID-19. We used the data from Section 7.6.1.1. The performance evaluation results on Groups A, C and D are summarized in Table 59 while the ROC curves for the time interval 1 are depicted in Figure 85. Due to the increased class imbalance in Groups C (43 targets over 171 controls) and D (36 targets over 178 controls), random downsampling with replacement was applied to yield equally numbered patients across the corresponding control and target groups. More specifically, the downsampled controls were matched according to age and gender, where the downsampling ratio was set to 1:1. The overall process was repeated ten times to avoid biases during the downsampling stage. A stratified 10-fold cross validation process was applied on each round and the performance evaluation results were averaged. According to Table 59, the performance of the GBT classifier was favorable, specifically in Groups A and D. The performance of the AI model in Group A yielded an AUC 0.84 in time interval INT1, 0.84 in time intervals INT1-INT2, 0.83 in time intervals INT1-INT3, and 0.81 in time intervals INT1-INT4 towards the classification of the patients who were not admitted in the ICU and survived. The AI model in Group C was able to classify the patients who were admitted in the ICU and

survived with an AUC 0.77 in INT1, 0.76 in INT1-INT2, 0.81 in INT1-INT3, and 0.8 in INT1-INT4. Finally, the AI model in Group D classified the patients who were not admitted in the ICU and died with an AUC 0.84 in time interval 1, 0.84 in INT1-INT2, 0.83 in INT1-INT3, and 0.81 in INT1-INT4. It should be noted that the missing values in INT3 and INT4 affected the performance of the AI models against those trained in INT1.

Table 59. Performance evaluation results across sequential time intervals for Groups A, C, and D. Group B was ignored due to the small number of patients (INT1: days 0 to 2, INT2: days 3 to 5, INT3: 6 to 8 and INT4: days 9 to 11).

INT1				
Groups	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**
Group A*	0.77	0.77	0.71	0.84
Group C**	0.73	0.73	0.75	0.77
Group D**	0.77	0.77	0.78	0.83
INT1-INT2				
Groups	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**
Group A*	0.79	0.79	0.71	0.84
Group C**	0.72	0.72	0.72	0.76
Group D**	0.77	0.77	0.77	0.84
INT1-INT3				
Groups	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**
Group A*	0.77	0.77	0.70	0.83
Group C**	0.78	0.77	0.80	0.81
Group D**	0.79	0.78	0.81	0.86
INT1-INT4				
Groups	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**
Group A*	0.77	0.77	0.69	0.82
Group C**	0.77	0.77	0.77	0.80
Group D**	0.81	0.80	0.82	0.85
* a stratified 10-fold cross validation procedure was used. ** random downsampling with replacement was applied to match the control group with the target group due to the increased class imbalance.				

According to Figure 85, the classification performance was favorable in all groups (in terms of the true positive rate versus the false positive rate), where the AUC score was 0.87, 0.79, and 0.88, for Groups A, C, and D, respectively.

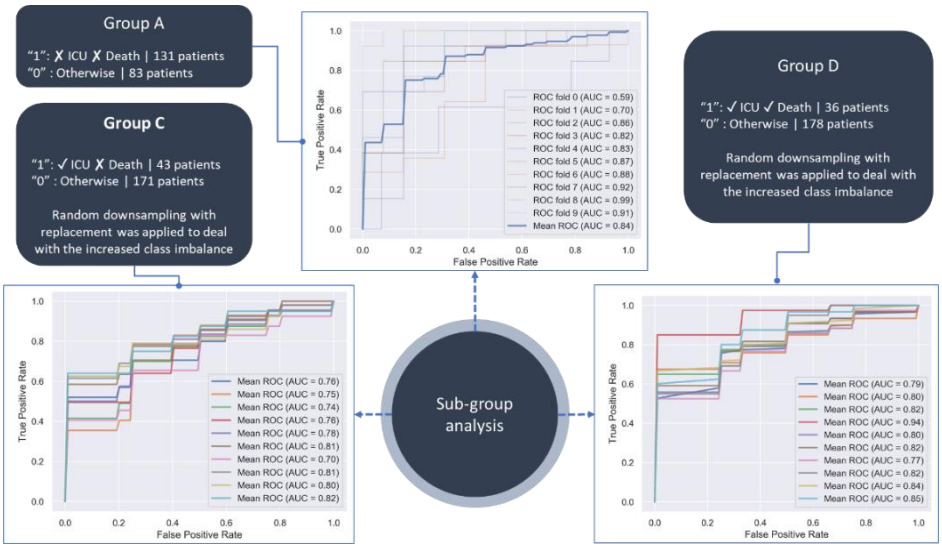


Figure 85. ROC curves of the GBT classifier on Groups A, C, and D in time interval 1.

The mean absolute Shapley values which quantify the average impact of each feature on the model's output magnitude are depicted in Figure 86 (on the left subpanel) along with the Shapley values that quantify the impact of the corresponding feature on the model output (on the right subpanel).

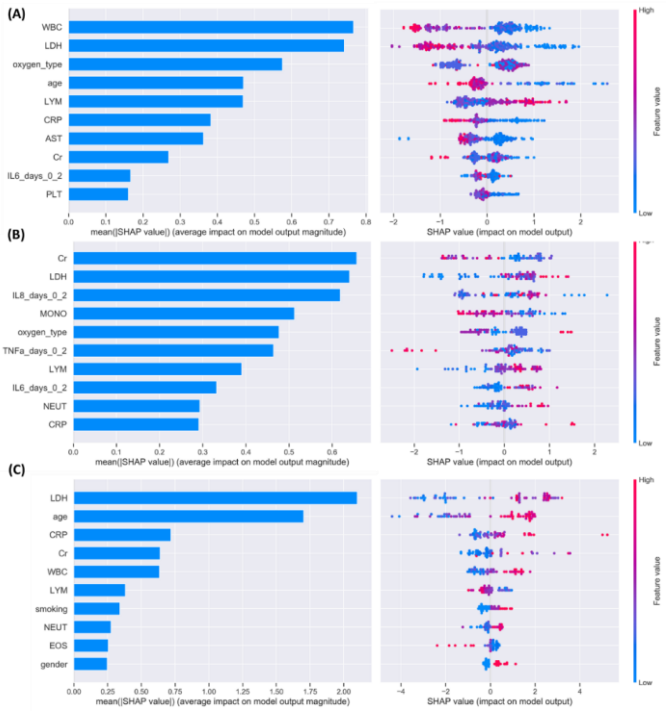


Figure 86. Prominent features across (A) Group A, (B) Group C, and (C) Group D, using the baseline data along with the cytokines from time interval 1 (INT1).

Since the objective function of the GBT classifier is set to the logistic loss, the Shapley values correspond to the log-odds. Thus, features that significantly affect the model's output from the base value (i.e., the average model output) to higher log-odds are depicted in red whereas features that affect the average model's output to lower log-odds are depicted in blue.

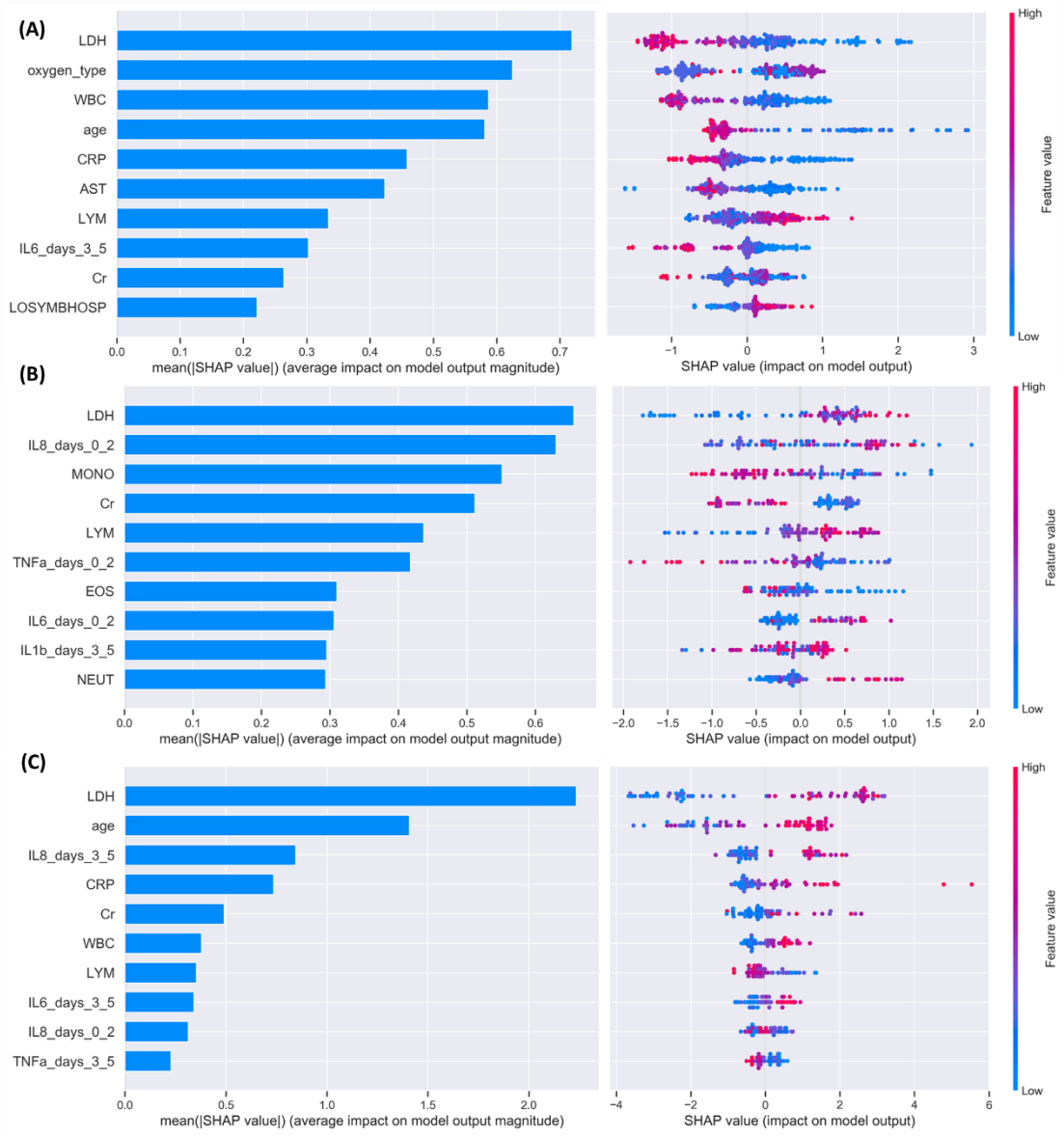


Figure 87. Prominent features across (A) Group A, (B) Group C, and (C) Group D, using the baseline data along with the cytokines from time intervals 1-2.

The average absolute Shapley values for Group A are depicted in Figure 86 (A) in a descending order (on the left subpanel) along with the Shapley values (on the right) which quantify the positive or negative impact of the 10 most prominent features on the

model's output. According to Figure 86 (A), WBC had the highest contribution to the decision-making process by affecting the model's output to higher log-odds for low white blood cell (WBC) values along with Lactate Dehydrogenase (LDH), age, C-reactive protein (CRP), Aspartate Aminotransferase (AST), number of platelets (PLT), and IL-6. Other features include the number of lymphocytes which affect the model's output to higher log-odds but for higher values.

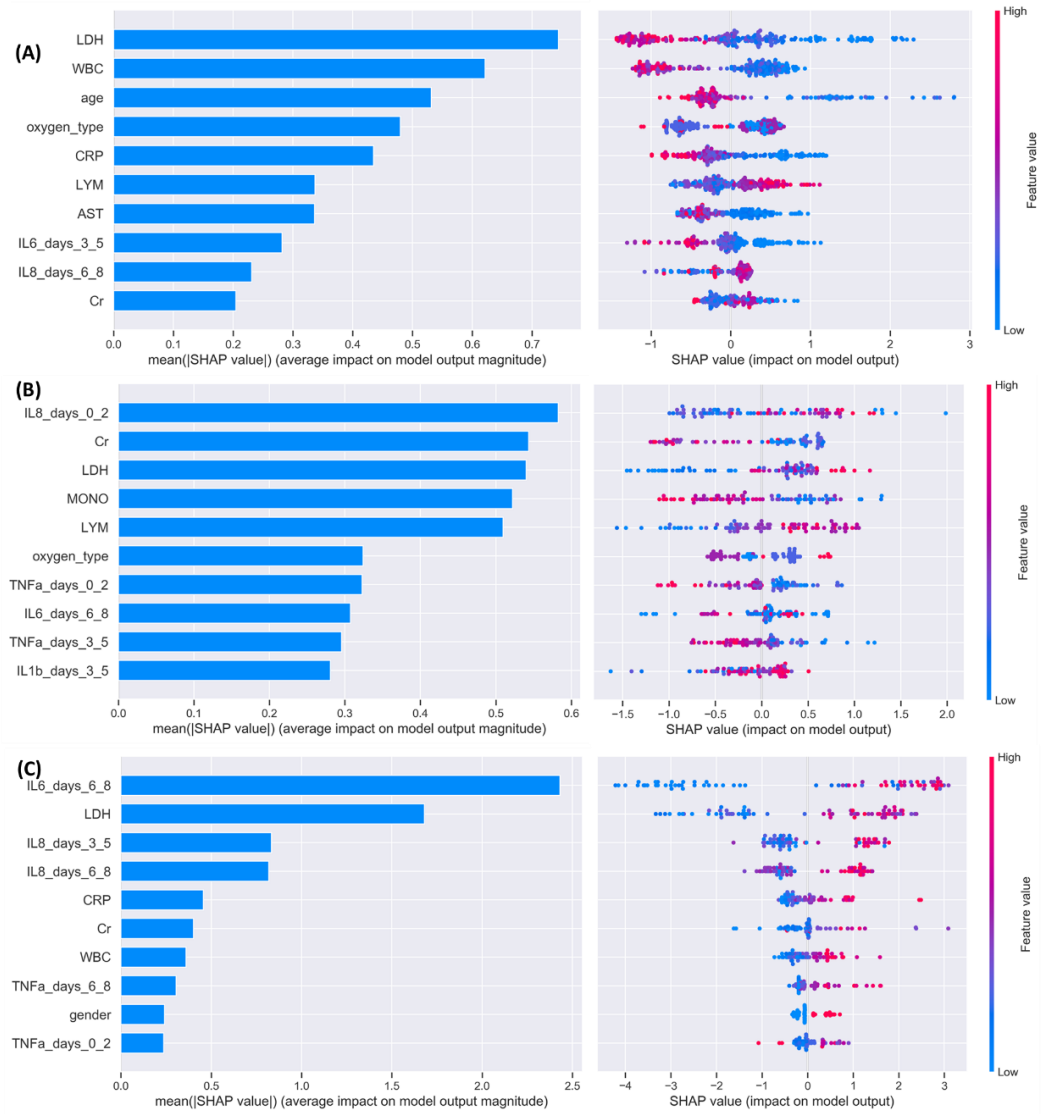


Figure 88. Prominent features across (A) Group A, (B) Group C, and (C) Group D, using the baseline data along with the cytokines from time intervals 1- 3.

Regarding Group C, (Figure 86 (B)) creatinine (Cr) and LDH had the highest contribution in the classification outcome, along with the IL-8, number of monocytes (MONO), oxygen type and TNF, where on one hand both low and high values of these features affect the model's output to higher log-odds but on the other hand small LDH,

IL-6, number of lymphocytes (LYM), WBC values affect the model's output to lower log-odds. Finally, according to Figure 86 (C), LDH had the highest impact during decision-making in Group D, along with age, CRP, Cr, and WBC, among others, where large values for age and WBC affect the model's output to higher log-odds but for higher values.

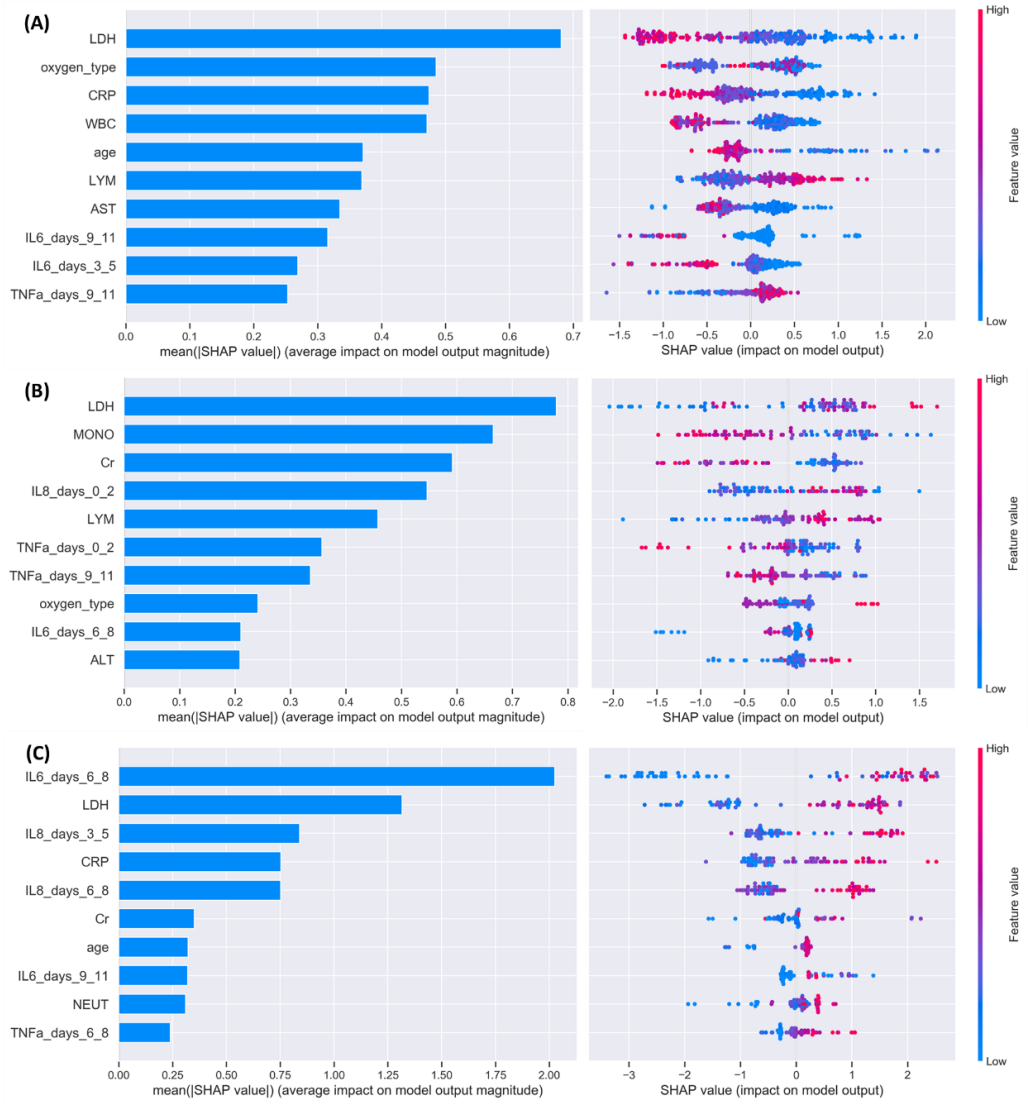


Figure 89. Prominent features across (A) Group A, (B) Group C, and (C) Group D, using the baseline data along with the cytokines from all time intervals.

According to Figure 87, the overall importance in Group A using the cytokines from INT2 is preserved, where the LDH, WBC, age, and CRP continue to appear as prominent, as well as, IL-6 but on INT2. Regarding Group B, the features LDH, IL-8, MONO, Cr and LYM have the highest contribution to the model's output. As far as Group D is concerned, the contribution of LDH, age, CRP, and Cr is also dominant.

According to Figure 88, the overall importance in Groups A and B using the cytokines from INT1-INT3 is preserved. Regarding Group D, the contribution of LDH, IL-6 in INT3 and IL-8 in INT2 and INT3 appear to be important affecting the model's output to higher log-odds but for higher values. Finally, according to Figure 89, the overall importance in Groups A and B using the cytokines from all time intervals is preserved with updates in the ranking order. As for Group D, the contribution of the LDH, IL-6 in INT3 and INT4 and IL-8 in INT2 and INT3 appear to be important.

To better understand the similarities among the Shapley values of each prominent feature, heatmaps were also derived (Figure 90), where the horizontal axis depicts the instances in ascending order, the vertical axis depicts the features ranked in descending order based on their classification importance, and the color coding corresponds to the Shapley explanation value levels across the instances in the whole dataset. Hierarchical clustering was then applied based on the explanation similarity of the most prominent features to identify activation patterns among the patients.

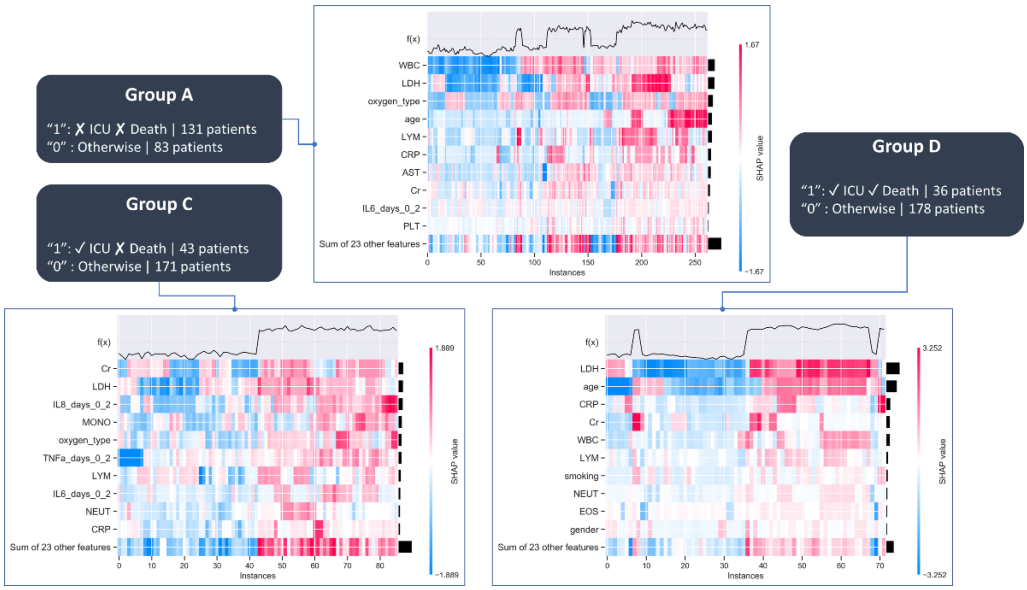


Figure 90. Heatmaps for Groups A, C and D, using the baseline data along with the cytokines from time interval 1 (INT1).

According to Figure 90, the instances that exhibit increased explanation values using the cytokines from INT1, include the WBC, LDH, and oxygen type, which implies that these features can be used to derive homogeneous clusters and are in concordance with the feature importance plots in Figure 86. A similar pattern is observed in Group C for LDH along with IL-8, oxygen type, and TNF which are also reported in Figure 86.

Regarding Group D, the LDH is an important factor for hierarchical clustering, along with the age which exhibits strong explanation similarities with the outcome. A similar behavior regarding the contribution of the prominent features from the Shapley explanation analysis to the patterns across the derived hierarchical clusters was observed in the case where the cytokine measurements from INT1-INT2, INT1-INT3, and INT1- INT4 are used.

Decision trees were induced to further enhance the interpretability of the groupwise AI models by capturing the decision pathways which are involved in the decision-making process (Figure 91). Towards this direction, the CART algorithm [416] was applied on the baseline and cytokine data from each individual group and across sequential time intervals to identify critical thresholds for the prominent features, i.e., the features which are highly involved in the decision-making process; excluding Group B due to the small number of patients (Section 7.6.1.1).

According to Figure 91, the decision-making process in Group A, using the baseline data and the cytokines from INT1, is based on WBC since it is the root of the induced decision tree. The threshold 7.58 in WBC indicates a critical value that determines whether the decision will be based on CRP in case it is less than (or equal to) 7.58, where additional emphasis is given on Cr (with a critical threshold at 1.5; values less than or equal to 1.5 are classified as positive) and PLT (with a critical threshold at 243.5; instances with values larger than 243.5 are classified as positive). Otherwise, the decision-making process follows the right pathway which is based on the lymphocyte count with a critical threshold at 1.405, where in the case that this is lower than or equal to 1.405 the decision is based on AST (values less than or equal to 22 are classified as positive) or on age (values less than or equal to 82.5 are classified as positive) in the case where the lymphocyte count is higher than 1.405. It is interesting that in the case where CRP is less than (or equal to) 3.637 and Cr is less than (or equal to) 1.5, the instance is classified as positive (i.e., no admission in the ICU and survival).

When CRP is larger than 3.637 and PLT is higher than 243.5, the instance is also classified as positive. In the case where WBC is higher than 7.56 the instance is classified as positive either when LYM is lower than (or equal to) 1.405 and AST is less than (or equal to) 22 or when LYM is higher than 1.404 and the age is less than (or equal to) 82.5.

As far as Group C is concerned (Figure 91), the decision-making process is based on LDH. The threshold 278 in LDH indicates a critical value that determines whether the decision will be based on IL-8 in time interval 1 with a threshold at 6.355, where emphasis is given on MONO (values less than or equal to 0.635 are classified as positive) in the case where IL-8 is less than or equal to 6.355 or again on MONO (values less than or equal to 0.305 are classified as positive) otherwise. Otherwise, the decision-making process follows the right pathway where emphasis is given on LYM with a critical threshold at 1.094 where in the case it is lower than 1.094 emphasis is given on IL6 at INT1 (values larger than 9.447 are classified as positive) otherwise on WBC (values less than or equal to 5.418 are classified as positive).

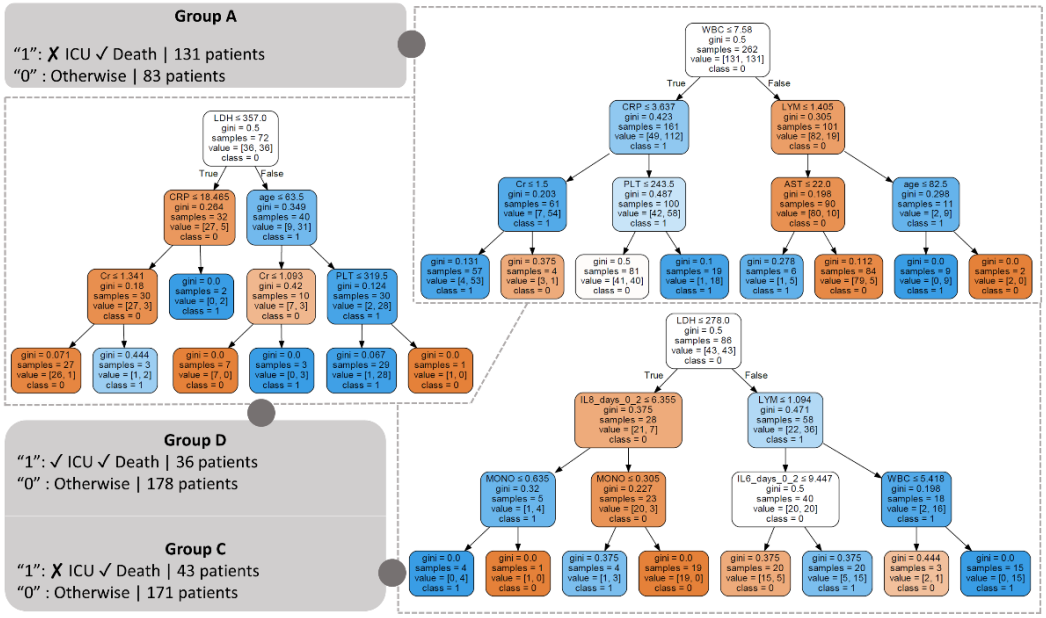


Figure 91. Induced decision trees across (A) Group A, (B) Group C, and (C) Group D using the baseline data along with the cytokines from the time interval 1 (INT1; days 0 to 2). In each case, blue color gradings denote instances which are classified as positive (i.e., outcome = "1") whereas orange color gradings denote otherwise (i.e., outcome = "0").

It is interesting that in the case where LDH is less than (or equal to) 278 and IL-8 is less than (or equal to) 6.355, and MONO is less than (or equal to) 0.635 the instance is classified as positive (i.e., admission in the ICU and survival). The same occurs in the case where IL6 is larger than 6.355 and MONO is less than 0.305. However, when LDH is larger than 278 and LYM is larger than 1.094 and WBC is larger than 5.418 the instance is classified as positive. The same occurs when LYM is less than (or equal to) 1.094 and IL-6 is larger than 9.447.

Regarding Group D (Figure 91), the decision-making process is once more based on the LDH. The threshold 357 in LDH indicates a critical value which determines whether the decision will be based on CRP (with a critical threshold at 18.465; values larger than 18.465 are classified as positive) in the case where the LDH is less than (or equal to) 357, where emphasis is given on Cr (values larger than 1.341 are classified as positive). Otherwise, the decision-making process follows the right pathway where the decision is based on age with a critical threshold at 63.5 years, where in the case it is larger than 63.5 emphasis is given on PLT (values larger than 319.15 are classified as positive) or on Cr (values larger than 1.093 are classified as positive) otherwise. The acronyms of the features which participate in the decision-making process (Figure 91) are described in Table 57.

We described a multimodal AI approach based on an anonymized dataset of 324 hospitalized patients who have been diagnosed with COVID-19, in Greece, that includes laboratory and clinical information, as well as, biological information across four time intervals. The pipeline utilizes explainable and interpretable AI models along with dynamic modeling methods to support decision making for ICU admission and/or mortality and shed light into the pathogenesis and clinical features of COVID-19. Data curation is first applied to overcome data incompatibilities and inconsistencies. Subgroup analysis is performed by dividing the curated data into four subclasses of interest based on the ICU admission and/or mortality. Gradient Boosting Trees (GBT) are trained on each subgroup to develop explainable AI models using concepts from coalition game theory to detect risk predictors for ICU admission and mortality, as well as, to evaluate the predictors across four time intervals.

Our results highlight the importance of LDH, IL-6, IL-8, Cr, number of monocytes, lymphocyte count, and TNF as risk predictors for ICU admission and survival, as well as, LDH, age, CRP, Cr, WBC, and lymphocyte count for mortality after ICU admission. These predictors were combined with those from the dynamic analysis of the biological data using Dynamic Bayesian Networks (DBNs) to formulate an ICU scoring index based on APACHE II [416], where the DBNs revealed notable dependencies between TNF and IL-6. To our knowledge, this is the first study that explores the interpretability of AI models and risk predictors for ICU admission and mortality of hospitalized COVID-19 patients with dynamically associated biological markers.

7.6.2.2. Case Study 2 – ICU admission and mortality prediction across 6 timepoints

The scope of this case study is to enrich time-series clinical and laboratory data with meta information from SOMs towards the improvement of the ICU admission and mortality classifiers in COVID-19. We used the data from Section 7.6.1.2. A 7x7 grid was utilized for the neuron training process. The latter was applied on the 32 continuous features with "fair" or "good" quality status at timepoints 1-4 like in the DBN analysis. Clusters with common patterns were further grouped into four super-clusters through hierarchical clustering. The distribution of the patients in each super-cluster is presented in Table 60, where the average number of patients is 117 (27.72%), 108 (25.6%), 88 (20.85%), and 109 (25.83%) in super-clusters 1, 2, 3, and 4, respectively. Statistically significant differences were identified in the patient distribution for features "Hct", "Lymph_abs_number", "Lymph_percent", "Neut_abs_number", "Neut_percent", "PO2_FiO2_ratio" regarding ICU admission and mortality. Additional differences among the patient subgroups were found in "AST" for ICU admission and in "ALP" and "LDH" for mortality.

Table 60. Number of patients assigned in each SOMs super-cluster for the most important features from the DBNs (p-values in bold denote significant differences among the distributions of the ICU against the non-ICU patients and the patients who survived against those who died).

Feature	**Patient distribution in each super-cluster**				**p-value***	
	C1	**C2**	**C3**	**C4**	**ICU**	**mortality**
ALP	88	223	83	28	0.732	**0.04**
AST	173	71	92	86	**0.005**	0.285
cardiac_frequency	86	68	145	123	0.905	0.103
Hct	107	167	44	104	**<0.001**	**0.0001**
LDH	80	61	101	180	0.061	**0.005**
Lymph_abs_number	82	82	84	174	**0.015**	**0.033**
Lymph_percent	102	105	79	136	**0.024**	**0.0007**
Neut_abs_number	130	148	95	49	**0.016**	**0.005**
Neut_percent	148	95	74	105	**0.0008**	**0.0003**
PO2_FiO2_ratio	132	74	87	129	**<0.001**	**0.004**
Tbil	166	89	79	88	1	0.319
Average patient distribution	117	108	88	109		
* A Fisher's exact test was applied where the confidence level was set to 95%.						

Three case studies were investigated which involve the classification of patients for ICU admission and mortality (Table 61) based on: (i) the 51 time-series clinical data across the first 4 timepoints with and without the inclusion of the 32 features with the clustering labels from the SOMs, (ii) the 11 features from the DBNs analysis with and without the clustering labels from the SOMs, and (iii) only with the clustering labels from the SOMs. In case study 1, the contribution of the clustering labels from the SOMs enhanced the sensitivity by 1% and the specificity by 2% of the classifier for ICU admission against the use of the time-series data only. In case study 2, the contribution of the clustering labels from the SOMs enhanced the sensitivity and specificity of the classifier for ICU admission by 4% compared against the use of the best features from the DBNs, as well as, by 3% in sensitivity and 2% in specificity for mortality (Table 61). In case study 3, the use of the clustering labels from the SOMs yielded favorable classification performance. According to Table 61, the performance of the classifiers was higher using the clustering labels from the SOMs for both mortality (in case study 1) and ICU admission (in case study 2), thus highlighting the positive impact of the DBNs and the SOMs during the training process. This can be also confirmed even in the case where no class imbalance handling is applied, where the performance of the classifiers remains higher using the clustering labels from the SOMs for both mortality (in case study 1) and ICU admission (in case study 3).

Table 61. Performance evaluation results from the GBT for ICU and mortality classification across different cases with donwsampling using the SOMs clustering labels from all the 32 continuous features (with blue color: specifications with the best or equal classification performance).

Case	Outcome	SOMs	accuracy	sensitivity	specificity	AUC
Case study 1*: 51 features across 4 timepoints with and without the clustering labels from the SOMs	**death**	no	0.74	0.74	0.76	0.83
	death	yes	0.74	0.74	0.76	0.83
	ICU	no	0.78	0.79	0.79	0.88
	ICU	yes	0.79	0.80	0.82	0.89
Case study 2*: 11 features from DBNs across 4 timepoints with and without the clustering labels from the SOMs	**death**	no	0.67	0.67	0.70	0.74
	death	yes	0.70	0.70	0.72	0.76
	ICU	no	0.78	0.79	0.78	0.87
	ICU	yes	0.83	0.83	0.82	0.91

Case	Outcome	SOMs	accuracy	sensitivity	specificity	AUC
Case study 3*: Only with the clustering labels from the SOMs	**death**	yes	0.67	0.67	0.68	0.74
	ICU	yes	0.80	0.80	0.82	0.86
*Random downsampling with replacement was applied to deal with the underlying class imbalance.						

The corresponding ROC curves are depicted in Figure 92 for ICU and mortality classification across the three case studies from Table 61. Regarding the performance of the classifier for ICU admission, the average ROC was 0.89 for case 1, 0.91 for case 2, and 0.86 for case 3. As far as mortality classification is concerned, the average ROC was 0.83 for case 1, 0.76 for case 2, and 0.74 for case study 3.

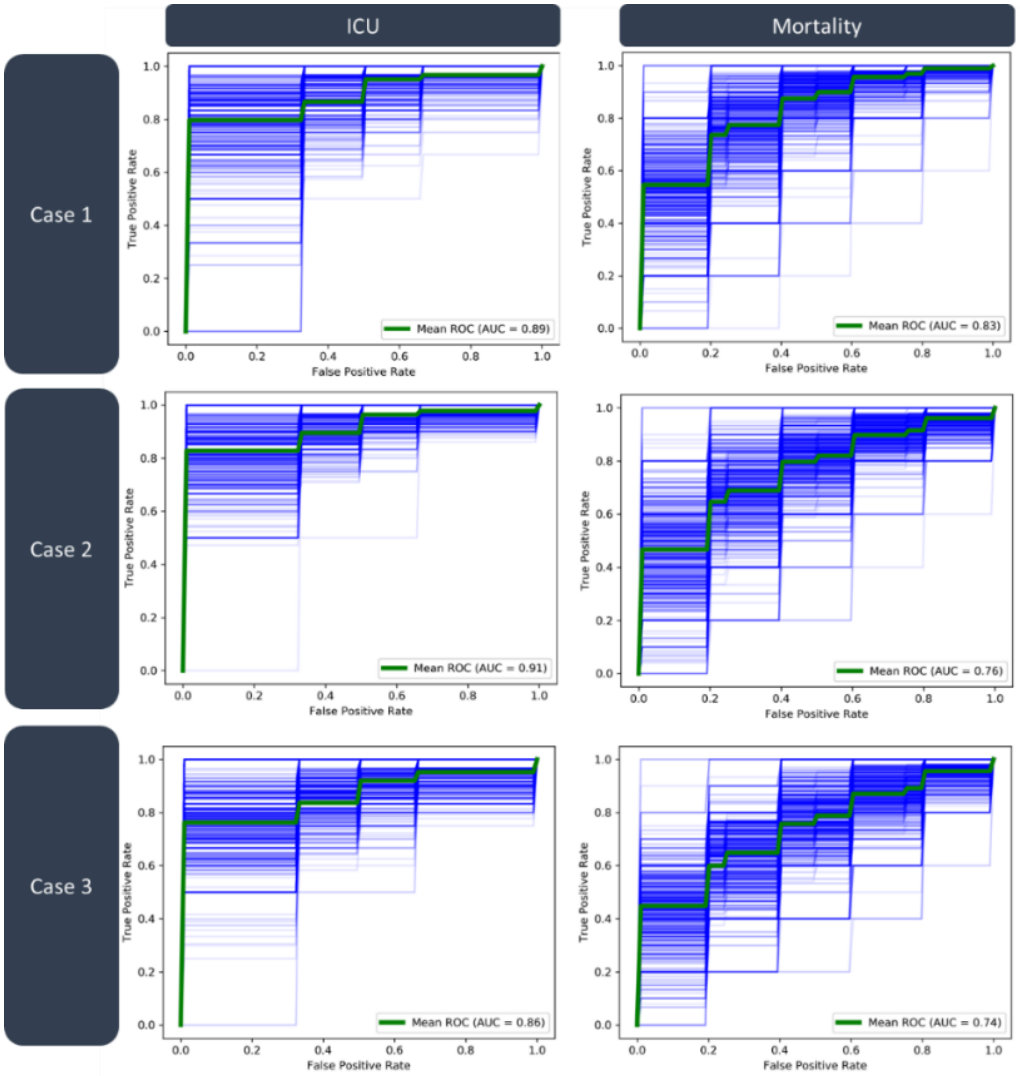


Figure 92. Performance evaluation results for the GBT with the clustering labels from the SOMs. The line in bold denotes the average ROC across 100 iterations of the downsampling process.

According to Figure 93, the risk factor analysis highlighted the following features as important (i.e., the top five features) for ICU admission in case study 1 (with the clustering labels from the SOMs): O2_supply_type_day5", "O2_supply_type_SOM", "SatO2_day7", "tachypnea_day5", and "SBP_day7". The rest of the features include "temperature_day7", "secondary_O2_supply_lit_SOM", "PCO2_day3", "K_day3", and "DBP_day3". Regarding mortality, the most informative features for decision making, include the: "Lymph_percent_day7", "Urea_day5, "ALP_day1", "Neut_percent_day7", and "Hb_day1". Additional features include the "tachypnea_day_3", "INR_day1", "PO2_FiO2_ratio_day5", "hs_TPN_day1", and "FiO2_day5". The important features with the "SOM" tag denote the features with the clustering labels.

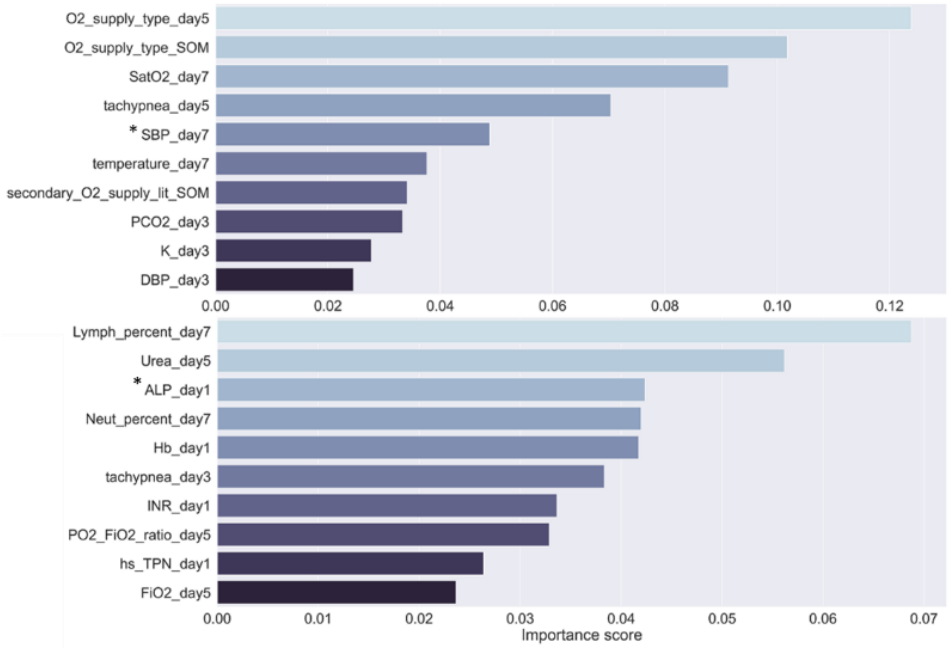


Figure 93. Feature importance for ICU admission (on top) and mortality (on bottom) from case study 1 with the clustering labels from the SOMs.

According to Figure 94, the risk factor analysis indicated the following features as important for ICU admission in case study 2 (with the clustering labels from the SOMs): "PO2_FiO2_ratio_day5", "Lymph_abs_number_day5", "O2_supply_type_SOM", "PO2_FiO2_ratio_day7", and "Lymph_percent_day3", among others. Regarding mortality, the most important features for decision making include the: "PO2_FiO2_ratio_day5", "Hct_day1", "ALP_day1", "LDH_day5", and "Neut_abs_number_day7", among others.

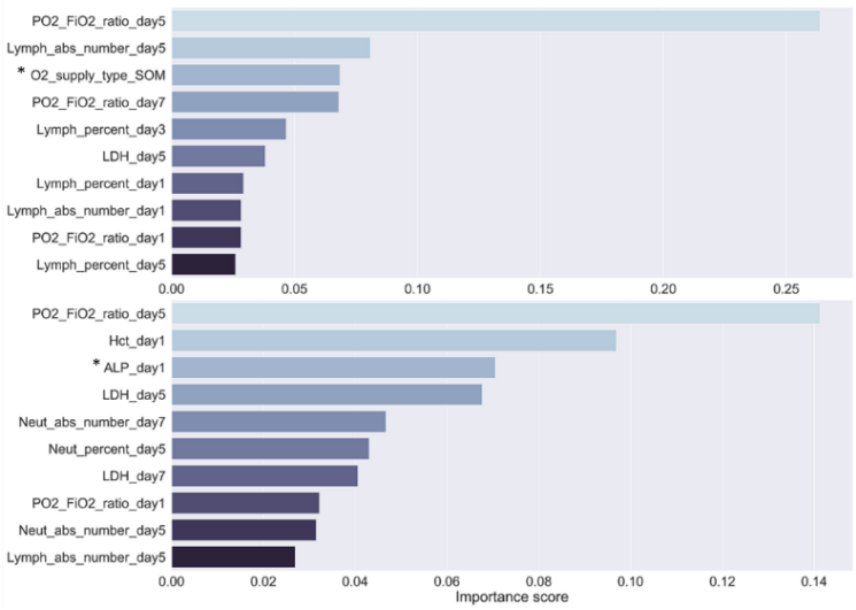


Figure 94. Feature importance for ICU admission (on top) and mortality (on bottom) from case study 2 with the clustering labels from the SOMs.

According to Figure 95, the analysis highlighted the following features as important for ICU admission in the case study 3: “O2_supply_type_SOM”, “temperature_SOM”, “secondary_O2_supply_lit_SOM”, “SatO2_SOM”, and “cardiac_frequency_SOM”, among others. Regarding mortality classification, the most important features include the: “SatO2_SOM”, “secondary_O2_supply_lit_SOM”, “Na_SOM”, “ALP_SOM, and “Creatinine_SOM”, among others.

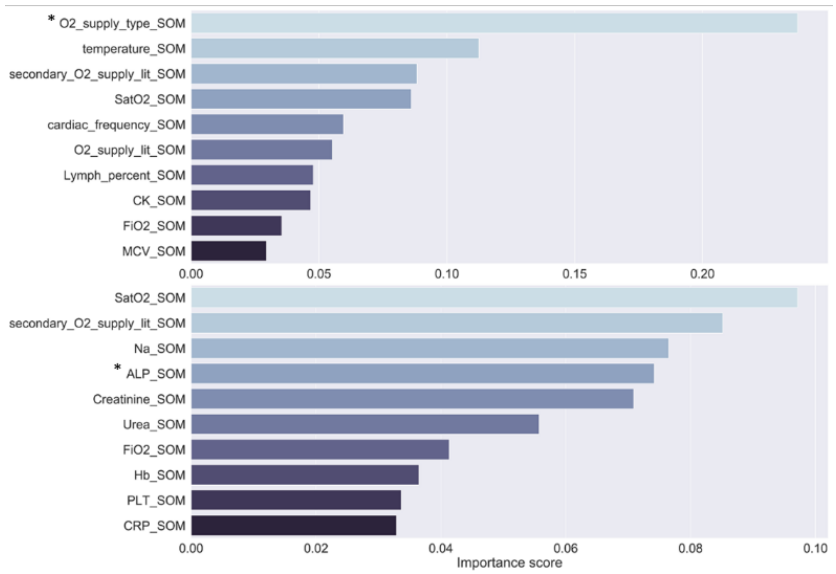


Figure 95. Feature importance for ICU admission (on top) and mortality (on bottom) from case study 3 with the clustering labels from the SOMs.

In all cases, the clustering labels from the SOMs regarding the O2 supply type and the feature "ALP" were prominent for ICU admission and mortality, respectively (these features have been denoted with asterisks in Figure 93, Figure 94, and Figure 95. An additional experiment was conducted to evaluate the contribution of baseline data including demographics (e.g., age, gender, patient history), clinical (e.g., fever, fatigue, dyspnea), and treatments (e.g., administration of various therapeutic treatments, such as, statin, betablocker, corticosteroids) in the case study where the GBTs achieved the best performance in case study 2. According to Table 62, the inclusion of demographics, clinical, and treatments did not yield any improvement in the performance of the classifier for ICU admission. On the other hand, the sensitivity of the classifier for mortality was improved by 4% using the demographic data. The specificity was improved by 4% in the case where the demographics are included and by 1% in the case where the baseline clinical data and the treatments were included.

Table 62. Performance evaluation results for case study 2 before and after the inclusion of demographics, clinical data, and treatments (with blue color: specifications with the best or equal classification performance).

Outcome	Accuracy	Sensitivity	Specificity	AUC
ICU admission				
Before	0.83	0.83	0.82	0.91
After adding demographic data	0.83	0.83	0.81	0.90
After adding clinical data	0.81	0.81	0.80	0.89
After adding treatments	0.83	0.83	0.82	0.91
Mortality				
Before	0.70	0.70	0.72	0.76
With demographic data	0.74	0.74	0.76	0.82
With clinical data	0.70	0.70	0.73	0.76
With treatments	0.70	0.70	0.73	0.76

Three case studies were conducted to evaluate the performance improvement in classifying the patient subgroups derived from the SOMs. Our results highlight the contribution of the extracted patient subgroups in the improvement of the classification performance for ICU admission up to sensitivity 0.83 and specificity 0.83, and for mortality up to sensitivity 0.74 and specificity 0.76. Additional baseline data were included in the input space to improve the performance of the classifiers, yielding an

increase of 4% in sensitivity and specificity for ICU admission and 3% in sensitivity and 2% in specificity. The risk factor analysis highlighted the number of lymphocytes, SatO2, PO2/FiO2, and O2 supply type as risk factors for ICU admission and the percentage of neutrophils and lymphocytes, PO2/FiO2, LDH, and ALP for mortality.

CHAPTER 8. Discussion

8.1. Technical impact

8.1.1. Curation

The absence of data quality control often leads to studies with inaccurate and incomplete data which are characterized by small statistical power. On the other hand, curation methods shall be used with caution since it is more likely to make things worse. In this thesis, we proposed an automated framework for medical data curation supporting clinical and genetic data across multiple time-points. Furthermore, we extended data standardization as a pre-harmonization process to make data harmonization, which follows, easier and faster. Through this procedure, we produced semantic relations between the fields of the raw dataset with those from a reference dataset and therefore enhance the semantic matching process for data harmonization.

The proposed framework consists of a three-layer architecture which is scalable and able to deal with incomplete terminologies, irrelevant terms, outliers, missing values, data categorization, and duplicated terms. In the core of this framework lies data standardization. The framework was evaluated on two anonymized clinical datasets from two cohorts of patients with pSS, highlighting the importance of the proposed framework for data quality assessment and data harmonization. The source code of the data curator has been made publicly available under the following github repository: https://github.com/vpz4/Data-curator, along with a brief user manual to further promote technical advancements in the field of data quality and control. The fact that the standardization procedure can use an XML representation of the reference model as

input, increases its overall performance and introduces the ontologies and semantics as a preliminary step for achieving medical data harmonization. Having two ontologies and seeking for a way to match these two ontologies is a typical semantic matching problem which is one way to achieve data harmonization.

Semantics have gained a lot of attention nowadays especially in computer science and linguistics for schema and ontology merging, data migration, query translation, agent communication, etc.. Currently, the data standardization module supports the basic XML format (the basis of almost all types of ontologies and markup languages) for the semantic representation of the reference model. The outcomes of this module are capable of assisting the semantic matching process which requires these matching pairs in order to semantically match each term of the input dataset with those from the reference model and thus enable data harmonization. The problem of harmonizing one dataset based on a standard one can be reduced to the standardization of the semantic representations of these two datasets so as to approximate the semantic matching problem.

8.1.2. Harmonization

Most of the current efforts towards data harmonization involve the definition of a global standard (common) procedure for data collection which is ideal in the case of prospective data (i.e., data that will be collected in the future). Moreover, they adopt a semi-automated strategy which is based on the extensive collaboration between the clinical and the technical experts to define pairing rules, i.e., a pre-defined set of rules for lexical matching, and focus on fast matchmaking services for biobanks and researchers. To this end, we focused on the development of a fully automated, hybrid terminology interlinking pipeline which utilizes both lexical and semantic analyzers on top of a knowledge base on obesity, including word embeddings, NLTK (Natural Language Toolkit) synonyms and FHIR related terminologies from international medical index repositories to automatically match terminologies across a variety of clinical data.

A complete medical corpus is constructed, and the hybrid data harmonizer was developed on top of the medical corpus to identify terminologies with both lexical and semantic overlap. The medical corpus was enriched with: (i) synonyms from Python's

NLTK toolkit, (ii) word embeddings from Latent Semantic Analysis (LSA), and (iii) international terminologies from external repositories under the aegis of the OHDSI Athena vocabulary, including ICD-10/11/12, SNOMED-CT, LOINC, ICPC, ATC, and OMOP terminologies to ensure FHIR compliance.

8.1.3. Synthetic data generation

Regarding the proposed MVND approach which is resilient against randomness, our results demonstrate the robustness and accuracy of the proposed method towards the generation of virtual clinical data for *in silico* clinical trials with high level of agreement between the densities and the distributions of the virtual and the real clinical datasets. The gof values of the proposed method were less than 0.2 compared to the average gof values obtained across 10 random executions (> 0.2 in some cases). Although the number of iterations needed to control for the randomness of the generated virtual distribution was large enough (~5000), the execution time was small considering the number of virtual patients and the number of iterations needed to assess for the "randomness" factor. The lack of significant computational complexity followed by an increase in the quality of the virtual data is an advantage of the proposed methodology, especially in the case of large-scale *in silico* clinical trials where the number of virtual patients to be generated is significantly larger. As a future work additional methods for virtual population generation, such as the Bayesian networks [348], [418] and the modified genetic function [419], along with neural network-based strategies [173] will be employed for comparison.

Regarding the application of tree ensembles for synthetic data generation in the domain of HCM, our results reveal the favorable performance of the unsupervised tree ensembles for virtual population generation which outperformed the rest of the virtual population generation methods having the smallest goodness of fit and Kullback-Leibler divergence values in both experimental case studies (Table 32). The histograms of the virtual data that were generated by the unsupervised tree ensembles can be found in Figure 63 for the HCM dataset. The histograms reflect a highly qualitative similarity between the real and the virtual distributions. The supervised tree ensembles had the second-best performance (Table 32). The results from the supervised RBF-based artificial neural networks (ANNs) are close to the two previous methods, with the Bayesian networks and the log-MVND trailing behind.

Regarding data augmentation, the performance of the HCM risk stratification model showed an increase by accuracy, 16.9% in sensitivity, 13.7% in specificity, and 20.1% in area under the curve against the one trained on the real HCM data (Table 39, Figure 74). A similar increase is also observed in the case of the AdaBoost (5.5% in accuracy, 5.3% in sensitivity, 6.3% in specificity, and 10.1% in AUC), as well as, in the case of the Random Forests (9.4% in accuracy, 10.1% in sensitivity, 7.2% in specificity, and 12.2% in AUC). In addition, the aggregation of the virtual data from the supervised tree ensembles with the real patient data yielded enhanced classification models at a similar extent (Table 39, Figure 74). Finally, the aggregation of the virtual data from the supervised RBF-based ANNs, the Bayesian networks and the Log-MVND with the real one yielded supervised machine learning models with partially enhanced performance while maintaining the increased performance than in the case of training on the real data only. Our study builds on principles from existing studies (Table 3) to develop a beyond the state-of-the art computational pipeline for clinical data augmentation. We extended the conventional statistical approaches, such as, the MVND and the Log-MVND, as well as, multivariate functions, such as, Bayesian methods, discrete re-sampling techniques, through machine learning based generators, such as, the tree ensembles, the RBF-based ANNs and the Bayesian networks to produce high-quality virtual patient data with increased similarity and decreased divergence with the real patient data.

As far as "smart" imputation is concerned, a case study was conducted in the context of *in silico* clinical trials for the HCM domain, where the real patient dataset was randomly contaminated with missing values for multiple ratios varying from 10% to 50%. The BGMM generator yielded 10000 virtual patient profiles compared against the rest of the generators with less than 0.02 KL divergence and average correlation difference. The PMS was calculated for each virtual patient profile and the best matching profiles were extracted as those with the smallest dissimilarity between the proposed values for imputation and the original ones. The BGMM generator provided imputed values with the lowest average SSAD (0.4) and average CD (0.02) with the real dataset which can be confirmed by the extracted heatmaps since the optimal profiles are separated by intense color coding. To our knowledge, this is the first "smart" method that offers explainable heatmaps compared against the existing frameworks [67], [69], [70], [136], [137], [139], [140] which provide either manual or

semi-automated workflows based on pre-defined semantic data models to address data completeness.

The proposed method for the estimation of the Dirichlet concentration of each component on the weight distribution yielded a stable number of components (24 components) across multiple virtual populations executions, where the prior structure of the GMM was defined according to the Dirichlet process mixture. The proposed BGMM with the optimal number of Gaussian components achieved the lowest goodness of fit values (less than 0.1) along with the UTE and the STE compared to the RBF-based ANN and the Bayesian networks (with average gof larger than 0.15). In addition, the proposed BGMM method yielded the lowest inter- and intra-correlation differences between the features in the virtual data (almost 0.01), in less execution time (0.4319 sec) than the STE (46.5373 sec), which had the second-best performance. This confirms the computational efficiency of the proposed BGMM approach towards the generation of large-scale virtual populations for *in-silico* clinical trials in HCM.

We also designed a robust and computationally efficient large scale virtual data generator to overcome the lack of sufficient population size and leverage the increased costs for patient recruitment for *in silico* clinical trials. Our intention was to resolve significant biases which are introduced by the estimation of the hyperparameters during the conventional BGMM training process. To do so, we introduced the BGMM-OCE, a computationally efficient and robust extension of the BGMM which was designed to: (i) avoid the use of an arbitrary number of Gaussian components through a computationally efficient spectral clustering stage based on the LOBPCG eigensolver, (ii) provide non-linear estimation of the gamma parameter through an exponential relationship with the optimal number of Gaussian components.

Through this way, the BGMM-OCE yields non-arbitrary estimations of the VI at reduced computational complexity. A case study was conducted to generate diverse virtual populations varying from small scale (e.g., 1000 virtual patients) to large scale (e.g., 30000 virtual patients), where the BGMM-OCE outperformed state-of-the art data generators, yielding the lowest coefficient of variation (0.046), goodness of fit (0.191), KL divergence (0.049), and inter- and intra- correlation differences (0.017, 0.016) at stable execution time (16.12 sec) for increased virtual populations. In addition, we have uploaded the BGMM-OCE script in a public GitHub repository:

https://github.com/vpz4/BGMM-OCE to increase the impact of our work and promote the development of a Python package.

In practice, spectral clustering is computationally demanding and particularly during the extensive evaluation of an increasing number of clusters (e.g., 2 to 30 or 50 clusters). To overcome this limitation, we used the LOBPCG method to extract fast estimations of the eigenvectors and eigenvalues by solving the minimum trace problem, rather than using the conventional ARPACK (ARnoldi PACKage) [420] and AMG (algebraic multigrid) [421] methods which are computationally demanding. Regarding the optimal component estimation process, the DBS was estimated for each clustering solution and the one with the highest DBS was extracted as the final one. To further reduce the complexity of the clustering evaluation process, we store the local maxima of the DBS and if there are no reported maxima after 5 clusters under evaluation, the process is terminated thus avoiding additional unnecessary clustering evaluations.

The cluster with the highest DBS is then extracted to define the number of Gaussian components in the BGMM training stage. In addition, the gamma parameter was exponentially related (non-linear) with the number of components, rather than inverse related (linear), to avoid linear assumptions. The BGMM-OCE places particular emphasis on the quality of the input data since lack of data quality makes data useless and reduces the statistical power of the outcomes. Thus, the quality of the real data is reflected on the virtually generated data. In this work, we extend a fully automated data curation pipeline presented in [66] to avoid data contamination by separating the features in the input space into two states; the “eligible” and the “non eligible”, based on the detected outliers and missing values. This separation provides a comprehensive view into the quality of the input data. Advanced outlier detection methods like the Isolation Forests were used to identify outliers with more than 80% accuracy and string-matching methods were applied to detect duplicated features. Imputation based on the kNN method was applied only on the “eligible” features, where applicable.

According to Table 63, the MVND and the log-MVND algorithms are fast and require only the mean vector and the covariance matrix as input but they are based on critical assumptions (e.g., normality). In the case of the Bayesian networks, although they provide explainable presentations of the conditional probabilities through the network, the number of all possible permutations of edges in one topology is infinite, the quality

of the virtual data is reduced, and the computational complexity is large. On the other hand, the STE and UTE yield virtual data with better quality, but they still have increased computational complexity for training and testing. In addition, they require a target feature which influences the associations of the virtual features and introduces biases in the generated data. The same stands for the ANN but it has reduced computational complexity. The UTE, STE, and ANN are unable to capture the inter- and intra- correlation differences like the GMM and the BGMM. As far as the GMM algorithm is concerned, it is more computationally efficient but requires multiple hyperparameters (number of Gaussian components, weight concentration parameter) which are arbitrarily defined and thus introduce biases.

However, the definition of the optimal number of Gaussian components and the estimation of the weight concentration parameter is a technical challenge. The BGMM-OCE overcomes this limitation by introducing a clustering stage based on the LOBPCG method prior to the BGMM training to estimate the optimal number of clusters as the one with the highest DBS across a set of predefined clusters. The best clustering solution is then set equal to the number of Gaussian components, and the weight concentration parameter is exponentially related to the number of Gaussian components instead of assuming linear dependencies.

Table 63. Comparison with the state-of-the-art virtual data generators.

Algorithm	**Study**	**Advantages**	**Issues**
MVND, log-MVDN	[164]–[166]	▪ fast execution time, ▪ requires only the mean vector and the covariance matrix of the real data.	▪ assumes that data are normally distributed, ▪ moderate virtual data quality (high GOF values).
Bayesian Networks	[166], [169], [171]	▪ explainable presentation of the conditional probabilities through the network,	▪ increased computational complexity, ▪ the conditional probabilities are modeled based on assumptions on the prior

Algorithm	Study	Advantages	Issues
		▪ accounts for conditional dependencies among the features.	distribution of the features (nodes), ▪ infinite number of topologies, ▪ bad virtual data quality (increased inter- and intra-correlation differences, KL-divergence, GOF, cV).
Supervised tree ensembles	[131], [173]	▪ robust training based on random forests, ▪ reduced GOF and KL-divergence.	▪ exponentially increasing computational complexity especially for large scale virtual population generation, ▪ moderate virtual data quality (increased intra- and inter-correlation differences, cV scores), ▪ requires a target feature, ▪ increased time for model validation.
Unsupervised tree ensembles	[131], [173]	▪ robust training based on density forests, ▪ reduced GOF and KL-divergence, ▪ resilient against outliers.	▪ exponentially increasing computational complexity especially for large scale virtual population generation, ▪ moderate virtual data quality (increased cV, inter- and intra-correlations), ▪ increased time for model validation.

Algorithm	Study	Advantages	Issues
Artificial Neural Networks with RBF-based kernels	[131], [172], [173]	▪ medium to high computational complexity, ▪ reduced KL-divergence.	▪ increased computational complexity for large scale virtual population generation, ▪ requires a target feature, ▪ moderate virtual data quality (increased cV, inter- and intra-correlation differences), ▪ increased time for model validation.
Gaussian Mixture Models	[174]	▪ good virtual data quality (reduced KL-divergence, inter- and intra- correlation differences), ▪ moderate execution time (maximizes the data likelihood), ▪ medium to high computational complexity.	▪ biases during the selection of the number of Gaussian components, ▪ increased GOF, ▪ biases in the weight concentration parameter, ▪ might yield specific structures that might or might not apply to the data maximizes the likelihood based on the expectation maximization (EM) approach.
Gaussian Mixture Models with variational inference	[130], [178]	▪ good virtual data quality (reduced KL-divergence, inter- and intra- correlation differences), ▪ fast execution time,	▪ increased GOF, ▪ biases from the arbitrary selection of the number of Gaussian components, ▪ biases from the arbitrary weight concentration parameter estimation.

Algorithm	Study	Advantages	Issues
		▪ uses variational inference (VI) which maximizes a lower bound on the model evidence to reduce the computational complexity.	
BGMM-OCE	[133]	▪ excellent data quality (reduced GOF, cV, KL-divergence, inter- and intra- correlation differences), ▪ optimal selection of the number of Gaussian components, ▪ robust estimation of the weight concentration parameter, ▪ fast execution time and stability across large scale virtual populations.	▪ non-significant variations in KL-divergence (but less than 0.05), ▪ moderate clustering evaluation execution time - especially for multiple clusters (can be leveraged using local maxima).

8.1.4. Federated/distributed learning

The problem of supervised learning towards predictive modeling for effective disease management in distributed environments was examined in the context of incremental learning by considering the problem of adjusting an initial data model on data that lie across multiple sites. The stochastic gradient descent approaches along with the gradient boosting schema yield an increased prediction performance for binary classification problems that are apt to the prediction of disease-oriented outcomes across distributed clinical data against conventional methods, such as, regression

models, neural networks, and Naïve Bayes. The proposed framework uses additive learning to incrementally develop binary prediction models across clinical data that are stored in private spaces in contrast to the meta-learning schema, where the classification algorithm runs individually on each site and the prediction outcomes of the separate classifiers are combined, a fact that yields inaccurate results due to the biases that are introduced by the integration stage. Moreover, the framework is cloud-based, a fact that overcomes the need to install any kind of software on premises and thus overcomes any potential threats for malicious attacks and privacy breach (in contrast to [202]–[205]).

The fact that the current incremental learning strategies are based on non-linear classifiers hampers the incorporation of more complex classifiers, such as, the convolutional neural networks, for deep learning, although the batch normalization technique has been recently proposed for dealing with the nonlinearities on each layer. Furthermore, the data on each site need to be harmonized otherwise the distributed learning schema will fail. In addition, the overall execution time for incremental learning is affected by the number of sites, due to the large number of communication links that need to be established between the central processing node and the sites (including the "handshaking" process for data access). This issue however can be reduced by the development of efficient network links within the platform.

The euroCAT platform [202]–[204] requires the installation of local servers on each hospital's premises, where the distributed learning algorithms include Bayesian networks and Support Vector Machines which were trained across 3 centers to predict dyspnea yielding modest prediction performance. In the PHT platform [203], a distributed logistic regression model was trained across 8 sites to predict post-treatment with adequate performance. In other studies, lymphoma classification models were trained across 4 pSS cohorts for federated lymphoma classification [129] and single cohorts were used to develop lymphoma classification models with reduced statistical power [244], [250].

Moreover, the existing federated learning frameworks like the Open Federated Learning (OpenFL) [205], the PySyft [201], the Open Federated Learning (OpenFL) [205], the FedML [200], the Paddle Federated Learning (PFL) [197], the TensorFlow Federated (TFF) is open source [199], the Flower [195] and the LEAF [194] are software frameworks which focus mainly on the installation of local servers on each

site and require the installation of a series of libraries on each premise. The proposed federated AI framework removes the need for the installation of any local servers or any type of software on premises through the development of a federated data management system that supports a large family of federated AI algorithms with small execution time complexity yielding robust and explainable AI models for lymphoma classification. In addition, the framework is cloud agnostic and thus can be adapted to any cloud infrastructure.

The proposed federated AI framework overcomes significant security threats that are posed by centralized analysis and deals with the development of accurate predictive models for disease management in distributed environments. The case study on pSS reveals promising results regarding the applicability of the framework towards the precise identification of pSS patients with lymphoma for effective patient stratification, where the class imbalance is large. Additional applications on multiple dispersed datasets are necessary to validate the applicability of the framework.

The federated AI framework can be used for the accurate risk prediction of lymphoma and thus contribute to the early lymphoma diagnosis in patients who have been diagnosed with pSS avoiding additional costs for biopsies. In addition, the AI model provides explainable scores which can be used by the clinician to assess the contribution of critical risk factors for lymphoma development and thus support the clinical decision-making process. The impaired 10-year survival of SS patients with MALT lymphomas and the association of lymphoma stage with the overall prognosis, point out the necessity for early lymphoma diagnosis and thus the development for lymphoma prediction models.

We also presented the FHBF algorithm to address critical overfitting effects during supervised learning tasks across heterogeneous clinical data with increased class imbalance in federated environments. To do so, a scaling parameter was first defined to adjust the shape of a hybrid loss function (based on the pre-defined dropout rate value) to avoid weight overfitting during the error reduction (boosting) process. Confound-based random downsampling with replacement was applied to yield 1:1 matched control and target populations in each federated database based on three confound factors. The downsampling process was repeated multiple times to avoid biases yielding an aggregated hybrid federated GBT (HFGBT) model on each iteration.

The HFGBTs from all downsampling iterations were then collected to formulate clusters of trees, where the weak clusters (i.e., those with log loss score less than the average) were discarded to enhance the classification performance by increasing its resilience against weak decisions. Explainability analysis was finally applied based on the SHAP approach to yield explainable outcomes. According to Table 64, the federated implementations of the conventional supervised learning algorithms which rely on SGD are easy to be implemented and deployed in federated environments, but they are prone to overfitting effects since they suffer by linearity assumptions (i.e., the data can be explained by linear relations of the features) and thus fail to capture complex data structures (i.e., the weights of the model tend to zero or infinity). The same stands for similar approaches like the logistic regression and the federated multi-layer perceptron which are based on linear functions.

Table 64. Comparison of the FHBF with federated implementations of existing supervised learning algorithms.

Algorithm	**Advantages**	**Weaknesses**
Federated SGD-based (e.g., Support Vector Machines, Logistic regression)	Easy to be implemented and deployed in federated environments, low computational complexity.	Poor classification performance, prone to overfitting during federated training.
Federated Multinomial Naïve Bayes (FMNB)	Easy to be implemented and deployed in federated environments, low computational complexity, immune against overfitting.	Probabilistic approach, biases are introduced in the results from several assumptions regarding the independence of the set of input features.
Federated Gradient Boosting Trees (FGBT)	Favorable classification performance due to the error reduction through boosting, easy deployment in federated environments, scalable, starts	Low to medium computational complexity especially with increasing number of boosting rounds.

Algorithm	Advantages	Weaknesses
	pruning trees backward based on the depth-first approach.	
Federated Gradient Boosting Trees with dropout rates (FDART)	Favorable classification performance due to the error reduction through boosting, easy deployment in federated environments, scalable, allows for the use of dropout rates which can significantly enhance the performance.	Low to medium computational complexity especially with increasing number of boosting stages, arbitrarily defined dropout rates can lead to overfitting and neglect the performance of the model.
Federated hybrid boosted forests (FHBF)	Increased classification performance, ideal in federated cases with increased class imbalance, allows for the use of dropout rates which can significantly enhance the performance, adjusted hybrid loss topology which avoids overfitting considering the dropout rate.	Medium to high computational complexity which can be affected by the number of iterations.

On the other hand, overfitting is less likely to occur in the federated multinomial Naïve Bayes (FMNB) algorithm since its hypothesis regarding the feature independence is strong. However, this assumption makes the FMNB more biased and less flexible and thus fails to capture complex data structures. Contrarily, more advanced algorithms like the federated gradient boosting tree methods have been proposed which combine sequentially connected weak tree learners (in the form of an ensemble) to create a strong learner, where each tree in the ensemble minimizes the prediction error of the previous tree. However, these methods tend to be biased since the trees that are added early in the ensemble have higher impact in the decision-making process than those added later in the ensemble.

The federated gradient boosting trees with dropouts solve this by introducing a dropout rate which accounts for an additional set "dropped" trees in the decision-making process. An arbitrary definition of this rate, however, can neglect the model's performance since a higher rate will include weak trees in the decision and thus lower the performance of the model by causing overfitting effects whereas a low rate might not yield any positive impact in the model's performance. To control for this effect, the FHBF utilizes a hybrid loss function with a scalable topology which can be adjusted according to the dropout rate to reduce overfitting.

In addition, the FHBF accounts for biases during the downsampling process by introducing a separation layer and a decision layer, where the former collects multiple instances of HFDARTs and drops instances with reduced score, whereas the decision layer includes only the "survivors" in the decision-making process. Although the performance of FHBF was higher in all cases, the execution time was higher which, however, can be leveraged by reducing the number of rounds. In each round, the computational complexity was similar to the FDART and FGBT algorithms.

The FHBF can be easily integrated in federated learning frameworks through a typical Python environment requiring no more than conventional libraries, such as, the 'numpy', 'scipy', 'xgboost', and 'pandas'. The fact that the algorithm was tested in a federated AI model deployment engine which was built under the open-source Nextcloud infrastructure supporting the WEBDAV (as defined in RFC 4918 by a working group of the Internet Engineering Task Force (IETF)) API [422] makes it compatible with Python and simplifies its integration to similar frameworks. To demonstrate the explainability and the clinical relevance of the model outcomes, we used the HFGBTs as learners in the FHBF algorithm, instead of HFDARTs, to overcome the fact that the SHAP package [401] does not support the 'dart' type of booster. In this case, the FHBF yielded explainable outcomes which are in line with similar findings in the literature [65], [224], [228], [233], [234], [238], [240], [241].

The FHBF is highly scalable since it can support both the FGBT and the FDART (using the xgboost 'gbtree' and 'dart' boosters) as base learners. In addition, the hybrid loss function can be used as an alternative to the existing loss functions that are used for binary classification tasks and are reported in the xgboost documentation [423], such as, the 'logistic', 'logitraw' and 'hinge' [423]. The selection of the hybrid loss function

is highly recommended to avoid overfitting effects that are introduced by the arbitrary definition of the dropouts. The dominance of the FHBF algorithm was demonstrated in two experimental phases involving the development of data intensive AI models for lymphoma classification across complex clinical data structures with increased class imbalance. In the second experimental phase, the class imbalance was even smaller to stress the performance of the algorithms.

To this end, and by taking into consideration the results from the two experimental phases, the FHBF algorithm showed the lowest average log loss distribution in the training and testing across all cases (Table 29, Figure 56, Figure 55), as well as, the best classification performance (Table 29, Figure 55), where the FDART with dropout rate 0.2 achieved the second-best performance in cases 1-8 (Table 29, Figure 55). On the other hand, in cases 1, 5 and 7 (Table 29, Figure 56, Figure 55), the FGBT and FDART had poor performance compared to the FHBF. Moreover, cases 5 and 7 (Table 29, Figure 56, Figure 55) demonstrated how the FHBF deals with overfitting effects against the FDART and FGBT implementations. In these cases, the existing state of the art implementations yielded specificity values close to 0.5 which is an indicator of biases during the weight update process. Through its straightforward decision layer, the FHBF was able to eliminate these "bad" clusters with the biased trees and thus prevented them from neglecting the model's performance. Although the federated AI model deployment time of the FHBF algorithm was larger than the FGBT and the FDART schemas, without however any significant differences, the number of rounds can be reduced to leverage the computational complexity.

8.2. Clinical impact

8.2.1. AD (pSS)

The data evaluation module was able to capture a first look into the dataset's structure and vocabulary. The data quality control module was able to identify outliers and missing values, as well as, detect fields with similar context and duplicated terms. The ability to choose among different outlier detection methods (z-scores, Grubb's test, IQR, LOF) increases the statistical power of the outcomes. The clinicians successfully validated the accuracy of the problematic fields which were correctly identified in both cohorts. In addition, the detected outliers helped the clinicians fix several discrepancies

or remove them where necessary, with a validity index of more than 90%. A large portion of the outliers in both cohorts were detected in features with binary values.

In most of the cases, binary outliers do not have any clinical importance (e.g., in "Gender" the majority of values is zero which denotes females so values with one (males) are highlighted). The largest portion of the missing values was physically explained by the data providers. Most of them had to do with follow-up calculations or records that have been lost in the past. Undoubtedly, the data evaluation report combined with the curated dataset have been proven useful for the clinicians during the data quality assessment process, reducing the time effort needed for manual data curation.

The data standardization module was able to identify and properly classify more than 85% of pSS-related terms for the UoA and HUA cohorts, based on knowledge from the reference model. This highlights the importance of the reference model which stands as a gold standard for matching similar terminologies across heterogeneous data and thus enables data harmonization. However, the percentage of matching terms can be greatly enhanced if the data standardization module receives as input the semantic representation of the raw dataset instead of (only) the clinical one. In addition, a semantic representation of the raw dataset can reduce information loss. An example of how an ontology can reduce information loss and improve the overall matching percentage can be seen in the HUA cohort.

The HUA cohort includes nine variables which are not stated in the reference model and are related to the various therapeutic prescriptions, such as, Methotrexate (MTX), Leflunomide, Cyslosporine, Azathiprine, Hydroxychloroquine (HCQ), Mycophenolate mofetil (MMF), Anti-TNFs, Rituximab (RTX), and Belimumab. These variables could be grouped into the class "Therapies" and then the semantic matching process would be able to match this class with the homonymous class of the reference model and thus reduce the information loss by 4% with an additional increase in the matching performance by 2%. As a matter of fact, the vocabulary could be enriched by adding the detailed (sub-)symptomatology related to the different ESSDAI domains so as to increase the matching percentage, as well as, include medical acronyms related to popular laboratory tests, such as, the "HBsAg" which stands for Hepatitis B, the "WBC" which stands for white blood cells, etc.

Sjögren's syndrome exists as a field with DOID: 12894 in the generalized disease ontology [424]. More specifically, it is a special case of the hypersensitivity reaction type II immune system disease and is registered with three exact synonyms, i.e., xerodermosteosis, Sicca syndrome, Sjögren syndrome. However, (a) no related domains exist and (b) there is no discrimination between primary and secondary Sjögren's syndrome. The current ontology is not based on the classic BFO (Basic Formal Ontology) type which is a high-level generic disease ontology type but rather is a disease-specific one, aiming to cover the knowledge domain of the pSS. Hence, it is a low-level, context-specific, biomedical ontology focusing on the domains of the syndrome. The proposed pSS reference model defines a set of minimum criteria necessary at diagnosis/follow-up including pathologic ocular involvement disease indicators, such as, Schrimer's, as well as several laboratory measures, such as, leukopenia, cryoglobulinemia, lip or parotid biopsy, ESSPRI and ESSDAI scores, etc., for improving data inclusion and quality. These will be further combined with existing reports of missing information (e.g., SNPs, etc.) from clinical partners, also by taking into consideration useful recommendations from the European League Against Rheumatism (EULAR) [397].

The presented ontology is the first biomedical ontology that covers a large portion of the pSS domain knowledge. It is primarily based on a pre-defined reference model (i.e., a set of variables and descriptions) that fulfills all the necessary requirements for the definition of a complete schema which is then evolved into an ontology. The ensuing ontology is an hierarchical model which consists of properly defined classes, data and object properties that are organized in a simple manner. According to the defined ontology, a pSS patient has demographic data, various laboratory measures, ESSDAI scores, and therapies, which are treated as subclasses of the main class patient. Each subclass includes further data properties (e.g., the laboratory tests include oral, ocular, OSS tests, etc.). This hierarchy is easier to follow and better to comprehend since its' context is more data driven (i.e., a low-level ontology), instead of the high-level BFO type. To this end, the pSS ontology has been published in the form of an .owl file on the following link: https://github.com/vpz4/PSS-Ontology.

The main reason behind the development of a general pSS ontology is to enable the mapping of existing pSS ontologies into a common ontology that will enable the

analysis of big heterogeneous medical data. As a matter of fact, future data harmonization and data integration procedures are going to be applied on such heterogeneous datasets, through ontology mapping (i.e., schematic interlinking) mechanisms. Therefore, the development of a common pSS ontology is a crucial step prior to the implementation of these mechanisms, in order to apply harmonization, integration and federated analysis on cohort data. Meanwhile, ontology mapping is a complex field since it does not only involve the transformation of one schema into a common one but also needs to comply with crucial data sharing regulations (e.g., the General Data Protection Regulation in Europe), especially when the involved ontologies originate from medical cohorts across different countries all over the world.

Our results reveal the favorable performance of the unsupervised tree ensembles for virtual population generation which outperformed the rest of the virtual population generation methods having the smallest goodness of fit and Kullback-Leibler divergence values in both experimental case studies (Section 7.1.3.1). The histograms of the virtual data that were generated by the unsupervised tree ensembles can be found in Section 7.1.3.1 for the pSS dataset. In both cases, the histograms reflect a highly qualitative similarity between the real and the virtual distributions. The supervised tree ensembles had the second-best performance (Section 7.1.3.1). The results from the supervised RBF-based artificial neural networks (ANNs) are close to the two previous methods, with the Bayesian networks and the log-MVND trailing behind. The dominance of the tree ensembles as a method for generating virtual data with increased level of agreement with the real data is in line with a recent study [131] which focuses on the generation of virtual data for in-silico cardiomyopathies drug development.

Our results also highlight the positive impact of augmenting the real with the virtual patient data which were generated by the "unsupervised" tree ensembles through data augmentation towards the development of robust disease classification and risk stratification models. The XGBoost algorithm was selected as a state-of-the art tree ensemble approach the value of which was demonstrated in previous studies [129], [250] for lymphoma classification in pSS. The performance of the lymphoma classification model in the pSS domain showed an increase by 10.9% in the classification accuracy, 10.7% in sensitivity, 11.5% in specificity, and 12.2% in area under a curve for lymphoma classification (Table 21, Figure 43) against the one trained

only on the real data. A similar increase is also observed in the case of the AdaBoost (7.1% in accuracy, 5.7% in sensitivity, 10% in specificity, and 6.5% in AUC), as well as, in the case of the Random Forests (9.5% in the accuracy, 8.9% sensitivity, 10.8% in specificity, and 11.2% in AUC).

In this thesis, we extended: (i) a previous study [250] by applying federated learning to develop disease prediction models across cohort data that are stored in private cloud databases instead of mining knowledge from an integrative database, and (ii) a second study [135] by recruiting two additional European cohorts on pSS (for the first time) towards the development of a preliminary federated lymphomagenesis progression model, and the discovery of prominent factors for lymphomagenesis progression. Our results reveal a 90% average overlap among the cohort data and confirm the dominance of the federated XGBoost schema for predicting lymphomagenesis with accuracy 0.848, sensitivity 0.833, specificity 0.849, area under the curve 0.868 along with the identification of lymphadenopathy, salivary gland enlargement, C4, and age at SS diagnosis as prominent factors for lymphomagenesis progression.

These risk factors are in line with previous findings [228], [233], [234], [240], which confirm the importance of C4, salivary gland enlargement and lymphadenopathy for lymphoma progression. Furthermore, the data curator was able to automatically identify inconsistent fields and provide adequate reports regarding the conformity and completeness of each feature, a fact that assisted the clinical experts during the inspection of the curated data and the data quality evaluation reports. Moreover, the combination of lexical matching with ontology matching was able to detect overlapping terminologies among the curated cohort data based on the FHIR-compliant, pSS reference ontology, overcoming significant computational barriers that are posed during the co-analysis of the harmonized cohort data and thus enhancing the statistical power of the clinical findings.

The biomarkers for lymphoma development include parotid or submandibular swelling, cryoglobulinemia, rheumatoid factor, and low C4 levels, among others, which have been validated in previous studies [238], [425], [426] highlighting the significance of parotid or submandibular gland swelling, low C4, rheumatoid factor and cryoglobulinemia for lymphoma development. In [425], [426] salivary gland swelling and cryoglobulinemia appear to be significantly higher in pSS patients evolving into

lymphoma compared to pSS controls. In fact, cryoglobulinemia can affect many extraglandular organs, such as, the kidney, the skin, and the peripheral nerves, leading to permanent damage. The impact of age of SS diagnosis was also highlighted as a prominent factor in [233], [234], where the time interval from pSS diagnosis to lymphoma has been stated as a biomarker for lymphoma prediction. Furthermore, patients with the presence of parotid or submandibular swelling, rheumatoid factor (RF), cryoglobulinemia, and low C4 tend to have higher impact for lymphoma development. This can be confirmed by the distribution of the samples in Figure 51 and Figure 52 which shift the ground truth to the right direction and thus have a positive predictive value for lymphoma development.

8.2.2. HCM

In this thesis, we examined the effectiveness of data augmentation in terms of enhancing the real clinical research databases with high-quality virtual data to enhance the performance of the HCM risk stratification models. To do so, a computational pipeline was developed, where high-quality virtual data are aggregated with the real data to yield robust lymphoma classification models, where the performance of each model was evaluated on testing instances of the real data to avoid any biases. The proposed pipeline was able to generate virtual distributions with increased similarity, correlation, and reduced divergence with the real distributions. The aggregation of the real with the virtual patient data yielded a notable increase in the classification accuracy, sensitivity, specificity, and area under the curve scores of the supervised machine learning models which were trained on the augmented clinical data compared to those trained on real data instances. The proposed methods could potentially provide significant insight in the field of virtual population generation to re-adjust the perspective of Clinical Trials (CTs) in other domains.

8.2.3. CVD

The presented CVD ontology is a first, hierarchical data model that covers a large portion of the CVD clinical domain knowledge. The ontology is based on a pre-defined reference model which was developed under the aegis of the TO_AITION project by harnessing knowledge from the Tampere University Hospital. The proposed CVD ontology fulfills all the necessary requirements for the CVD domain knowledge. This

hierarchy is easier to follow and better to comprehend since its' context is more data driven. The hybrid data harmonization process was able to utilize the CVD ontology as a reference ontology to enrich the current medical corpus and harmonize terminologies across two large European clinical centers. The proposed ontology was published in the following github repository to promote research in CVD: https://github.com/vpz4/TO_AITION/blob/main/urn_webprotege_ontology_3b32db46-2123-4003-ba85-2923003cfd0c.owl.

8.2.4. MD

The MD ontology is a first, hierarchical data model that covers a large portion of the MD clinical domain knowledge, which is a rather unmapped domain. The ontology is based on a pre-defined reference model which was developed under the aegis of the TO_AITION project by harnessing knowledge from the University of Amsterdam. The proposed MD ontology fulfills all the necessary requirements for the MD domain knowledge. This hierarchy is easier to follow and better to comprehend since its' context is more data driven. The hybrid data harmonization process was able to utilize the MD ontology as a reference ontology to enrich the current medical corpus and harmonize terminologies across two large European clinical centers. In addition, the hybrid data harmonization process was able to identify cross-matches among the CVD and MD domains since the MD domain is a more complex domain compared to the CVD and the majority of the MD-related terminologies in the two European centers were highly associated with CVD-oriented terminologies. The proposed ontology was published in the following github repository to further promote research in MD and better understand the underlying clinical associations and phenotypes among CVD and MD:https://github.com/vpz4/TO_AITION/blob/main/urn_webprotege_ontology_692fe63d-ca65-47e8-8a51-6ff347eaea3a.owl.

8.2.5. SAIDs (Kawasaki)

Our results reveal five prominent genes for KD diagnosis which are proposed for the first time, namely the HLA-DQB1, HLA-DRA, ZBTB48, TNFRSF13C, and CASD1. The KD classifiers which were trained on the proposed genes yielded better performance against those trained on the known ones, in terms of increased accuracy, sensitivity, specificity, and AUC. In the common platform analysis, the sample size in

GPL6271 was considered as adequate for the application of the proposed computational workflow due to the significant lack of available KD patients, in terms of time-series expression profiling. To further test the discrimination performance of the proposed set of diagnostic biomarkers across other types of similar diseases, a cross-platform analysis was also conducted through the transformation and subsequent integration of six datasets from two different platforms (GPL570, GPL10558). The integrated dataset included 1,347 patient samples, where the non-KD group included patients with SJIA and SLE, which are characterized by certain clinical similarities with the KD patients. To our knowledge, this is the first data-driven workflow which constructs SOMs on the three clinical phases of KD based on time-series gene expression data towards the discovery of five candidate diagnostic biomarkers for KD with increased discrimination performance against other analogous diseases. The Self-Organizing Maps were constructed in a straightforward way to enable the clustering of the KD patients across the three clinical phases of KD, in a two-stage manner; the inter-phase and the intra-phase clustering. The two-stage clustering process yielded homogeneous and concise clusters of patients which were subsequently merged to identify four super-clusters. The derived super-clusters were able to categorize the available KD patients into four subgroups with similar genetic profiles across the whole duration of the disease and not on a single clinical phase to better comprehend the mechanisms of KD onset. The super-clusters were utilized, in a data-driven way, to extract the most prominent genes through FDR-based feature selection yielding statistically significant genes for KD diagnosis.

Both the boosting classifiers highlighted the impact of the proposed genes against the known KD genes, specifically in the Acute and Subacute phases, yielding an average increase by 4.40% in the accuracy, 5.52% in sensitivity, 3.57% in specificity, and 2.85% in the AUC. The performance of the AdaBoost on the proposed set of genes is significantly higher in all clinical phases of Kawasaki compared against the known set of genes. This increase, however, is not observed in the Convalescent phase for the GBT schema. These imply that the proposed set of genes can be used to shed light into the underlying pathogenic mechanisms and genetic basis of the KD onset with favorable precision in the first two phases of the disease. On the other hand, the known KD genes can be used to understand the evolvement of KD in the second clinical phase, where the patients already start to exhibit clinical manifestations and thus the pathophysiology is already observed.

Regarding the cross-platform analysis, the boosting classifiers yielded an average increase by 2.30% in the accuracy, 2.20% in sensitivity, 4.70% specificity, and 1.70% in AUC, across the two boosting classifiers. This suggests that the proposed diagnostic biomarkers for KD present a notable discrimination performance of KD patients even in cases where the control group consists of patients that exhibit clinical similarities with KD. Finally, in both types of analyses, the gene expression data in the acute phase contribute most to KD prediction than those in the sub-acute and convalescent phases (Table 54, Table 56) which is in line with the fact that early identification and timely IVIG (intravenous immunoglobulin) treatment is the best policy to treat KD.

The potential relation of the proposed genes with KD according to previous works reported in the literature is presented in Table 65. Specifically, for the HLA class II genes, like HLA-DQB1 and HLA-DRA, certain Single Nucleotide Polymorphisms have been associated with KD diagnosis in Genome Wide Association Studies (GWAS) reports [45]. Moreover, zinc finger proteins, like the ZBTB48, have been found to be down-regulated in KD patients [46], while increased TNFRSF13C gene expression has been associated with induced inflammation in RAW 264.7 cells [47]. Finally, several studies have indicated the role of CASD1 in the immune system [48-50]. These five genes are reported as biomarkers for KD diagnosis for the first time in the literature using data-driven analysis instead of the conventional laboratory analysis.

Table 65. Relation of the proposed set of genes with KD studies in the literature.

ID_REF	GB_LIST	Gene ID	Description
15658	AI431505	HLA-DQB1	Association of the SNPs in HLA class II genes were documented as susceptibility genes of KD in GWAS reports [427]
15660	AI434629	HLA-DRA	
22055	AA810410	ZBTB48	Zinc finger protein 124 (circZNF124) has been found to be significantly down-regulated in untreated patients with Kawasaki disease [428]

ID_REF	GB_LIST	Gene ID	Description
26049	AA864899	TNFRSF13C	TNFRSF13C is a target gene of miR-122 in RAW 264.7 cells' inflammatory responses [429]
35359	AI250844	CASD1	The role of CAS1 protein has been associated with the immune system in various works [430]–[432]

8.2.6. COVID-19

In this work, we developed a multimodal data analytics pipeline which utilizes explainable and interpretable AI models along with dynamic modeling methods on curated clinical data to understand the pathogenesis and risk factors of COVID-19 regarding ICU admission and mortality. The extracted risk factors for ICU admission and/or mortality were combined with the APACHE-II score, which has been reported in [433], [434] as one of the most contributory variables for the risk prediction of COVID-19, to develop an ICU scoring index with accuracy 0.9 based on IL-6, IL-8, IL-1b and TNF and thus quantify the severity of the disease. Our results highlight the importance of LDH, age, CRP, WBC, IL-6, IL-8, Cr, number of monocytes, lymphocyte count, and TNF as risk predictors for ICU admission (and survival) and mortality in the ICU, among others. A similar picture is observed in the case of time intervals INT1-INT2, INT1-INT3, and INT1-INT4 with updates in the ranking order.

The proposed method focuses on the detection of explainable risk predictors for ICU admission and/or mortality, as well as, to the identification of an ICU scoring index which complements the APACHE-II score compared to the workflows that are presented in [435], [436] and focus only on the extraction of risk factors. Furthermore, the proposed approach avoids the application of conventional multivariate regression analysis like in [437]–[441] since these types of methods are based on statistical assumptions regarding the independence of the input factors and thus reduce the statistical power of the outcomes. In addition, the AI model validation process is not based on random splits of the data as in [441], [442] nor on the application of bagging

methods as in [435], [442], [443] which introduce biases during the assembly stage and the performance evaluation of the AI models.

The classification performance of the AI model in Group A regarding the patients who were not admitted in the ICU and survived was favorable, where the classification performance in time intervals INT3 and INT4 was less than INT1 since patients in Group A did not remain in the hospital for many days and thus most of the cytokine measures in the future time intervals were missing. Regarding Group C, the performance of the AI model in INT2 was lower than INT1 due to the higher percentage of missing cytokines in INT2. The same occurred in INT4 when compared against INT3. The AI model in Group D was not affected by the missing cytokine measurements, as in the previous groups, since the number of patients who were submitted in the ICU and died was small and easily separable thus the impact of the missing cytokines in future time intervals was irrelevant in this case.

To further highlight the prediction performance of the proposed AI model we compared it against four other machine learning schemas, including the Logistic Regression (LR), the Support Vector Machines (SVM), the AdaBoost and the Naïve Bayes (NB). The prediction performance results are summarized in Table 59. According to Table 59, the GBT had the best performance in all cases and for all groups under investigation.

Table 66. Comparison with existing state-of-the-art studies.

Study	Dataset	Method	Outcomes
[437]	Electronic medical records with symptoms, signs, and laboratory findings from 244 hospitalized COVID-19 patients in China.	Multivariate logistic regression analysis was used to identify risk factors for mortality.	Risk factors: Disease severity, gender, white blood cell count and age as risk factors, C reactive protein (CRP).

Study	Dataset	Method	Outcomes
[438]	Clinical data from 663 COVID-19 patients in China.	Multivariate logistic regression analysis to model the disease severity.	Risk factors: Sex, disease severity, expectoration, muscle ache, and decreased albumin.
[439]	Clinical data from 3,894 COVID-19 patients in Italy.	Machine learning (random forest) and Cox survival analysis were used to identify risk factors for mortality in the hospital.	Risk factors: Impaired renal function, elevated C-reactive protein, and advanced age.
[440]	Medical records from 4,404 COVID-19 patients in China.	Exploratory multivariate analysis was applied to identify predictors of ICU care and mechanical ventilation.	Lower oxygen saturations were associated with need for ICU and invasive mechanical ventilation, and with death. High respiratory rates were associated with the need for ICU care.
[441]	Medical records from 4,997 COVID-19 patients in the U.S.	Multivariate logistic regression was applied to predict ICU admission and death.	Risk factors for ICU admission: lactate dehydrogenase, procalcitonin, pulse oxygen saturation, smoking history, lymphocyte count. Risk factors for mortality: heart failure, procalcitonin, lactate dehydrogenase, chronic obstructive pulmonary disease,

Study	Dataset	Method	Outcomes
			pulse oxygen saturation, heart rate, age. The risk score model yielded good accuracy with an AUC of 0.74 for predicting ICU admission and 0.83 for predicting mortality with a train test split method.
[442]	1,270 COVID-19 patients in the U.S.	Multi-tree extreme gradient boosting (XGBoost) was used to detect prominent features for COVID-19 mortality.	Risk factors: disease severity, age, levels of high-sensitivity C-reactive protein (hs-CRP), lactate dehydrogenase (LDH), ferritin, and interleukin-10 (IL-10). XGBoost model predicted death risk accurately with >0.9 precision and >0.85 sensitivity with a train test split method.
[443]	Clinical and laboratory data from 214 COVID-19 patients in China.	Random forest (RF) algorithm to differentiate severe and no severe COVID-19 clinical types based on multiple medical features and provide reliable predictions of the clinical type of the disease.	Risk factors: age, hypertension, cardiovascular disease, gender, diabetes, absolute neutrophil count, IL-6, LDH with 0.97 predictive accuracy using a train test split method.

Study	Dataset	Method	Outcomes
[444]	Clinical data from 162 hospitalized COVID-19 patients.	Artificial neural networks (ANNs) and bagging methods to predict the risk for critical COVID-19.	Risk factors: white blood cell count, time from symptoms to admission, oxygen saturation and blood lymphocytes count, APACHE II.
[435]	Clinical and laboratory data from 635 COVID-19 patients.	Multivariate and machine learning algorithms, such as, the decision trees, RF, GBT and ANNs were applied to predict risk factors for ICU admission and mortality.	Risk factors for mortality: age, procalcitonin, C-creative protein, lactate dehydrogenase, D-dimer, and lymphocytes. Risk factors for ICU admission: procalcitonin, lactate dehydrogenase, C-creative protein, pulse oxygen saturation, temperature, and ferritin.
[436]	Clinical data from 516 COVID-19 patients.	To produce models of mortality or criticality (mortality or ICU admission) in a development cohort using machine learning algorithms (e.g., XGBoost, Random Forests).	Risk factors for mortality: Age, diastolic pressure, O2 Sat, BMI, AST, creatinine, CRP, ferritin, platelet, RDW, WBC. Risk factors for criticality: Age, O2 Sat, ALT, AST, creatinine, CRP, ferritin, platelet, RDW, WBC, neutrophil/lymphocyte ratio. Prediction of mortality with 0.89 AUC and ICU admission with 0.79 AUC.
[416]	Clinical and biological data across four time points from 324	A multimodal data analytics pipeline which utilizes explainable and interpretable AI	Risk factors for mortality: LDH, IL-6, IL-8, Cr, number of monocytes, lymphocyte count, and TNF.

Study	Dataset	Method	Outcomes
	COVID-19 patients	models along with dynamic modeling methods to identify risk factors of COVID-19 regarding ICU admission and mortality and develop an ICU scoring index.	Risk factors for ICU admission and survival: LDH, age, CRP, Cr, WBC, lymphocyte count for mortality in the ICU, with prediction accuracy 0.79 and 0.81, respectively. These risk factors were combined with dynamically associated biological markers to develop an ICU scoring index with accuracy 0.9.

Regarding the findings of the explainability analysis (Figure 86, Figure 87, Figure 88, Figure 89, Figure 90), the importance of LDH has been confirmed in [434], [445], [446], [447]–[449] as an independent risk factor for the severity and mortality of COVID-19. In addition, IL-6 has been linked to severity and duration of hospitalization in [450]. Furthermore, IL-6 has been identified in [451] as a disease severity predictor for COVID-19 and in [452] as a key factor, among numerous cytokines and chemokines, the treatment of which can reduce mortality in COVID-19 patients. The importance of IL-8 has been highlighted in [453] along with other circulating cytokines, including IP-10 (CXCL10), MCP1 (CCL2), and RANTES (CCL5). CRP levels have been positively associated with the severity of COVID-19 in [454], [455], [456], [457]–[459], [460], where the elevated levels of CRP and IL-6 have been proposed as predictors for mechanical ventilation in COVID-19 [457].

Age is a major predictor of mortality especially in older patients and has been considered as a key factor for the definition of various scoring systems for COVID-19 [458], [459]. The importance of IL-6 and IL-8 has been also stated in [460] in which the profiling of serum cytokines IL-6 and IL-8 have been identified as disease severity predictors for COVID-19. Increased cytokine levels, including TNF and IL-6 have been also reported in [450], [461] as risk factors for severity and mortality in COVID-19. Creatinine has been identified as an independent risk factor for predicting adverse outcomes in COVID-19 patients [462] but has been reported only on a few case studies

in the literature. The diagnostic and predictive role of the lymphocyte-to-monocyte ratio, the neutrophil-to-lymphocyte ratio, and the platelet-to-lymphocyte ratio in COVID-19 patients has been reported in [463].

The induced decision trees (Figure 91) confirm the importance of LDH and WBC counts in the decision-making process across Groups A, C, and D. Furthermore, the decision trees have highlighted the importance of CRP [464] and number of lymphocytes [463] as prominent factors for mortality in COVID-19. The clinical significance of the WBC morphology has been noted in [465] and its diagnostic and prognostic value in COVID-19 patients has been highlighted in [466]. The profiling of cytokines has also revealed IL-8 (apart from IL-6) as a disease severity predictor which is in line with the findings reported in [467]–[469]. Regarding the number of neutrophils (NEUT), the neutrophil-to-lymphocyte ratio has been found as an independent risk factor for mortality in hospitalized patients with COVID-19 [470]. The AST and ALT levels have been also associated with the mortality in COVID-19 patients [471]. In addition, the PLT count is related with the prediction of severe illness in COVID-19 [472].

Critical thresholds of the above risk predictors were identified by the induced decision trees using the baseline data and the cytokines from INT1 (Figure 91). More specifically, in Group A, the threshold 7.58 in WBC counts determines whether the decision will be based on CRP in case it is less than (or equal to) 7.58, where emphasis is given to Cr and PLT or LYM count, AST and age. Regarding Group C, the threshold 278 in the LDH determines whether the decision will be based on IL-6, and the number of monocytes in case it is less than (or equal to) 278 or on LYM, IL-6 and WBC, otherwise. As for Group D, the threshold 357 in LDH indicates a critical value which determines whether the decision will be based on CRP and Cr in case it is lower than (or equal to) 257, or on age, Cr and PLT count otherwise.

To capture an overall picture of the prominent risk factors for ICU admission and mortality a multiclass problem was also investigated using a Random Forests classifier which was trained on the 214 patients to solve a four-class classification problem, where class “0” denotes the patients who belong in Group A, class “1” those in Group B, class “2” those in Group C, and class “3” those in Group D. Feature ranking was measured based on the Gini index across the total number of instances. The obtained

classification performance was 0.71. The set of prominent features include the CRP, WBC, LDH, IL-6 in INT2 and INT3, AST and age, which are in line with the findings in the subgroup analysis, along with NEUT (number of neutrophils) and ALT (Alanine transaminase) which appear to be important, as well.

The ICU scoring index analysis which was conducted in Groups C and D using only the cytokines that were identified as important from the DBN analysis and the explainability analysis, yielded significant risk predictors for ICU admission and mortality. More specifically, the proposed method was able to recursively identify the APACHE-II, IL-1b, and IL-8 as those contributing most to the classification accuracy in Group C and for INT1, which suggests that these features can be used as disease severity predictors for ICU scoring and thus can determine the admission of hospitalized patients with COVID-19 in the ICU. In addition, the APACHE-II, IL-1b, and IL-6 were also highlighted as risk predictors for ICU admission and mortality in Group D. Regarding the rest of the time interval combinations in Group C, it is interesting to note that cytokines measured in previous time intervals continue to remain prominent in future time intervals as well (e.g., IL-6 from INT1 remains important in INT1-INT2, IL-6 from INT2 preserves its importance in INT1-INT3, and TNF from INT2 in INT1-INT4). As for Group D, a similar pattern is observed, where IL-6 from INT1 remains important even when the cytokines from INT1-INT2 and INT1-INT3 are used, TNF from INT1 remains important when using INT1-INT3, and IL-8 from INT2 remains prominent in the analysis even when using INT1-INT4 along with IL-1b measured in INT3.

Altogether, our results from the baseline data and cytokines from INT1 highlight the importance of LDH, IL-6, IL-8, Cr, number of monocytes, lymphocyte count, and TNF as risk predictors for ICU admission and survival, as well as, LDH, age, CRP, Cr, WBC, and lymphocyte count as risk predictors for mortality after ICU admission, among others. Based on DBN modeling the prediction of probable and reasonable trajectories was provided over time, considering the measurement of the four cytokines in discrete time points. Moreover, the model revealed the probabilistic relationships among risk factors of COVID-19 regarding ICU admission and mortality. For instance, we found that IL-6 influences the levels of TNF in the last time point and more dependencies were evidenced over time between TNF and IL-8. The most important features from

the DBN analysis were finally combined with the risk predictors from Shapley explanation analysis to extend the clinical impact of the APACHE-II score towards the development of a scoring index based on IL-6, IL-8, IL-1b and TNF during the admission of hospitalized COVID-19 patients in the ICU across different time intervals of the disease.

In addition, we presented a straightforward workflow which uses SOMs to derive homogeneous clusters of patients with COVID-19 based on a subset of features that have the highest degree and connectivity across multiple timepoints. The clustering labels from the SOMs were used to enrich the existing time-series clinical and laboratory data with meta information yielding an increase in the performance of classification models for ICU admission and mortality. Our results highlight the contribution of the extracted patient subgroups from the SOMs along with the dynamically associated features with increased connectivity from the DBNs towards the improvement of the classification performance for ICU admission (sensitivity 0.83; specificity 0.83) and mortality (sensitivity 0.74; specificity 0.76). The number of lymphocytes, SatO2, PO2/FiO2, and O2 supply type were highlighted as prominent risk factors for ICU admission and the percentage of neutrophils and lymphocytes, PO2/FiO2, LDH, and ALP for mortality, among others.

Significant differences were identified in the patient distribution across the four super-clusters from the SOMs analysis and particularly for the features “Hct”, “Lymph_abs_number”, “Lymph_percent”, “Neut_abs_number”, “Neut_percent” and “PO2_FiO2_ratio”, regarding ICU admission and mortality. Additional significant differences were detected in “AST” for ICU admission and in “ALP” and “LDH” for mortality. The most important features were utilized in the SOMs to extract homogeneous clusters of COVID-19 patients with common clinical profiles. Subsequently, MARTs were trained on the aggregated features from the DBNs and the new features from the SOMs yielding robust classifiers for ICU admission and mortality with an increase by 1% in sensitivity and 2% in specificity for ICU admission in case study 1, as well as an increase by 4% in sensitivity and specificity for ICU admission and by 3% in sensitivity and 2% in specificity in case study 2, compared to the classifiers trained with the clustering labels from the SOMs. The contribution of demographics-related data yielded an increase in accuracy by 4% and in AUC by 6%

for mortality (Table 62) which suggests that age has a high impact on mortality in hospitalized COVID-19 patients.

The number of lymphocytes, SatO2, PO2/FiO2 and O2 supply type were highlighted as major risk factors for ICU admission and the percentage of neutrophils and lymphocytes, PO2/FiO2, LDH, and ALP for mortality. According to Table 4, our findings are in line with related risk factors that were reported in the literature. More specifically, the neutrophils infiltration has been found to drive necroinflammation during coronavirus in [473] whereas in [474] the neutrophil to lymphocyte ratio has been highlighted as a risk factor for the severity of COVID-19. Additional risk factors for mortality include the "Hb" which has been highlighted also in [475] as an independent risk factor for the mortality in COVID-19 patients and the "INR" which has been linked with COVID-19 severity in [476].

The prognostic value of troponin elevation has been identified in [477] and particularly in patients with underlying cardiovascular diseases. The "PO2_FiO2_ratio" along with the "FiO2" have been identified as independent risk factors for in-hospital mortality in patients with COVID-19 [478]. Likewise, LDH has been found as an independent risk factor of severe COVID-19 in [434] while tachypnea and low SBP have been strongly associated with in-hospital mortality in COVID-19 [479]. Additional risk factors for ICU admission include the supply oxygen type which is highly associated with COVID-19 severity, and SatO2 which serves as a predictor of mortality in adult patients with COVID-19 [480]. The relationship between mortality and ALP has also been demonstrated in [481], [482] which underline the clinical need for further investigation of elevated serum alkaline phosphatase levels as a mechanism of liver injury in COVID-19. In addition, this study goes beyond the state of the art by combining DBNs with SOMs and trajectories to derive homogeneous clusters of patients with COVID-19 based on a subset of features that have the highest degree and connectivity across multiple timepoints.

Unlike the existing studies (Table 67) which focus on the direct application of machine learning algorithms for the development of ICU admission and mortality classifiers and the detection of related risk factors, the proposed approach places particular emphasis on the dynamic modeling of features across multiple time-points to extract the most informative ones. The latter are utilized to derive homogeneous clusters of COVID-19

patients with similar clinical profiles based on the SOMs and the trajectory analysis. The extracted clustering information is then combined with the input data to enhance the robustness of the classifiers for ICU admission and mortality.

Table 67. Comparison with the state-of-the-art studies for ICU admission and mortality in COVID-19.

Study	Method	Risk factors
[483]	Ensemble-based algorithms to predict ICU admission and mortality across 3597 COVID-19 patients.	Risk factors: CRP, LDH, O2 saturation for ICU admission and neutrophil and lymphocytes for mortality.
[484]	Random forests for risk stratification based on time-series data across 1987 unique patients diagnosed with COVID-19.	A risk prioritization tool that predicts the need for ICU admission within 24h to optimize the flow of operations within the hospitals.
[485]	Ensemble learning to objectively identify an optimal combination of factors that predicts ICU admissions across 733 COVID-19 patients.	The number of lymphocytes was involved in all prediction tasks with the highest AUC score.
[486]	Multipurpose algorithms (boosting ensembles, artificial neural networks) to estimate the risk of ICU admission or mortality among 3623 patients with COVID-19.	The final model achieved good discrimination for the external validation set (AUC 0.821). A cut-off of 0.4 yields sensitivity and specificity 0.71 and 0.78, respectively.
[487]	Predict the risk for COVID-19 severity by training multipurpose algorithms across 3280 patients.	High predictive performance (average ROC 0.92) with the following risk factors: lymphocytes, C-reactive protein, and Braden Scale.

Study	Method	Risk factors
[488]	GBTs were trained on 1270 COVID-19 patients from Wuhan to detect risk factors.	Age, CRP, and LDH were identified as prominent features for COVID-19 mortality.
[443]	Bagging methods were applied on clinical data from 362 patients with confirmed COVID-19.	Age, hypertension, gender, diabetes, absolute neutrophil count, IL-6, and LDH were identified as risk factors for COVID-19 severity.
[417]	DBNs combined with SOMs to derive homogeneous clusters of patients with COVID-19 which were used to enrich the existing time-series clinical and laboratory data with meta information to increase the performance of classification models for ICU admission and mortality.	Risk factors: number of lymphocytes, SatO2, PO2/FiO2, and O2 supply type as risk factors for ICU admission and the percentage of neutrophils and lymphocytes, PO2/FiO2, LDH, and ALP for mortality. Classification performance for ICU admission with sensitivity: 0.83 and specificity: 0.83 (AUC 0.91), and mortality with sensitivity: 0.74 and specificity: 0.76 (AUC 0.83).

Moreover, the scope of the proposed framework can be extended to include genome wide association studies (GWAS), where data curation methods can be used to enhance the quality of genomic data. Since the proposed framework is easily scalable, it can be updated to include functionalities for outlier detection, de-duplication, and imputation in genetic data, including DNA and RNA sequencing data, among others, where the existence of missing values, outliers, and similarities can lead to falsified associations between genetic variants (e.g., single-nucleotide polymorphisms (SNPs)) and traits (e.g., diseases). This is present in case-control studies, where common variants are examined between a case group (a group of individuals under a common disease or condition) and a control group (a group of healthy individuals). However, important genetic-related information needs to be added as a next step in the near future in order

to provide a much more detailed and accurate biomedical ontology with pSS genetic associations. Such an ontology not only will be able to completely describe the pSS domain knowledge but also lead to the creation of a much more effective schematic interlinking mechanism for conducting genetic analysis on integrated data across various genetic databases.

CHAPTER 9. CONCLUSIONS AND FUTURE WORK

9.1. Conclusions

The proposed automated framework for medical data curation is an integrated, web-based data quality assessment strategy which offers a fully automated REST API service that combines both univariate and multivariate methods for outlier detection, "smart" imputation, and de-duplication, as well as, terminology-based data standardization, yielding clinician-friendly data quality assessment reports that promote the use and re-use of high-quality data. The framework is easily scalable and can be incorporated into any medical platform that deals with big data analytics, as part of their data quality assessment strategy. For this reason, the source code of the data curator has been made publicly available under the following github repository: https://github.com/vpz4/Data-curator, along with a brief user manual. Additional applications on clinical databases are necessary to further evaluate the framework's efficacy and reliability, as well as, include more functionalities for outlier and similarity detection. Until now, the framework is executed in the form of a REST API service and efforts are needed to publish the service in the form of a user-friendly front-end web interface. The fact that the proposed framework introduces a reference model as a standard model for data standardization can be generalized for different types of diseases due to the scalability it offers and its increased FAIRification potential. The results of this framework can be also combined with semantic matching algorithms to enable data harmonization in different domains varying from autoimmune diseases to cardiovascular diseases and genomics. To this end, the data standardization report can be presented in the form of a drop-down menu, where the clinician will be able to select

the best match with the standard term(s) according to the Health Level-7 (HL7) standards, such as, the FHIR protocol. According to our findings from the case studies (CHAPTER 7), the proposed data curation framework enhanced the data quality in all clinical domains under investigation (pSS, HCM, CVD, MD, SAIDs, COVID-19) highlighting its increased clinical impact towards the effective quality control across complex clinical data structures.

The proposed hybrid data harmonization method was compared against the application of conventional lexical analysis and manual mapping (definition of pairing rules) across 8090 terminologies from the CVD domain and 998 terminologies from the mental disorders' domain, yielding a set of individual and cross domain matched terminologies with 85% precision (in average) and 10% higher performance than conventional lexical analysis. In addition, the proposed method is highly scalable and can be applied to any clinical domain given a reference ontology as input. The overall value of the proposed method lies on the fact that it can be used to deal with open issues and unmet needs in various clinical domains which enhances its scientific and clinical impact. The hybrid data harmonization method can be generalized to any clinical domain, as long as, a reference ontology is provided as input towards the construction of a medical corpus which will in turn enable the application of lexical and semantic matching algorithms. As a future work, we are planning to include more harmonized cohort data on the private cloud spaces to further enhance the statistical power of the lymphoma prediction models in the pSS domain, as well as, interlink the corpus with other medical index repositories and include deep learning algorithms in the distributed data analytics module to enhance the robustness of the disease prediction models. According to our findings from the case studies (CHAPTER 7), the proposed hybrid data harmonization workflow was able to diminish the underlying data heterogeneity across three diverse clinical domains (pSS, CVD, MD) highlighting its performance towards the accurate harmonization of complex clinical data structures across multiple clinical domains through its ability to be applied in a cross-domain manner.

The BGMM-OCE algorithm is an extension of the conventional BGMM which aims to address open issues regarding hyperparameter estimation in BGMM which is a crucial technical challenge. BGMM-OCE introduces a highly efficient spectral clustering stage based on the LOBPCG method to cluster patients with similar profiles within the input

data towards the estimation of the optimal number of Gaussian components. To do so, the algorithm estimates the DBS across a set of clusters under evaluation. To avoid computational burden during the clustering evaluation stage, we store the local maxima of the DBS and if there are no reported maxima after 5 clusters under evaluation, the process is terminated. The cluster with the highest DBS is then extracted to define the number of Gaussian components in the BGMM training stage and the gamma parameter was exponentially related (non-linear) with the number of components to avoid linear assumptions. The robustness of the BGMM-OCE was demonstrated through a large-scale study in the context of *in silico* clinical trials for HCM towards the generation of 30000 virtual patients. The proposed method yielded the lowest average inter- and intra-correlation differences and average coefficient of variation which suggests that it can capture hidden similarity patterns among the real and the virtual data with reduced dispersity. The outcomes of the proposed pipeline are promising since the existing lack of population size in HCM obscure the development of robust disease classification and risk stratification models.

In addition, the BGMM-OCE is highly sustainable since it can be applied in any clinical domain. As a future work, we plan to apply the BGMM-OCE across other clinical domains to populate medical databases with insufficient population size and make *in silico* clinical trials feasible, as well as, to test its accuracy for data imputation across highly complex clinical data structures and test its effectiveness in combining *in silico* and *in vivo* data into augmented clinical trials [489]. According to our findings from the case studies (CHAPTER 7), the proposed synthetic data generator was able generate high-quality synthetic data for *in silico* clinical trials in HCM, where the cost for drug testing is high and the population size is small, highlighting its reduced computational complexity and robustness towards large-scale, high-quality synthetic data generation.

Regarding data augmentation, we focused on the generation of high-quality synthetic data to enhance the performance of the conventional supervised machine learning models for lymphoma classification and HCM risk stratification in two rare clinical domains. The proposed computational pipeline can be deployed for the augmentation of clinical data although medical imaging features extracted from radiomics analysis can also be used as input. To our knowledge, this is the first computational pipeline which aggregates high-quality virtual data with real data to deal with clinical unmet

needs in two rare clinical domains, including the development of robust lymphoma classification and HCM risk stratification models. The data quality control module enhanced the quality of the raw clinical data through the removal of outliers and duplicated fields. The virtual population generation module provided straightforward virtual data generators, where the tree ensemble generators were extended to avoid overfitting effects during the generation stage yielding virtual data with increased quality in terms of increased convergence with the real data. A similar strategy was developed for the ANNs using Gaussian kernels as activation functions to deal with overfitting during the training stage.

The "hybrid" machine learning module utilizes supervised machine learning algorithms on the aggregated real and high-quality virtual data to enhance the performance of the lymphoma classification and HCM risk stratification models. Although the application of the proposed pipeline has a strong potential towards the improvement of the existing disease classification and risk stratification models in other clinical domains, emphasis must be given on its concise application in each clinical domain of interest. Although the virtual population generators and specifically the tree ensembles and the ANNs have been adjusted to resolve overfitting effects during the training stage, emphasis should be given on the precise definition of the data types of the input features to avoid the generation of virtual data with heterogeneous data structure. Finally, the statistical power of the augmented clinical data must be sufficient for the application of the hybrid machine learning module to yield robust disease classification and risk stratification models. According to our findings from the case studies (CHAPTER 7), the proposed data augmentation pipeline enhanced the performance of the existing lymphoma classification models in pSS and HCM risk stratification models in the homonymous domains, highlighting its increased predictive value towards the improvement of the disease classification and risk stratification models.

In this thesis, we also presented the FHBF algorithm as a new paradigm towards the design, development, and deployment of robust and unbiased supervised machine learning models across federated databases with highly imbalanced clinical data structures, where the arbitrary selection of dropout rates (in FDART) combined with the increased class imbalance can cause overfitting effects. To this end, we first defined a hybrid loss with a configurable topology which accounts for overfitting effects. The

customized loss function was then introduced into the FDART schema by estimating the gradient and hessian vectors. Confound-based random downsampling with replacement was applied on each federated database to match the target population with the controls with respect to three confound factors (age at disease diagnosis, gender, disease duration). The process was repeated K times, where the HFGBTs from each round are assembled to formulate a cluster of trees in the form of a forest. The log loss score was computed for each cluster of HFGBTs to identify and discard "weak" clusters, where the final decision-making process was based on a majority voting schema based on the predictions of the trees across the most dominant clusters. The FHBF dominated the existing state of the art federated learning schemas in terms of classification accuracy and reduced log loss during training and testing under two experimental cases involving the development of six federated AI models for the classification of rare lymphoma types across a Pan-European data hub with rare autoimmune diseases with increased class imbalance.

According to our findings from the case studies (CHAPTER 7), the proposed federated AI modeling framework is highly scalable and generalizable since it can support federated learning through a central node which communicates with private nodes (in the cloud) or distributed nodes (in the case of local distributed database management system). The proposed federated AI framework was successfully utilized towards the development of lymphomagenesis models in pSS, where the proposed FHBF algorithm yielded highly-robust supervised learning models resilient against overfitting effects along with explainable risk factors for lymphomagenesis compared to existing state-of-the art implementations.

We also focused on the development of a straightforward, multimodal, and explainable AI model to predict the risk of intensive care and mortality across multiple timepoints with accuracy 0.79 and 0.81, respectively. The extracted biomarkers were combined with the APACHE-II score to formulate a highly robust ICU scoring index with accuracy up to 0.9. In addition, we identified major factors for ICU admission and mortality, including the number of lymphocytes, PO2/FiO2, percentage of neutrophils and lymphocytes, LDH, and ALP at the baseline or during the follow-up as prominent for ICU admission and mortality in COVID-19. The contribution of the extracted clusters yielded an improved classification performance both for ICU admission

(sensitivity 0.83, specificity 0.83) and mortality (sensitivity 0.74, specificity 0.76). Thorough investigation of the derived patient subgroups (i.e., clusters and trajectories) would permit the identification of major factors at the baseline or during the follow-up period that contribute to risk stratification of COVID-19 patients. The sensitivity of the classifier for mortality was improved by 4% using demographic-related data while the specificity was improved by 4% in the case where the baseline clinical data are included and by 3% in the case where the demographics and the therapies-related data were incorporated. The features "number of lymphocytes", "SatO2", "PO2/FiO2" and "O2 supply type" were highlighted as risk factors for ICU admission and the percentage of neutrophils and lymphocytes, PO2/FiO2, LDH and ALP for mortality, among others.

Although most of the existing studies (Table 66, Table 67) focus on the development of ICU admission and mortality classifiers without taking into consideration the underlying dynamic associations among the data, the proposed method combines dynamic modeling with clustering analysis to identify subgroups of COVID-19 patients with common clinical profiles which are in turn utilized for the development of robust classifiers for ICU admission and mortality. According to our findings from the case studies (CHAPTER 7), the proposed multimodal AI modeling pipeline was able to identify hidden patterns among diverse COVID-19 patient subgroups. This information was used to enrich the performance of the classifiers for ICU admission and mortality yielding not only classifiers with improved performance but also explainable risk factors for ICU admission and mortality followed by an enhanced ICU scoring index complementing the APACHE-II score.

9.2. Future work

As a future work, we plan to extend the functionalities of the medical data curation service by integrating the proposed "smart" data imputer in the existing REST API service of the data curator. In addition, we plan to apply the hybrid data harmonizer across multiple databases in other domains, such as, cancer, as well as, in other medical data types (e.g., omics) to further evaluate the consistency of the produced harmonized data after the execution of the data harmonization process. Although a preliminary version of the hybrid data harmonization service is available, we plan to develop a front-end user interface so that anyone will be able to apply the data harmonization REST API service in a pre-defined reference ontology for the disease of interest. The proposed

BGGM-OCE algorithm could potentially provide significant insight in the field of virtual population generation to re-adjust the perspective of *in silico* Clinical Trials (CTs) in other rare diseases, as well as, to evaluate the approach in time-series data, including gene expression microarray data. To this end, we also plan to deploy artificial neural networks (ANNs) with multiple hidden layers towards the generation of even more robust clinical data for *in-silico* clinical trials. Furthermore, we plan to apply the proposed data augmentation pipeline in other clinical domains to enhance the population size of clinical research databases with reduced statistical power.

In addition, we plan to expand the hybrid machine learning module of the proposed computational pipeline for data augmentation using deep learning algorithms to support the extraction of biomarkers from time-series gene expression data, as well as, to enhance the applicability of the data quality control module to support the curation of complex genetic data structures. As far as federated/distributed learning is concerned, we plan to include additional cohort data for the application of the federated AI modeling framework towards the development of even more robust federated learning classifiers, which, however, would not only be restricted to the domain of autoimmune diseases but also to other domains, such as, cancer, cardiovascular diseases, etc. In addition, we plan to further enhance the performance of the federated AI models by including genetic data (e.g., FMS-like tyrosine kinase 3 ligand). Although the proposed classifiers do not exist in distributed libraries like Apache Spark's MLlib [490] we plan to conduct a comparison study in the future.

Moreover, we plan to extend the explainability of the classifiers by measuring the impact of each ensemble in the decision-making process and explore new utilities to avoid biases during the training stage. We also plan to test the performance of the FHBF in extreme data mining applications under the PRECIOUS (MIS: 5047133) hyper convergence infrastructure. The proposed federated/distributed learning framework could be also applied on more SAIDs-oriented genetic data to provide new insights on the underlying pathogenic mechanisms and biomarkers of SAIDs, such as, the Cryopyrin-Associated Autoinflammatory Syndromes (CAPS), the Hyperimmunoglobulinemia D syndrome (HIDS), and the Pharyngitis and cervical Adenitis (PFAPA), among others. We plan to utilize the proposed multimodal AI-based data analysis pipeline across a larger sample of hospitalized COVID-19 patients in the

future and analyze follow-up data across more time points, further enhancing the statistical power of the outcomes. We also plan to explore the fusion of the available clinical and laboratory related data with RNA-sequencing (transcriptomic) data and/or imaging-based features to shed light into the genetic mechanisms and underlying associations of the existing risk prediction factors of COVID-19 for ICU admission and mortality.

Literature

[1] S. J. Miah, E. Camilleri, and H. Q. Vu, “Big Data in Healthcare Research: A survey study,” *J. Comput. Inf. Syst.*, vol. 62, no. 3, pp. 480–492, May 2022, doi: 10.1080/08874417.2020.1858727.

[2] A. Rehman, S. Naz, and I. Razzak, “Leveraging big data analytics in healthcare enhancement: trends, challenges and opportunities,” *Multimed. Syst.*, Jan. 2021, doi: 10.1007/s00530-020-00736-8.

[3] I. J. Borges do Nascimento *et al.*, “Impact of Big Data Analytics on People’s Health: Overview of Systematic Reviews and Recommendations for Future Studies,” *J. Med. Internet Res.*, vol. 23, no. 4, p. e27275, Apr. 2021, doi: 10.2196/27275.

[4] S. Shilo, H. Rossman, and E. Segal, “Axes of a revolution: challenges and promises of big data in healthcare,” *Nat. Med.*, vol. 26, no. 1, pp. 29–38, Jan. 2020, doi: 10.1038/s41591-019-0727-5.

[5] M. Naeem *et al.*, “Trends and Future Perspective Challenges in Big Data,” in *Advances in Intelligent Data Analysis and Applications*, vol. 253, J.-S. Pan, V. E. Balas, and C.-M. Chen, Eds. Singapore: Springer Singapore, 2022, pp. 309–325. doi: 10.1007/978-981-16-5036-9_30.

[6] M. Karatas, L. Eriskin, M. Deveci, D. Pamucar, and H. Garg, “Big Data for Healthcare Industry 4.0: Applications, challenges and future perspectives,” *Expert Syst. Appl.*, vol. 200, p. 116912, Aug. 2022, doi: 10.1016/j.eswa.2022.116912.

[7] H. Li and P. Boulanger, “A Survey of Heart Anomaly Detection Using Ambulatory Electrocardiogram (ECG),” *Sensors*, vol. 20, no. 5, p. 1461, Mar. 2020, doi: 10.3390/s20051461.

[8] M. Soufineyestani, D. Dowling, and A. Khan, “Electroencephalography (EEG) Technology Applications and Available Devices,” *Appl. Sci.*, vol. 10, no. 21, p. 7453, Oct. 2020, doi: 10.3390/app10217453.

[9] M. J. Brookes *et al.*, “Magnetoencephalography with optically pumped magnetometers (OPM-MEG): the next generation of functional neuroimaging,” *Trends Neurosci.*, p. S0166223622001023, Jun. 2022, doi: 10.1016/j.tins.2022.05.008.

[10] R. M. Enoka, “Physiological validation of the decomposition of surface EMG signals,” *J. Electromyogr. Kinesiol.*, vol. 46, pp. 70–83, Jun. 2019, doi: 10.1016/j.jelekin.2019.03.010.

[11] R. Reda, M. Tantawi, H. shedeed, and M. F. Tolba, “Analyzing Electrooculography (EOG) for Eye Movement Detection,” in *The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2019)*, vol. 921, A. E. Hassanien, A. T. Azar, T. Gaber, R. Bhatnagar, and M. F. Tolba, Eds. Cham: Springer International Publishing, 2020, pp. 179–189. doi: 10.1007/978-3-030-14118-9_18.

[12] N. Adra *et al.*, “Optimal spindle detection parameters for predicting cognitive performance,” *Sleep*, vol. 45, no. 4, p. zsac001, Apr. 2022, doi: 10.1093/sleep/zsac001.

[13] S. K. Prabhakar, H. Rajaguru, C. Kim, and D.-O. Won, “A Fusion-Based Technique With Hybrid Swarm Algorithm and Deep Learning for Biosignal Classification,” *Front. Hum. Neurosci.*, vol. 16, p. 895761, Jun. 2022, doi: 10.3389/fnhum.2022.895761.

[14] A. H. Kashou and P. A. Noseworthy, “Electrocardiographic biosignals to predict atrial fibrillation: Are we there yet?,” *J. Electrocardiol.*, vol. 70, pp. 37–38, Jan. 2022, doi: 10.1016/j.jelectrocard.2021.11.033.

[15] R. Soundararajan *et al.*, “Deeply Trained Real-Time Body Sensor Networks for Analyzing the Symptoms of Parkinson’s Disease,” *IEEE Access*, vol. 10, pp. 63403–63421, 2022, doi: 10.1109/ACCESS.2022.3181985.

[16] P. J. Withers *et al.*, “X-ray computed tomography,” *Nat. Rev. Methods Primer*, vol. 1, no. 1, p. 18, Dec. 2021, doi: 10.1038/s43586-021-00015-4.

[17] H. Chandarana, H. Wang, R. H. N. Tijssen, and I. J. Das, “Emerging role of MRI in radiation therapy: Emerging Role of MRI in Radiation Therapy,” *J. Magn. Reson. Imaging*, vol. 48, no. 6, pp. 1468–1478, Dec. 2018, doi: 10.1002/jmri.26271.

[18] J. M. Hooker and R. E. Carson, “Human Positron Emission Tomography Neuroimaging,” *Annu. Rev. Biomed. Eng.*, vol. 21, no. 1, pp. 551–581, Jun. 2019, doi: 10.1146/annurev-bioeng-062117-121056.

[19] O. Israel *et al.*, “Two decades of SPECT/CT – the coming of age of a technology: An updated review of literature evidence,” *Eur. J. Nucl. Med. Mol. Imaging*, vol. 46, no. 10, pp. 1990–2012, Sep. 2019, doi: 10.1007/s00259-019-04404-6.

[20] M. L. Elliott, A. R. Knodt, and A. R. Hariri, “Striving toward translation: strategies for reliable fMRI measurement,” *Trends Cogn. Sci.*, vol. 25, no. 9, pp. 776–787, Sep. 2021, doi: 10.1016/j.tics.2021.05.008.

[21] M. J. Smith, S. A. Hayward, S. M. Innes, and A. S. C. Miller, “Point-of-care lung ultrasound in patients with COVID -19 – a narrative review,” *Anaesthesia*, vol. 75, no. 8, pp. 1096–1104, Aug. 2020, doi: 10.1111/anae.15082.

[22] Y. Jin, Y. Yin, C. Li, H. Liu, and J. Shi, “Non-Invasive Monitoring of Human Health by Photoacoustic Spectroscopy,” *Sensors*, vol. 22, no. 3, p. 1155, Feb. 2022, doi: 10.3390/s22031155.

[23] Quaresima and Ferrari, “A Mini-Review on Functional Near-Infrared Spectroscopy (fNIRS): Where Do We Stand, and Where Should We Go?,” *Photonics*, vol. 6, no. 3, p. 87, Aug. 2019, doi: 10.3390/photonics6030087.

[24] R. M. Sherman and S. L. Salzberg, “Pan-genomics in the human genome era,” *Nat. Rev. Genet.*, vol. 21, no. 4, pp. 243–254, Apr. 2020, doi: 10.1038/s41576-020-0210-7.

[25] T. Züllig and H. C. Köfeler, “HIGH RESOLUTION MASS SPECTROMETRY IN LIPIDOMICS,” *Mass Spectrom. Rev.*, vol. 40, no. 3, pp. 162–176, May 2021, doi: 10.1002/mas.21627.

[26] N. Pappireddi, L. Martin, and M. Wühr, “A Review on Quantitative Multiplexed Proteomics,” *ChemBioChem*, vol. 20, no. 10, pp. 1210–1224, May 2019, doi: 10.1002/cbic.201800650.

[27] S. Alseekh *et al.*, “Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices,” *Nat. Methods*, vol. 18, no. 7, pp. 747–756, Jul. 2021, doi: 10.1038/s41592-021-01197-1.

[28] M. Kato and R. Natarajan, “Epigenetics and epigenomics in diabetic kidney disease and metabolic memory,” *Nat. Rev. Nephrol.*, vol. 15, no. 6, pp. 327–345, Jun. 2019, doi: 10.1038/s41581-019-0135-6.

[29] L. M. LaFave, R. E. Savage, and J. D. Buenrostro, “Single-Cell Epigenomics Reveals Mechanisms of Cancer Progression,” *Annu. Rev. Cancer Biol.*, vol. 6, no. 1, pp. 167–185, Apr. 2022, doi: 10.1146/annurev-cancerbio-070620-094453.

[30] A. Kulkarni, A. G. Anderson, D. P. Merullo, and G. Konopka, “Beyond bulk: a review of single cell transcriptomics methodologies and applications,” *Curr. Opin. Biotechnol.*, vol. 58, pp. 129–136, Aug. 2019, doi: 10.1016/j.copbio.2019.03.001.

[31] S. Morganti *et al.*, “Complexity of genome sequencing and reporting: Next generation sequencing (NGS) technologies and implementation of precision medicine in real life,” *Crit. Rev. Oncol. Hematol.*, vol. 133, pp. 171–182, Jan. 2019, doi: 10.1016/j.critrevonc.2018.11.008.

[32] R. Stark, M. Grzelak, and J. Hadfield, “RNA sequencing: the teenage years,” *Nat. Rev. Genet.*, vol. 20, no. 11, pp. 631–656, Nov. 2019, doi: 10.1038/s41576-019-0150-2.

[33] F. Anyakudo, E. Adams, and A. Van Schepdael, “Thin-Layer Chromatography–Flame Ionization Detection,” *Chromatographia*, vol. 83, no. 2, pp. 149–157, Feb. 2020, doi: 10.1007/s10337-019-03849-z.

[34] J. Wang and S. Nie, “Application of atomic force microscopy in microscopic analysis of polysaccharide,” *Trends Food Sci. Technol.*, vol. 87, pp. 35–46, May 2019, doi: 10.1016/j.tifs.2018.02.005.

[35] M. Corrales, S. Doizi, Y. Barghouthy, H. Kamkoum, B. Somani, and O. Traxer, “Ultrasound or Fluoroscopy for Percutaneous Nephrolithotomy Access, Is There Really a Difference? A Review of Literature,” *J. Endourol.*, vol. 35, no. 3, pp. 241–248, Mar. 2021, doi: 10.1089/end.2020.0672.

[36] D. Jain *et al.*, “Immunocytochemistry for predictive biomarker testing in lung cancer cytology,” *Cancer Cytopathol.*, vol. 127, no. 5, pp. 325–339, May 2019, doi: 10.1002/cncy.22137.

[37] W. C. C. Tan *et al.*, “Overview of multiplex immunohistochemistry/immunofluorescence techniques in the era of cancer immunotherapy,” *Cancer Commun.*, vol. 40, no. 4, pp. 135–153, Apr. 2020, doi: 10.1002/cac2.12023.

[38] M. I. Tariq, V. E. Balas, and S. Tayyaba, *Security and Privacy Trends in Cloud Computing and Big Data*, 1st ed. Boca Raton: CRC Press, 2022. doi: 10.1201/9781003107286.

[39] Y. Zhong, L. Chen, C. Dan, and A. Rezaeipanah, “A systematic survey of data mining and big data analysis in internet of things,” *J. Supercomput.*, Jun. 2022, doi: 10.1007/s11227-022-04594-1.

[40] R. Kleiman *et al.*, “Comparison of electrocardiograms (ECG) waveforms and centralized ECG measurements between a simple 6-lead mobile ECG device and a standard 12-lead ECG,” *Ann. Noninvasive Electrocardiol.*, vol. 26, no. 6, Nov. 2021, doi: 10.1111/anec.12872.

[41] L. M. Silva, K. M. S. Silva, W. G. Lira-Bandeira, A. C. Costa-Ribeiro, and S. A. Araújo-Neto, “Localizing the Primary Motor Cortex of the Hand by the 10-5 and 10-20 Systems for Neurostimulation: An MRI Study,” *Clin. EEG Neurosci.*, vol. 52, no. 6, pp. 427–435, Oct. 2021, doi: 10.1177/1550059420934590.

[42] R. Schofield *et al.*, “Image reconstruction: Part 1 – understanding filtered back projection, noise and image acquisition,” *J. Cardiovasc. Comput. Tomogr.*, vol. 14, no. 3, pp. 219–225, May 2020, doi: 10.1016/j.jcct.2019.04.008.

[43] S. Zhang and Y. Xia, “CT image reconstruction algorithms: A comprehensive survey,” *Concurr. Comput. Pract. Exp.*, vol. 33, no. 8, Apr. 2021, doi: 10.1002/cpe.5506.

[44] X. Zhang, J. Sun, and X. Cao, “Robust direction-of-arrival estimation based on sparse asymptotic minimum variance,” *J. Eng.*, vol. 2019, no. 21, pp. 7815–7821, Nov. 2019, doi: 10.1049/joe.2019.0720.

[45] C. Yang, H. Lan, F. Gao, and F. Gao, “Review of deep learning for photoacoustic imaging,” *Photoacoustics*, vol. 21, p. 100215, Mar. 2021, doi: 10.1016/j.pacs.2020.100215.

[46] A. A. Rezigalla, “Observational Study Designs: Synopsis for Selecting an Appropriate Study Design,” *Cureus*, Jan. 2020, doi: 10.7759/cureus.6692.

[47] L. Martinengo *et al.*, “Prevalence of chronic wounds in the general population: systematic review and meta-analysis of observational studies,” *Ann. Epidemiol.*, vol. 29, pp. 8–15, Jan. 2019, doi: 10.1016/j.annepidem.2018.10.005.

[48] A. George, T. S. Stead, and L. Ganti, “What’s the Risk: Differentiating Risk Ratios, Odds Ratios, and Hazard Ratios?,” *Cureus*, Aug. 2020, doi: 10.7759/cureus.10047.

[49] M. J. Stensrud and M. A. Hernán, “Why Test for Proportional Hazards?,” *JAMA*, vol. 323, no. 14, p. 1401, Apr. 2020, doi: 10.1001/jama.2020.1267.

[50] the SEED Lifecourse Sciences Theme *et al.*, “Retention strategies in longitudinal cohort studies: a systematic review and meta-analysis,” *BMC Med. Res. Methodol.*, vol. 18, no. 1, p. 151, Dec. 2018, doi: 10.1186/s12874-018-0586-7.

[51] J. W. Song and K. C. Chung, “Observational Studies: Cohort and Case-Control Studies:,” *Plast. Reconstr. Surg.*, vol. 126, no. 6, pp. 2234–2242, Dec. 2010, doi: 10.1097/PRS.0b013e3181f44abc.

[52] V. C. Pezoulas, T. P. Exarchos, and D. I. Fotiadis, *Medical data sharing, harmonization and analytics*. London, United Kingdom ; San Diego, CA:

Academic Press, 2020.

[53] N. Ahmed, Ed., *Clinical biochemistry*, Second edition. Oxford: Oxford University Press, 2017.

[54] E. Lambrinou, T. B. Hansen, and J. W. Beulens, "Lifestyle factors, self-management and patient empowerment in diabetes care," *Eur. J. Prev. Cardiol.*, vol. 26, no. 2_suppl, pp. 55–63, Dec. 2019, doi: 10.1177/2047487319885455.

[55] G. Currie, K. E. Hawk, E. Rohren, A. Vial, and R. Klein, "Machine Learning and Deep Learning in Medical Imaging: Intelligent Imaging," *J. Med. Imaging Radiat. Sci.*, vol. 50, no. 4, pp. 477–487, Dec. 2019, doi: 10.1016/j.jmir.2019.09.005.

[56] K. M. Tsiouris, V. C. Pezoulas, M. Zervakis, S. Konitsiotis, D. D. Koutsouris, and D. I. Fotiadis, "A Long Short-Term Memory deep learning network for the prediction of epileptic seizures using EEG signals," *Comput. Biol. Med.*, vol. 99, pp. 24–37, Aug. 2018, doi: 10.1016/j.compbiomed.2018.05.019.

[57] I. R. Management Association, Ed., *Healthcare Policy and Reform: Concepts, Methodologies, Tools, and Applications*. IGI Global, 2019. doi: 10.4018/978-1-5225-6915-2.

[58] C. Krittanawong *et al.*, "Integration of novel monitoring devices with machine learning technology for scalable cardiovascular management," *Nat. Rev. Cardiol.*, vol. 18, no. 2, pp. 75–91, Feb. 2021, doi: 10.1038/s41569-020-00445-9.

[59] B. Li and M. D. Ritchie, "From GWAS to Gene: Transcriptome-Wide Association Studies and Other Methods to Functionally Understand GWAS Discoveries," *Front. Genet.*, vol. 12, p. 713230, Sep. 2021, doi: 10.3389/fgene.2021.713230.

[60] O. Azimzadeh, M. Gomolka, M. Birschwilks, S. Saigusa, B. Grosche, and S. Moertl, "Advanced Omics and Radiobiological Tissue Archives: The Future in the Past," *Appl. Sci.*, vol. 11, no. 23, p. 11108, Nov. 2021, doi: 10.3390/app112311108.

[61] P. Tolani, S. Gupta, K. Yadav, S. Aggarwal, and A. K. Yadav, "Big data, integrative omics and network biology," in *Advances in Protein Chemistry and Structural Biology*, vol. 127, Elsevier, 2021, pp. 127–160. doi: 10.1016/bs.apcsb.2021.03.006.

[62] J. Vaught, P. Hainaut, M. Pasterk, and K. Zatloukal, "The Future of Biobanking: Meeting Tomorrow's Challenges," in *Biobanking of Human Biospecimens*, P. Hainaut, J. Vaught, K. Zatloukal, and M. Pasterk, Eds. Cham: Springer International Publishing, 2021, pp. 187–197. doi: 10.1007/978-3-030-55901-

4_11.

[63] T. Hulsen, “Sharing Is Caring—Data Sharing Initiatives in Healthcare,” *Int. J. Environ. Res. Public. Health*, vol. 17, no. 9, p. 3046, Apr. 2020, doi: 10.3390/ijerph17093046.

[64] M. Janssen, P. Brous, E. Estevez, L. S. Barbosa, and T. Janowski, “Data governance: Organizing data for trustworthy Artificial Intelligence,” *Gov. Inf. Q.*, vol. 37, no. 3, p. 101493, Jul. 2020, doi: 10.1016/j.giq.2020.101493.

[65] V. C. Pezoulas *et al.*, “Addressing the clinical unmet needs in primary Sjögren’s Syndrome through the sharing, harmonization and federated analysis of 21 European cohorts,” *Comput. Struct. Biotechnol. J.*, vol. 20, pp. 471–484, 2022, doi: 10.1016/j.csbj.2022.01.002.

[66] V. C. Pezoulas *et al.*, “Medical data quality assessment: On the development of an automated framework for medical data curation,” *Comput. Biol. Med.*, vol. 107, pp. 270–283, Apr. 2019, doi: 10.1016/j.compbiomed.2019.03.001.

[67] F. Fox, V. R. Aggarwal, H. Whelton, and O. Johnson, “A Data Quality Framework for Process Mining of Electronic Health Record Data,” in *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, New York, NY, Jun. 2018, pp. 12–21. doi: 10.1109/ICHI.2018.00009.

[68] N. Martin, “Data Quality in Process Mining,” in *Interactive Process Mining in Healthcare*, C. Fernandez-Llatas, Ed. Cham: Springer International Publishing, 2021, pp. 53–79. doi: 10.1007/978-3-030-53993-1_5.

[69] V. Huser, M. G. Kahn, J. S. Brown, and R. Gouripeddi, “Methods for examining data quality in healthcare integrated data repositories,” in *Biocomputing 2018*, Kohala Coast, Hawaii, USA, Jan. 2018, pp. 628–633. doi: 10.1142/9789813235533_0059.

[70] B. Murray *et al.*, “Accessible data curation and analytics for international-scale citizen science datasets,” *Sci. Data*, vol. 8, no. 1, p. 297, Dec. 2021, doi: 10.1038/s41597-021-01071-x.

[71] M. A. Dakka *et al.*, “Automated detection of poor-quality data: case studies in healthcare,” *Sci. Rep.*, vol. 11, no. 1, p. 18005, Dec. 2021, doi: 10.1038/s41598-021-97341-0.

[72] A. Ismail, A. Shehab, and I. M. El-Henawy, “Healthcare Analysis in Smart Big Data Analytics: Reviews, Challenges and Recommendations,” in *Security in Smart Cities: Models, Applications, and Challenges*, A. E. Hassanien, M.

Elhoseny, S. H. Ahmed, and A. K. Singh, Eds. Cham: Springer International Publishing, 2019, pp. 27–45. doi: 10.1007/978-3-030-01560-2_2.

[73] K. D. Kourou *et al.*, "Cohort Harmonization and Integrative Analysis From a Biomedical Engineering Perspective," *IEEE Rev. Biomed. Eng.*, vol. 12, pp. 303–318, 2019, doi: 10.1109/RBME.2018.2855055.

[74] K. Adhikari *et al.*, "Data Harmonization and Data Pooling from Cohort Studies: A Practical Approach for Data Management," *Int. J. Popul. Data Sci.*, vol. 6, no. 1, Nov. 2021, doi: 10.23889/ijpds.v6i1.1680.

[75] S. J. Grannis *et al.*, "Evaluating the effect of data standardization and validation on patient matching accuracy," *J. Am. Med. Inform. Assoc.*, vol. 26, no. 5, pp. 447–456, May 2019, doi: 10.1093/jamia/ocy191.

[76] S. Kalkman, M. Mostert, C. Gerlinger, J. J. M. van Delden, and G. J. M. W. van Thiel, "Responsible data sharing in international health research: a systematic review of principles and norms," *BMC Med. Ethics*, vol. 20, no. 1, p. 21, Mar. 2019, doi: 10.1186/s12910-019-0359-9.

[77] S. Kalkman, J. van Delden, A. Banerjee, B. Tyl, M. Mostert, and G. van Thiel, "Patients' and public views and attitudes towards the sharing of health data for research: a narrative review of the empirical evidence," *J. Med. Ethics*, vol. 48, no. 1, p. 3, Jan. 2022, doi: 10.1136/medethics-2019-105651.

[78] D. L. Heymann, "Data sharing and outbreaks: best practice exemplified," *The Lancet*, vol. 395, no. 10223, pp. 469–470, Feb. 2020, doi: 10.1016/S0140-6736(20)30184-7.

[79] C. V. Cosgriff, D. K. Ebner, and L. A. Celi, "Data sharing in the era of COVID-19," *Lancet Digit. Health*, vol. 2, no. 5, p. e224, May 2020, doi: 10.1016/S2589-7500(20)30082-0.

[80] N. Gruschka, V. Mavroeidis, K. Vishi, and M. Jensen, "Privacy Issues and Data Protection in Big Data: A Case Study Analysis under GDPR," in *2018 IEEE International Conference on Big Data (Big Data)*, Seattle, WA, USA, Dec. 2018, pp. 5027–5033. doi: 10.1109/BigData.2018.8622621.

[81] N. Rieke *et al.*, "The future of digital health with federated learning," *Npj Digit. Med.*, vol. 3, no. 1, p. 119, Dec. 2020, doi: 10.1038/s41746-020-00323-1.

[82] I. Goldstein, C. S. Spatt, and M. Ye, "Big Data in Finance," *Rev. Financ. Stud.*, vol. 34, no. 7, pp. 3213–3225, Jun. 2021, doi: 10.1093/rfs/hhab038.

[83] M. Riba, C. Sala, D. Toniolo, and G. Tonon, "Big Data in Medicine, the Present

and Hopefully the Future," *Front. Med.*, vol. 6, p. 263, Nov. 2019, doi: 10.3389/fmed.2019.00263.

[84] M. Javaid, A. Haleem, R. P. Singh, and R. Suman, "Significant Applications of Big Data in Industry 4.0," *J. Ind. Integr. Manag.*, vol. 06, no. 04, pp. 429–447, Dec. 2021, doi: 10.1142/S2424862221500135.

[85] M. Younas, "Research challenges of big data," *Serv. Oriented Comput. Appl.*, vol. 13, no. 2, pp. 105–107, Jun. 2019, doi: 10.1007/s11761-019-00265-x.

[86] F. Cappa, R. Oriani, E. Peruffo, and I. McCarthy, "Big Data for Creating and Capturing Value in the Digitalized Environment: Unpacking the Effects of Volume, Variety, and Veracity on Firm Performance*," *J. Prod. Innov. Manag.*, vol. 38, no. 1, pp. 49–67, Jan. 2021, doi: 10.1111/jpim.12545.

[87] B. Jabir, F. Noureddine, and K. Rahmani, "Big Data Analytics Opportunities and Challenges for the Smart Enterprise," in *Distributed Sensing and Intelligent Systems*, M. Elhoseny, X. Yuan, and S. Krit, Eds. Cham: Springer International Publishing, 2022, pp. 833–845. doi: 10.1007/978-3-030-64258-7_70.

[88] J. Buhmann *et al.*, "Automatic detection of synaptic partners in a whole-brain Drosophila electron microscopy data set," *Nat. Methods*, vol. 18, no. 7, pp. 771–774, Jul. 2021, doi: 10.1038/s41592-021-01183-7.

[89] A. Petrenko, E. Wijmans, B. Shacklett, and V. Koltun, "Megaverse: Simulating Embodied Agents at One Million Experiences per Second," in *Proceedings of the 38th International Conference on Machine Learning*, Jul. 2021, vol. 139, pp. 8556–8566. [Online]. Available: https://proceedings.mlr.press/v139/petrenko21a.html

[90] H. Li, Z. Chen, Q. Gong, and Z. Jia, "Voxel-wise meta-analysis of task-related brain activation abnormalities in major depressive disorder with suicide behavior," *Brain Imaging Behav.*, vol. 14, no. 4, pp. 1298–1308, Aug. 2020, doi: 10.1007/s11682-019-00045-3.

[91] C. Yu and L. Huang, "Cross-Linking Mass Spectrometry: An Emerging Technology for Interactomics and Structural Biology," *Anal. Chem.*, vol. 90, no. 1, pp. 144–165, Jan. 2018, doi: 10.1021/acs.analchem.7b04431.

[92] I. Bludau, "Discovery–Versus Hypothesis–Driven Detection of Protein–Protein Interactions and Complexes," *Int. J. Mol. Sci.*, vol. 22, no. 9, 2021, doi: 10.3390/ijms22094450.

[93] M. Freeman, *Human rights*, Fourth edition. Cambridge, UK ; Medford, MA, USA:

Polity Press, 2022.

[94] E. Politou, E. Alepis, and C. Patsakis, “Forgetting personal data and revoking consent under the GDPR: Challenges and proposed solutions,” *J. Cybersecurity*, vol. 4, no. 1, p. tyy001, Jan. 2018, doi: 10.1093/cybsec/tyy001.

[95] C. Ryngaert and M. Taylor, “The GDPR as Global Data Protection Regulation?,” *AJIL Unbound*, vol. 114, pp. 5–9, 2020, doi: 10.1017/aju.2019.80.

[96] J. H. Gerards, *General principles of the European Convention on Human Rights law*. Cambride, United Kingdom ; New York: Cambridge University Press, 2019.

[97] J. E. Szalados, “Regulations and Regulatory Compliance: False Claims Act, Kickback and Stark Laws, and HIPAA,” in *The Medical-Legal Aspects of Acute Care Medicine: A Resource for Clinicians, Administrators, and Risk Managers*, J. E. Szalados, Ed. Cham: Springer International Publishing, 2021, pp. 277–313. doi: 10.1007/978-3-030-68570-6_12.

[98] W. Moore and S. Frye, “Review of HIPAA, Part 1: History, Protected Health Information, and Privacy and Security Rules,” *J. Nucl. Med. Technol.*, vol. 47, no. 4, pp. 269–272, Dec. 2019, doi: 10.2967/jnmt.119.227819.

[99] H. Aftab, K. Gilani, J. Lee, L. Nkenyereye, S. Jeong, and J. Song, “Analysis of identifiers in IoT platforms,” *Digit. Commun. Netw.*, vol. 6, no. 3, pp. 333–340, Aug. 2020, doi: 10.1016/j.dcan.2019.05.003.

[100] B. B. Gupta and M. Quamara, “An overview of Internet of Things (IoT): Architectural aspects, challenges, and protocols,” *Concurr. Comput. Pract. Exp.*, vol. 32, no. 21, Nov. 2020, doi: 10.1002/cpe.4946.

[101] X. Yao, F. Farha, R. Li, I. Psychoula, L. Chen, and H. Ning, “Security and privacy issues of physical objects in the IoT: Challenges and opportunities,” *Digit. Commun. Netw.*, vol. 7, no. 3, pp. 373–384, Aug. 2021, doi: 10.1016/j.dcan.2020.09.001.

[102] Y. S. Abdulsalam and M. Hedabou, “Security and Privacy in Cloud Computing: Technical Review,” *Future Internet*, vol. 14, no. 1, 2022, doi: 10.3390/fi14010011.

[103] W. G. Voss, “European Union Data Privacy Law Reform,” *Bus. Lawyer*, vol. 72, no. 1, pp. 221–234, 2016.

[104] G. Liu, “Data quality problems troubling business and financial researchers: A literature review and synthetic analysis,” *J. Bus. Finance Librariansh.*, vol. 25, no. 3–4, pp. 315–371, Oct. 2020, doi: 10.1080/08963568.2020.1847555.

[105] C. Shao, Y. Yang, S. Juneja, and T. GSeetharam, “IoT data visualization for business intelligence in corporate finance,” *Inf. Process. Manag.*, vol. 59, no. 1, p. 102736, Jan. 2022, doi: 10.1016/j.ipm.2021.102736.

[106] O. Azeroual, G. Saake, and M. Abuosba, “Data Quality Measures and Data Cleansing for Research Information Systems,” 2019, doi: 10.48550/ARXIV.1901.06208.

[107] S. Hoffman and A. Podgurski, “The Use and Misuse of Biomedical Data: Is Bigger Really Better?,” *Am. J. Law Med.*, vol. 39, no. 4, pp. 497–538, 2013, doi: 10.1177/009885881303900401.

[108] J. F. Hair and M. Sarstedt, “Data, measurement, and causal inferences in machine learning: opportunities and challenges for marketing,” *J. Mark. Theory Pract.*, vol. 29, no. 1, pp. 65–77, Jan. 2021, doi: 10.1080/10696679.2020.1860683.

[109] I. Fortier *et al.*, “Maelstrom Research guidelines for rigorous retrospective data harmonization,” *Int. J. Epidemiol.*, vol. 46, no. 1, pp. 103–105, Feb. 2017, doi: 10.1093/ije/dyw075.

[110] I. Fortier, N. Dragieva, M. Saliba, C. Craig, and P. J. Robson, “Harmonization of the Health and Risk Factor Questionnaire data of the Canadian Partnership for Tomorrow Project: a descriptive analysis,” *CMAJ Open*, vol. 7, no. 2, pp. E272–E282, Apr. 2019, doi: 10.9778/cmajo.20180062.

[111] E. R. van den Heuvel, L. E. Griffith, N. Sohel, I. Fortier, G. Muniz-Terrera, and P. Raina, “Latent variable models for harmonization of test scores: A case study on memory,” *Biom. J.*, vol. 62, no. 1, pp. 34–52, Jan. 2020, doi: 10.1002/bimj.201800146.

[112] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, and F. Wang, “Federated Learning for Healthcare Informatics,” *J. Healthc. Inform. Res.*, vol. 5, no. 1, pp. 1–19, Mar. 2021, doi: 10.1007/s41666-020-00082-4.

[113] S. R. Ali, J. Bryce, Y. Kodra, D. Taruscio, L. Persani, and S. F. Ahmed, “The Quality Evaluation of Rare Disease Registries—An Assessment of the Essential Features of a Disease Registry,” *Int. J. Environ. Res. Public. Health*, vol. 18, no. 22, 2021, doi: 10.3390/ijerph182211968.

[114] F. Pappalardo, G. Russo, F. M. Tshinanu, and M. Viceconti, “In silico clinical trials: concepts and early adoptions,” *Brief. Bioinform.*, vol. 20, no. 5, pp. 1699–1708, Sep. 2019, doi: 10.1093/bib/bby043.

[115] K.-K. Mak and M. R. Pichika, “Artificial intelligence in drug development: present status and future prospects,” *Drug Discov. Today*, vol. 24, no. 3, pp. 773–780, Mar. 2019, doi: 10.1016/j.drudis.2018.11.014.

[116] N. Noorbakhsh-Sabet, R. Zand, Y. Zhang, and V. Abedi, “Artificial Intelligence Transforms the Future of Health Care,” *Am. J. Med.*, vol. 132, no. 7, pp. 795–801, Jul. 2019, doi: 10.1016/j.amjmed.2019.01.017.

[117] S. Kumar, A. K. Bharti, and R. Amin, “Decentralized secure storage of medical records using Blockchain and IPFS : A comparative analysis with future directions,” *Secur. Priv.*, vol. 4, no. 5, Sep. 2021, doi: 10.1002/spy2.162.

[118] N. J. Podlesny, A. V. D. M. Kayem, and C. Meinel, “Towards Identifying De-anonymisation Risks in Distributed Health Data Silos,” in *Database and Expert Systems Applications*, Cham, 2019, pp. 33–43.

[119] S. Dash, S. K. Shakyawar, M. Sharma, and S. Kaushik, “Big data in healthcare: management, analysis and future prospects,” *J. Big Data*, vol. 6, no. 1, p. 54, Jun. 2019, doi: 10.1186/s40537-019-0217-0.

[120] M. Tzanou, *Health data privacy under the GDPR big data challenges and regulatory responses*. 2021. Accessed: Jul. 20, 2022. [Online]. Available: http://www.vlebooks.com/vleweb/product/openreader?id=none&isbn=9780429666568

[121] D. Kaur, S. Uslu, K. J. Rittichier, and A. Durresi, “Trustworthy Artificial Intelligence: A Review,” *ACM Comput. Surv.*, vol. 55, no. 2, pp. 1–38, Mar. 2023, doi: 10.1145/3491209.

[122] I. G. Cohen, T. Evgeniou, S. Gerke, and T. Minssen, “The European artificial intelligence strategy: implications and challenges for digital health,” *Lancet Digit. Health*, vol. 2, no. 7, pp. e376–e379, Jul. 2020, doi: 10.1016/S2589-7500(20)30112-6.

[123] N. A. Smuha, “The EU Approach to Ethics Guidelines for Trustworthy Artificial Intelligence,” *Comput. Law Rev. Int.*, vol. 20, no. 4, pp. 97–106, Aug. 2019, doi: 10.9785/cri-2019-200402.

[124] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in *Advances in Neural Information Processing Systems*, 2017, vol. 30. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[125] V. C. Pezoulas *et al.*, “Enhancing medical data quality through data curation: A case study in primary Sjögren’s syndrome,” *Clin Exp Rheumatol.*, vol. 37, no. 3, pp. 90–96, 2019.

[126] V. C. Pezoulas *et al.*, “A hybrid data harmonization workflow using word embeddings for the interlinking of heterogeneous cross-domain clinical data structures,” in *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2021, pp. 1–4.

[127] V. C. Pezoulas *et al.*, “Towards the Establishment of a Biomedical Ontology for the Primary Sjögren’s Syndrome,” *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Int. Conf.*, vol. 2018, pp. 4089–4092, Jul. 2018, doi: 10.1109/EMBC.2018.8513349.

[128] V. C. Pezoulas *et al.*, “Dealing with Open Issues and Unmet Needs in Healthcare Through Ontology Matching and Federated Learning,” in *8th European Medical and Biological Engineering Conference*, vol. 80, T. Jarm, A. Cvetkoska, S. Mahnič-Kalamiza, and D. Miklavcic, Eds. Cham: Springer International Publishing, 2021, pp. 306–313. doi: 10.1007/978-3-030-64610-3_36.

[129] V. C. Pezoulas *et al.*, “Overcoming the Barriers That Obscure the Interlinking and Analysis of Clinical Data Through Harmonization and Incremental Learning,” *IEEE Open J. Eng. Med. Biol.*, vol. 1, pp. 83–90, 2020, doi: 10.1109/OJEMB.2020.2981258.

[130] V. C. Pezoulas, G. I. Grigoriadis, N. S. Tachos, F. Barlocco, I. Olivotto, and D. I. Fotiadis, “Variational Gaussian Mixture Models with robust Dirichlet concentration priors for virtual population generation in hypertrophic cardiomyopathy: a comparison study,” in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2021, pp. 1674–1677.

[131] V. C. Pezoulas, G. I. Grigoriadis, N. S. Tachos, F. Barlocco, I. Olivotto, and D. I. Fotiadis, “Generation of virtual patient data for in-silico cardiomyopathies drug development using tree ensembles: a comparative study,” in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2020, pp. 5343–5346.

[132] V. Pezoulas, N. Tachos, and D. Fotiadis, “Generation of Virtual Patients for in Silico Cardiomyopathies Drug Development,” in *2019 IEEE 19th International*

Conference on Bioinformatics and Bioengineering (BIBE), Athens, Greece, Oct. 2019, pp. 671–674. doi: 10.1109/BIBE.2019.00126.

[133] V. C. Pezoulas, N. S. Tachos, G. Gkois, I. Olivotto, F. Barlocco, and D. I. Fotiadis, "Bayesian Inference-Based Gaussian Mixture Models With Optimal Components Estimation Towards Large-Scale Synthetic Data Generation for in Silico Clinical Trials," *IEEE Open J. Eng. Med. Biol.*, 2022.

[134] V. C. Pezoulas *et al.*, "A computational pipeline for data augmentation towards the improvement of disease classification and risk stratification models: A case study in two clinical domains," *Comput. Biol. Med.*, vol. 134, p. 104520, Jul. 2021, doi: 10.1016/j.compbiomed.2021.104520.

[135] V. C. Pezoulas, T. P. Exarchos, K. D. Kourou, A. G. Tzioufas, S. De Vita, and D. I. Fotiadis, "Utilizing Incremental Learning for the Prediction of Disease Outcomes Across Distributed Clinical Data: A Framework and a Case Study," in *XV Mediterranean Conference on Medical and Biological Engineering and Computing – MEDICON 2019*, Cham, 2020, pp. 823–831. doi: 10.1007/978-3-030-31635-8_98.

[136] N. G. Weiskopf and C. Weng, "Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research," *J. Am. Med. Inform. Assoc.*, vol. 20, no. 1, pp. 144–151, Jan. 2013, doi: 10.1136/amiajnl-2011-000681.

[137] A. P. Reimer, A. Milinovich, and E. A. Madigan, "Data quality assessment framework to assess electronic medical record data for use in research," *Int. J. Med. Inf.*, vol. 90, pp. 40–47, Jun. 2016, doi: 10.1016/j.ijmedinf.2016.03.006.

[138] M. G. Kahn, M. A. Raebel, J. M. Glanz, K. Riedlinger, and J. F. Steiner, "A Pragmatic Framework for Single-site and Multisite Data Quality Assessment in Electronic Health Record-based Clinical Research," *Med. Care*, vol. 50, pp. S21–S29, Jul. 2012, doi: 10.1097/MLR.0b013e318257dd67.

[139] F. A. Satti *et al.*, "Semantic Bridge for Resolving Healthcare Data Interoperability," in *2020 International Conference on Information Networking (ICOIN)*, Barcelona, Spain, Jan. 2020, pp. 86–91. doi: 10.1109/ICOIN48656.2020.9016461.

[140] S. Bauermeister *et al.*, "The Dementias Platform UK (DPUK) Data Portal," *Eur. J. Epidemiol.*, vol. 35, no. 6, pp. 601–611, Jun. 2020, doi: 10.1007/s10654-020-00633-4.

[141] I. Fortier *et al.*, “Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies,” *Int. J. Epidemiol.*, vol. 39, no. 5, pp. 1383–1393, Oct. 2010, doi: 10.1093/ije/dyq139.

[142] C. Pang *et al.*, “SORTA: a system for ontology-based re-coding and technical annotation of biomedical phenotype data,” *Database*, vol. 2015, p. bav089, Jan. 2015, doi: 10.1093/database/bav089.

[143] C. Pang *et al.*, “BiobankConnect: software to rapidly connect data elements for pooled analysis across biobanks using ontological and lexical indexing,” *J. Am. Med. Inform. Assoc.*, vol. 22, no. 1, pp. 65–75, Jan. 2015, doi: 10.1136/amiajnl-2013-002577.

[144] F. Giunchiglia, A. Autayeu, and J. Pane, “S-Match: An open source framework for matching lightweight ontologies,” *Semantic Web*, vol. 3, no. 3, pp. 307–317, 2012, doi: 10.3233/SW-2011-0036.

[145] M. Ehrig and Y. Sure-Vetter, “FOAM - Framework for Ontology Alignment and Mapping - Results of the Ontology Alignment Evaluation Initiative,” 2005.

[146] C. Pang *et al.*, “BiobankUniverse: automatic matchmaking between datasets for biobank data discovery and integration,” *Bioinformatics*, vol. 33, no. 22, pp. 3627–3634, Nov. 2017, doi: 10.1093/bioinformatics/btx478.

[147] I. Fortier *et al.*, “Is rigorous retrospective harmonization possible? Application of the DataSHaPER approach across 53 large studies,” *Int. J. Epidemiol.*, vol. 40, no. 5, pp. 1314–1328, Oct. 2011, doi: 10.1093/ije/dyr106.

[148] S. Scholtens *et al.*, “Cohort Profile: LifeLines, a three-generation cohort study and biobank,” *Int. J. Epidemiol.*, vol. 44, no. 4, pp. 1172–1180, Aug. 2015, doi: 10.1093/ije/dyu229.

[149] F. Giunchiglia, M. Yatskevich, P. Avesani, and P. Shivaiko, “A large dataset for the evaluation of ontology matching,” *Knowl. Eng. Rev.*, vol. 24, no. 2, pp. 137–157, 2009, doi: 10.1017/S026988890900023X.

[150] “Ontology Alignment Evaluation Initiative::Home.” http://oaei.ontologymatching.org/ (accessed Jul. 20, 2022).

[151] D. Doiron *et al.*, “Data harmonization and federated analysis of population-based studies: the BioSHaRE project,” *Emerg. Themes Epidemiol.*, vol. 10, no. 1, p. 12, Nov. 2013, doi: 10.1186/1742-7622-10-12.

[152] D. Doiron, Y. Marcon, I. Fortier, P. Burton, and V. Ferretti, “Software Application Profile: Opal and Mica: open-source software solutions for

epidemiological data management, harmonization and dissemination," *Int. J. Epidemiol.*, vol. 46, no. 5, pp. 1372–1378, Oct. 2017, doi: 10.1093/ije/dyx180.

[153] E. M. Jones, N. A. Sheehan, N. Masca, S. E. Wallace, M. J. Murtagh, and P. R. Burton, "DataSHIELD – shared individual-level analysis without sharing the data: a biostatistical perspective," *Nor. Epidemiol.*, vol. 21, no. 2, Apr. 2012, doi: 10.5324/nje.v21i2.1499.

[154] R. Wilson *et al.*, "DataSHIELD â€" new directions and dimensions," *Data Sci. J.*, vol. 16, 2017.

[155] "SNOMED Home page," *SNOMED*. https://www.snomed.org/ (accessed Jul. 20, 2022).

[156] "ICD-11." https://icd.who.int/en (accessed Jul. 20, 2022).

[157] "Human Phenotype Ontology." https://hpo.jax.org/app/ (accessed Jul. 20, 2022).

[158] The Gene Ontology Consortium, "The Gene Ontology Resource: 20 years and still GOing strong," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D330–D338, Jan. 2019, doi: 10.1093/nar/gky1055.

[159] J. M. Kush, K. E. Masyn, M. Amin-Esmaeili, R. Susukida, H. C. Wilcox, and R. J. Musci, "Utilizing Moderated Non-linear Factor Analysis Models for Integrative Data Analysis: A Tutorial," *Struct. Equ. Model. Multidiscip. J.*, pp. 1–16, May 2022, doi: 10.1080/10705511.2022.2070753.

[160] R. D. Gibbons, M. C. Perraillon, and J. B. Kim, "Item response theory approaches to harmonization and research synthesis," *Health Serv. Outcomes Res. Methodol.*, vol. 14, no. 4, pp. 213–231, Dec. 2014, doi: 10.1007/s10742-014-0125-x.

[161] K. S. Chan, A. L. Gross, L. E. Pezzin, J. Brandt, and J. D. Kasper, "Harmonizing Measures of Cognitive Performance Across International Surveys of Aging Using Item Response Theory," *J. Aging Health*, vol. 27, no. 8, pp. 1392–1414, Dec. 2015, doi: 10.1177/0898264315583054.

[162] T. L. Heafner and P. G. Fitchett, "US history content knowledge and associated effects of race, gender, wealth, and urbanity: Item Response Theory (IRT) modeling of NAEP-USH achievement," *J. Soc. Stud. Res.*, vol. 42, no. 1, pp. 11–25, Jan. 2018, doi: 10.1016/j.jssr.2017.01.001.

[163] S. J. Tannenbaum, N. H. G. Holford, H. Lee, C. C. Peck, and D. R. Mould, "Simulation of Correlated Continuous and Categorical Variables using a Single

Multivariate Distribution," *J. Pharmacokinet. Pharmacodyn.*, vol. 33, no. 6, pp. 773–794, Dec. 2006, doi: 10.1007/s10928-006-9033-1.

[164] D. Teutonico *et al.*, "Generating Virtual Patients by Multivariate and Discrete Re-Sampling Techniques," *Pharm. Res.*, vol. 32, no. 10, pp. 3228–3237, Oct. 2015, doi: 10.1007/s11095-015-1699-x.

[165] R. Allen, T. Rieger, and C. Musante, "Efficient Generation and Selection of Virtual Populations in Quantitative Systems Pharmacology Models," *CPT Pharmacomet. Syst. Pharmacol.*, vol. 5, no. 3, pp. 140–146, Mar. 2016, doi: 10.1002/psp4.12063.

[166] J. D. Silverman, K. Roche, Z. C. Holmes, L. A. David, and S. Mukherjee, "Bayesian Multinomial Logistic Normal Models through Marginally Latent Matrix-T Processes.," *J Mach Learn Res*, vol. 23, p. 7:1-7:42, 2022.

[167] G. Smania and E. N. Jonsson, "Conditional distribution modeling as an alternative method for covariates simulation: Comparison with joint multivariate normal and bootstrap techniques," *CPT Pharmacomet. Syst. Pharmacol.*, vol. 10, no. 4, pp. 330–339, Apr. 2021, doi: 10.1002/psp4.12613.

[168] S. G. Boettcher and C. Dethlefsen, "deal: A Package for Learning Bayesian Networks," *J. Stat. Softw.*, vol. 8, no. 20, pp. 1–40, Dec. 2003, doi: 10.18637/jss.v008.i20.

[169] M. Tsagris, "A New Scalable Bayesian Network Learning Algorithm with Applications to Economics," *Comput. Econ.*, vol. 57, no. 1, pp. 341–367, Jan. 2021, doi: 10.1007/s10614-020-10065-7.

[170] M. Sood *et al.*, "Realistic simulation of virtual multi-scale, multi-modal patient trajectories using Bayesian networks and sparse auto-encoders," *Sci. Rep.*, vol. 10, no. 1, p. 10971, Jul. 2020, doi: 10.1038/s41598-020-67398-4.

[171] M. Krauss and A. Schuppert, "Assessing interindividual variability by Bayesian-PBPK modeling," *Drug Discov. Today Dis. Models*, vol. 22, pp. 15–19, Dec. 2016, doi: 10.1016/j.ddmod.2017.08.001.

[172] M. Robnik-Sikonja, "Data Generators for Learning Systems Based on RBF Networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 5, pp. 926–938, May 2016, doi: 10.1109/TNNLS.2015.2429711.

[173] M. R. Šikonja, "Dataset comparison workflows," *Int. J. Data Sci.*, vol. 3, no. 2, p. 126, 2018, doi: 10.1504/IJDS.2018.092282.

[174] A. Budiarto, B. Mahesworo, A. A. Hidayat, I. Nurlaila, and B. Pardamean,

"Gaussian Mixture Model Implementation for Population Stratification Estimation from Genomics Data," *Procedia Comput. Sci.*, vol. 179, pp. 202–210, 2021, doi: 10.1016/j.procs.2020.12.026.

[175] W. Yuan, B. Eckart, K. Kim, V. Jampani, D. Fox, and J. Kautz, "DeepGMR: Learning Latent Gaussian Mixture Models for Registration," in *Computer Vision – ECCV 2020*, Cham, 2020, pp. 733–750.

[176] T. L. Athey, T. Liu, B. D. Pedigo, and J. T. Vogelstein, "AutoGMM: Automatic and Hierarchical Gaussian Mixture Modeling in Python," 2019, doi: 10.48550/ARXIV.1909.02688.

[177] Y. Lai *et al.*, "Extended variational inference for gamma mixture model in positive vectors modeling," *Neurocomputing*, vol. 432, pp. 145–158, Apr. 2021, doi: 10.1016/j.neucom.2020.12.042.

[178] S. Amudala, S. Ali, F. Najar, and N. Bouguila, "Variational Inference of Finite Generalized Gaussian Mixture Models," in *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, Xiamen, China, Dec. 2019, pp. 2433–2439. doi: 10.1109/SSCI44817.2019.9002852.

[179] O. Arenz, M. Zhong, and G. Neumann, "Efficient Gradient-Free Variational Inference using Policy Search," Jul. 2018.

[180] L. U. Khan, W. Saad, Z. Han, E. Hossain, and C. S. Hong, "Federated Learning for Internet of Things: Recent Advances, Taxonomy, and Open Challenges," *IEEE Commun. Surv. Tutor.*, vol. 23, no. 3, pp. 1759–1799, 2021, doi: 10.1109/COMST.2021.3090430.

[181] A. A. Benczúr, L. Kocsis, and R. Pálovics, "Online Machine Learning in Big Data Streams," 2018, doi: 10.48550/ARXIV.1802.05872.

[182] J. Vanschoren, "Meta-Learning: A Survey," 2018, doi: 10.48550/ARXIV.1810.03548.

[183] S. Sra, S. Nowozin, and S. J. Wright, *Optimization for machine learning*. Cambridge, Mass.: MIT Press, 2012.

[184] Y. Luo, L. Yin, W. Bai, and K. Mao, "An Appraisal of Incremental Learning Methods," *Entropy*, vol. 22, no. 11, Art. no. 11, Nov. 2020, doi: 10.3390/e22111190.

[185] Y. Wu *et al.*, "Large Scale Incremental Learning," 2019, pp. 374–382. Accessed: Jul. 21, 2022. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2019/html/Wu_Large_Scale_Incr

emental_Learning_CVPR_2019_paper.html

[186] J. Lin, C. Zhong, D. Hu, C. Rudin, and M. Seltzer, “Generalized and Scalable Optimal Sparse Decision Trees,” in *Proceedings of the 37th International Conference on Machine Learning*, Nov. 2020, pp. 6150–6160. Accessed: Jul. 21, 2022. [Online]. Available: https://proceedings.mlr.press/v119/lin20g.html

[187] J. Xu, C. Xu, B. Zou, Y. Y. Tang, J. Peng, and X. You, “New Incremental Learning Algorithm With Support Vector Machines,” *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 49, no. 11, pp. 2230–2241, Nov. 2019, doi: 10.1109/TSMC.2018.2791511.

[188] H. J. Kim and J. Chang, “Integrating Incremental Feature Weighting into NaÏve Bayes Text Classifier,” in *2007 International Conference on Machine Learning and Cybernetics*, Aug. 2007, vol. 2, pp. 1137–1143. doi: 10.1109/ICMLC.2007.4370315.

[189] T. Olsson, “Incremental Clustering of Source Code : a Machine Learning Approach,” 2022, Accessed: Jul. 21, 2022. [Online]. Available: http://urn.kb.se/resolve?urn=urn:nbn:se:lnu:diva-110142

[190] E. Jothimurugesan, A. Tahmasbi, P. Gibbons, and S. Tirthapura, “Variance-Reduced Stochastic Gradient Descent on Streaming Data,” in *Advances in Neural Information Processing Systems*, 2018, vol. 31. Accessed: Jul. 21, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2018/hash/cebd648f9146a6345d604ab093b02c73-Abstract.html

[191] P. Zhou, X. Yuan, and J. Feng, “New Insight into Hybrid Stochastic Gradient Descent: Beyond With-Replacement Sampling and Convexity,” in *Advances in Neural Information Processing Systems*, 2018, vol. 31. Accessed: Jul. 21, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2018/hash/67e103b0761e60683e83c559be18d40c-Abstract.html

[192] S. T. Nguyen, H. Y. Kwak, S. Y. Lee, and G. Y. Gim, “Featured Hybrid Recommendation System Using Stochastic Gradient Descent,” *Int. J. Networked Distrib. Comput.*, vol. 9, no. 1, pp. 25–32, Jan. 2021, doi: 10.2991/ijndc.k.201218.004.

[193] R. K. Vinayak and R. Gilad-Bachrach, “DART: Dropouts meet Multiple Additive Regression Trees,” in *Proceedings of the Eighteenth International*

Conference on Artificial Intelligence and Statistics, Feb. 2015, pp. 489–497. Accessed: Jul. 21, 2022. [Online]. Available: https://proceedings.mlr.press/v38/korlakaivinayak15.html

[194] S. Caldas *et al.*, "LEAF: A Benchmark for Federated Settings," 2018, doi: 10.48550/ARXIV.1812.01097.

[195] D. J. Beutel *et al.*, "Flower: A Friendly Federated Learning Research Framework," 2020, doi: 10.48550/ARXIV.2007.14390.

[196] X. Zhu, J. Wang, Z. Hong, T. Xia, and J. Xiao, "Federated Learning of Unsegmented Chinese Text Recognition Model," in *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, Portland, OR, USA, Nov. 2019, pp. 1341–1345. doi: 10.1109/ICTAI.2019.00186.

[197] "PaddlePaddle/PaddleFL." PaddlePaddle, Jul. 20, 2022. Accessed: Jul. 21, 2022. [Online]. Available: https://github.com/PaddlePaddle/PaddleFL

[198] I. Kholod *et al.*, "Open-Source Federated Learning Frameworks for IoT: A Comparative Review and Analysis," *Sensors*, vol. 21, no. 1, Art. no. 1, Jan. 2021, doi: 10.3390/s21010167.

[199] "TensorFlow Federated," *TensorFlow*. https://www.tensorflow.org/federated (accessed Jul. 21, 2022).

[200] C. He *et al.*, "FedML: A Research Library and Benchmark for Federated Machine Learning." arXiv, Nov. 08, 2020. doi: 10.48550/arXiv.2007.13518.

[201] P.-S. Lin, M.-C. Kao, W.-Y. Liang, and S.-H. Hung, "Performance Analysis and Optimization for Federated Learning Applications with PySyft-based Secure Aggregation," in *2020 International Computer Symposium (ICS)*, Dec. 2020, pp. 191–196. doi: 10.1109/ICS51289.2020.00046.

[202] T. M. Deist *et al.*, "Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT," *Clin. Transl. Radiat. Oncol.*, vol. 4, pp. 24–31, Jun. 2017, doi: 10.1016/j.ctro.2016.12.004.

[203] T. M. Deist *et al.*, "Distributed learning on 20 000+ lung cancer patients – The Personal Health Train," *Radiother. Oncol.*, vol. 144, pp. 189–200, Mar. 2020, doi: 10.1016/j.radonc.2019.11.019.

[204] A. Jochems *et al.*, "Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital – A real life proof of concept," *Radiother. Oncol.*, vol. 121, no. 3, pp. 459–467, Dec. 2016, doi:

10.1016/j.radonc.2016.10.002.

[205] “OpenFL - Creative expression for desktop, mobile, web and console platforms.” https://www.openfl.org/ (accessed Jul. 21, 2022).

[206] “PyTorch.” https://www.pytorch.org (accessed Jul. 21, 2022).

[207] “GitHub - PaddlePaddle/Paddle: PArallel Distributed Deep LEarning: Machine Learning Framework from Industrial Practice.” https://github.com/PaddlePaddle/Paddle (accessed Jul. 21, 2022).

[208] L. Collins, H. Hassani, A. Mokhtari, and S. Shakkottai, “FedAvg with Fine Tuning: Local Updates Lead to Representation Learning.” arXiv, May 26, 2022. doi: 10.48550/arXiv.2205.13692.

[209] K. Bonawitz, F. Salehi, J. Konečný, B. McMahan, and M. Gruteser, “Federated Learning with Autotuned Communication-Efficient Secure Aggregation.” arXiv, Nov. 29, 2019. doi: 10.48550/arXiv.1912.00131.

[210] Y. Zhao *et al.*, “Local Differential Privacy-Based Federated Learning for Internet of Things,” *IEEE Internet Things J.*, vol. 8, no. 11, pp. 8836–8853, Jun. 2021, doi: 10.1109/JIOT.2020.3037194.

[211] R. Dobson and G. Giovannoni, “Multiple sclerosis – a review,” *Eur. J. Neurol.*, vol. 26, no. 1, pp. 27–40, 2019, doi: 10.1111/ene.13819.

[212] M. P. McGinley, C. H. Goldschmidt, and A. D. Rae-Grant, “Diagnosis and Treatment of Multiple Sclerosis: A Review,” *JAMA*, vol. 325, no. 8, pp. 765–779, Feb. 2021, doi: 10.1001/jama.2020.26858.

[213] K. D. Deane and V. M. Holers, “Rheumatoid Arthritis Pathogenesis, Prediction, and Prevention: An Emerging Paradigm Shift,” *Arthritis Rheumatol.*, vol. 73, no. 2, pp. 181–193, 2021, doi: 10.1002/art.41417.

[214] Y. Tanaka, “State-of-the-art treatment of systemic lupus erythematosus,” *Int. J. Rheum. Dis.*, vol. 23, no. 4, pp. 465–471, 2020, doi: 10.1111/1756-185X.13817.

[215] A. Dhana, H. Yen, H. Yen, and E. Cho, “All-cause and cause-specific mortality in psoriasis: A systematic review and meta-analysis,” *J. Am. Acad. Dermatol.*, vol. 80, no. 5, pp. 1332–1343, May 2019, doi: 10.1016/j.jaad.2018.12.037.

[216] O. Bonhomme *et al.*, “Biomarkers in systemic sclerosis-associated interstitial lung disease: review of the literature,” *Rheumatology*, vol. 58, no. 9, pp. 1534–1546, Sep. 2019, doi: 10.1093/rheumatology/kez230.

[217] A. C. Dong and A. Stagnaro-Green, “Differences in Diagnostic Criteria Mask the True Prevalence of Thyroid Disease in Pregnancy: A Systematic Review and

Meta-Analysis," *Thyroid*, vol. 29, no. 2, pp. 278–289, Feb. 2019, doi: 10.1089/thy.2018.0475.

[218] M. Sebode, C. Weiler-Normann, T. Liwinski, and C. Schramm, "Autoantibodies in Autoimmune Liver Disease—Clinical and Diagnostic Relevance," *Front. Immunol.*, vol. 9, 2018, Accessed: Jul. 22, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fimmu.2018.00609

[219] R. Ranchal *et al.*, "Disrupting Healthcare Silos: Addressing Data Volume, Velocity and Variety With a Cloud-Native Healthcare Data Ingestion Service," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 11, pp. 3182–3188, Nov. 2020, doi: 10.1109/JBHI.2020.3001518.

[220] A. Margheri, M. Masi, A. Miladi, V. Sassone, and J. Rosenzweig, "Decentralised provenance for healthcare data," *Int. J. Med. Inf.*, vol. 141, p. 104197, Sep. 2020, doi: 10.1016/j.ijmedinf.2020.104197.

[221] "Data silos undermine efforts to characterize, predict, and mitigate dementia-related missing person incidents - Antonio Miguel Cruz, Samantha Marshall, Christine Daum, Hector Perez, John Hirdes, Lili Liu, 2022." https://journals.sagepub.com/doi/full/10.1177/08404704221106156 (accessed Jul. 22, 2022).

[222] A. G. Tzioufas and H. M. Moutsopoulos, "Sjögren's Syndrome," in *European Handbook of Dermatological Treatments*, A. D. Katsambas, T. M. Lotti, C. Dessinioti, and A. M. D'Erme, Eds. Berlin, Heidelberg: Springer, 2015, pp. 883–889. doi: 10.1007/978-3-662-45139-7_89.

[223] F. B. Vivino *et al.*, "Sjogren's syndrome: An update on disease pathogenesis, clinical manifestations and treatment," *Clin. Immunol. Orlando Fla*, vol. 203, pp. 81–121, Jun. 2019, doi: 10.1016/j.clim.2019.04.009.

[224] L. Chatzis, P. G. Vlachoyiannopoulos, A. G. Tzioufas, and A. V. Goules, "New frontiers in precision medicine for Sjogren's syndrome," *Expert Rev. Clin. Immunol.*, vol. 17, no. 2, pp. 127–141, Feb. 2021, doi: 10.1080/1744666X.2021.1879641.

[225] Y. Shapira, N. Agmon-Levin, and Y. Shoenfeld, "Geoepidemiology of autoimmune rheumatic diseases," *Nat. Rev. Rheumatol.*, vol. 6, no. 8, Art. no. 8, Aug. 2010, doi: 10.1038/nrrheum.2010.86.

[226] A. Travaglino *et al.*, "Sjögren Syndrome in Primary Salivary Gland Lymphoma: A Systematic Review and Meta-Analysis," *Am. J. Clin. Pathol.*, vol. 153, no. 6,

pp. 719–724, May 2020, doi: 10.1093/ajcp/aqaa005.

[227] P. Soret *et al.*, "A new molecular classification to drive precision treatment strategies in primary Sjögren's syndrome," *Nat. Commun.*, vol. 12, no. 1, Art. no. 1, Jun. 2021, doi: 10.1038/s41467-021-23472-7.

[228] A. V. Goules and A. G. Tzioufas, "Lymphomagenesis in Sjögren's syndrome: Predictive biomarkers towards precision medicine," *Autoimmun. Rev.*, vol. 18, no. 2, pp. 137–143, Feb. 2019, doi: 10.1016/j.autrev.2018.08.007.

[229] M. Bombardieri *et al.*, "One year in review 2020: Pathogenesis of primary Sjögren's syndrome," *Clin Exp Rheumatol*, vol. 38, no. 4, pp. S3–S9, 2020.

[230] Y. Park, J. Lee, S.-H. Park, and S. Kwok, "Male patients with primary Sjögren's syndrome: A distinct clinical subgroup?," *Int. J. Rheum. Dis.*, vol. 23, no. 10, pp. 1388–1395, 2020, doi: 10.1111/1756-185X.13940.

[231] A. Papageorgiou, M. Voulgarelis, and A. G. Tzioufas, "Clinical picture, outcome and predictive factors of lymphoma in Sjögren syndrome," *Autoimmun. Rev.*, vol. 14, no. 7, pp. 641–649, Jul. 2015, doi: 10.1016/j.autrev.2015.03.004.

[232] R. Solans-Laqué, A. López-Hernandez, J. Angel Bosch-Gil, A. Palacios, M. Campillo, and M. Vilardell-Tarres, "Risk, Predictors, and Clinical Characteristics of Lymphoma Development in Primary Sjögren's Syndrome," *Semin. Arthritis Rheum.*, vol. 41, no. 3, pp. 415–423, Dec. 2011, doi: 10.1016/j.semarthrit.2011.04.006.

[233] L. Chatzis *et al.*, "Sjögren's Syndrome: The Clinical Spectrum of Male Patients," *J. Clin. Med.*, vol. 9, no. 8, Art. no. 8, Aug. 2020, doi: 10.3390/jcm9082620.

[234] L. Chatzis *et al.*, "A biomarker for lymphoma development in Sjogren's syndrome: Salivary gland focus score," *J. Autoimmun.*, vol. 121, p. 102648, Jul. 2021, doi: 10.1016/j.jaut.2021.102648.

[235] S. Fragkioudaki, C. P. Mavragani, and H. M. Moutsopoulos, "Predicting the risk for lymphoma development in Sjogren syndrome: An easy tool for clinical use," *Medicine (Baltimore)*, vol. 95, no. 25, p. e3766, Jun. 2016, doi: 10.1097/MD.0000000000003766.

[236] J. P. A. Ioannidis, V. A. Vassiliou, and H. M. Moutsopoulos, "Long-term risk of mortality and lymphoproliferative disease and predictive classification of primary Sjögren's syndrome," *Arthritis Rheum.*, vol. 46, no. 3, pp. 741–747, 2002, doi: 10.1002/art.10221.

[237] M. R. Hillen, F. Barone, T. R. Radstake, and J. A. van Roon, “Towards standardisation of histopathological assessments of germinal centres and lymphoid structures in primary Sjögren’s syndrome,” *Ann. Rheum. Dis.*, vol. 75, no. 6, pp. e31–e31, Jun. 2016, doi: 10.1136/annrheumdis-2016-209475.

[238] S. De Vita and S. Gandolfo, “Predicting lymphoma development in patients with Sjögren’s syndrome,” *Expert Rev. Clin. Immunol.*, vol. 15, no. 9, pp. 929–938, Sep. 2019, doi: 10.1080/1744666X.2019.1649596.

[239] A. Alunno, M. C. Leone, R. Giacomelli, R. Gerli, and F. Carubbi, “Lymphoma and Lymphomagenesis in Primary Sjögren’s Syndrome,” *Front. Med.*, vol. 5, 2018, Accessed: Jul. 22, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmed.2018.00102

[240] C. Baldini, F. Ferro, E. Elefante, and S. Bombardieri, “Biomarkers for Sjögren’s syndrome,” *Biomark. Med.*, vol. 12, no. 3, pp. 275–286, Mar. 2018, doi: 10.2217/bmm-2017-0297.

[241] A. V. Goules *et al.*, “Sjögren’s syndrome towards precision medicine: the challenge of harmonisation and integration of cohorts,” *Clin Exp Rheumatol*, vol. 37, no. Suppl 118, pp. S175-84, 2019.

[242] A. G. Tzioufas and A. V. Goules, “The necessity of novel biomarkers in primary Sjögren’s syndrome,” *Clin Exp Rheumatol*, vol. 37, no. Suppl 118, pp. 16–8, 2019.

[243] S. Retamozo, P. Brito-Zerón, and M. Ramos-Casals, “Prognostic markers of lymphoma development in primary Sjögren syndrome,” *Lupus*, vol. 28, no. 8, pp. 923–936, Jul. 2019, doi: 10.1177/0961203319857132.

[244] G. Nocturne *et al.*, “Rheumatoid Factor and Disease Activity Are Independent Predictors of Lymphoma in Primary Sjögren’s Syndrome,” *Arthritis Rheumatol.*, vol. 68, no. 4, pp. 977–985, 2016, doi: 10.1002/art.39518.

[245] G. Ingravallo *et al.*, “Primary Breast Extranodal Marginal Zone Lymphoma in Primary Sjögren Syndrome: Case Presentation and Relevant Literature,” *J. Clin. Med.*, vol. 9, no. 12, Art. no. 12, Dec. 2020, doi: 10.3390/jcm9123997.

[246] “Sjögren Syndrome and Cancer - Rheumatic Disease Clinics.” https://www.rheumatic.theclinics.com/article/S0889-857X(20)30056-9/fulltext (accessed Jul. 22, 2022).

[247] A. Papageorgiou *et al.*, “Predicting the Outcome of Sjogren’s Syndrome-Associated Non-Hodgkin’s Lymphoma Patients,” *PLOS ONE*, vol. 10, no. 2, p. e0116189, Feb. 2015, doi: 10.1371/journal.pone.0116189.

[248] C. Baldini, F. Ferro, N. Luciano, S. Bombardieri, and E. Grossi, “Artificial neural networks help to identify disease subsets and to predict lymphoma in primary Sjögren’s syndrome,” *Clin. Exp. Rheumatol.*, vol. 36 Suppl 112, no. 3, pp. 137–144, Jun. 2018.

[249] M. Jiang, Y. Li, C. Jiang, L. Zhao, X. Zhang, and P. E. Lipsky, “Machine Learning in Rheumatic Diseases,” *Clin. Rev. Allergy Immunol.*, vol. 60, no. 1, pp. 96–110, Feb. 2021, doi: 10.1007/s12016-020-08805-6.

[250] V. C. Pezoulas, T. P. Exarchos, A. G. Tzioufas, S. De Vita, and D. I. Fotiadis, “Predicting lymphoma outcomes and risk factors in patients with primary Sjögren’s Syndrome using gradient boosting tree ensembles,” in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Jul. 2019, pp. 2165–2168. doi: 10.1109/EMBC.2019.8857557.

[251] K. Mc Namara, H. Alzubaidi, and J. K. Jackson, “Cardiovascular disease as a leading cause of death: how are pharmacists getting involved?,” *Integr. Pharm. Res. Pract.*, vol. 8, pp. 1–11, Feb. 2019, doi: 10.2147/IPRP.S133088.

[252] M. A. Khan *et al.*, “Global Epidemiology of Ischemic Heart Disease: Results from the Global Burden of Disease Study,” *Cureus*, vol. 12, no. 7, Jul. 2020, doi: 10.7759/cureus.9349.

[253] C. M. Wolf, “Hypertrophic cardiomyopathy: genetics and clinical perspectives,” *Cardiovasc. Diagn. Ther.*, vol. 9, no. Suppl 2, pp. S388–S415, Oct. 2019, doi: 10.21037/cdt.2019.02.01.

[254] “Hypertrophic Cardiomyopathy: Clinical Update | JACC: Heart Failure.” https://www.jacc.org/doi/abs/10.1016/j.jchf.2018.02.010 (accessed Jul. 22, 2022).

[255] C. V. Tuohy, S. Kaul, H. K. Song, B. Nazer, and S. B. Heitner, “Hypertrophic cardiomyopathy: the future of treatment,” *Eur. J. Heart Fail.*, vol. 22, no. 2, pp. 228–240, 2020, doi: 10.1002/ejhf.1715.

[256] M. de O. Antunes and T. L. Scudeler, “Hypertrophic cardiomyopathy,” *Int. J. Cardiol. Heart Vasc.*, vol. 27, p. 100503, Mar. 2020, doi: 10.1016/j.ijcha.2020.100503.

[257] B. J. Maron, E. J. Rowin, and M. S. Maron, “Global Burden of Hypertrophic Cardiomyopathy,” *JACC Heart Fail.*, vol. 6, no. 5, pp. 376–378, May 2018, doi: 10.1016/j.jchf.2018.03.004.

[258] C. Rucinski *et al.*, “A Population-Based Registry of Patients With Inherited

Cardiac Conditions and Resuscitated Cardiac Arrest," *J. Am. Coll. Cardiol.*, vol. 75, no. 21, pp. 2698–2707, Jun. 2020, doi: 10.1016/j.jacc.2020.04.004.

[259] Z. Cheng, T. Fang, J. Huang, Y. Guo, M. Alam, and H. Qian, "Hypertrophic Cardiomyopathy: From Phenotype and Pathogenesis to Treatment," *Front. Cardiovasc. Med.*, vol. 8, p. 722340, Oct. 2021, doi: 10.3389/fcvm.2021.722340.

[260] A. Malhotra and S. Sharma, "Hypertrophic Cardiomyopathy in Athletes," *Eur. Cardiol. Rev.*, vol. 12, no. 2, pp. 80–82, Dec. 2017, doi: 10.15420/ecr.2017:12:1.

[261] M. S. Maron, "Clinical Utility of Cardiovascular Magnetic Resonance in Hypertrophic Cardiomyopathy," *J. Cardiovasc. Magn. Reson.*, vol. 14, no. 1, p. 13, Feb. 2012, doi: 10.1186/1532-429X-14-13.

[262] H. H. Khachfe, H. A. Salhab, M. Y. Fares, and H. M. Khachfe, "Current State of Hypertrophic Cardiomyopathy Clinical Trials," *Glob. Heart*, vol. 14, no. 3, pp. 317–325, Sep. 2019, doi: 10.1016/j.gheart.2019.07.005.

[263] N. Maurizi *et al.*, "Long-term Outcomes of Pediatric-Onset Hypertrophic Cardiomyopathy and Age-Specific Risk Factors for Lethal Arrhythmic Events," *JAMA Cardiol.*, vol. 3, no. 6, pp. 520–525, Jun. 2018, doi: 10.1001/jamacardio.2018.0789.

[264] G. Andries, S. Yandrapalli, S. S. Naidu, and J. A. Panza, "Novel Pharmacotherapy in Hypertrophic Cardiomyopathy," *Cardiol. Rev.*, vol. 26, no. 5, pp. 239–244, Oct. 2018, doi: 10.1097/CRD.0000000000000211.

[265] "Pharmacological treatment options for hypertrophic cardiomyopathy: high time for evidence | European Heart Journal | Oxford Academic." https://academic.oup.com/eurheartj/article/33/14/1724/528196 (accessed Jul. 22, 2022).

[266] C. Dehner, R. Fine, and M. A. Kriegel, "The Microbiome in Systemic Autoimmune Disease – Mechanistic Insights from Recent Studies," *Curr. Opin. Rheumatol.*, vol. 31, no. 2, pp. 201–207, Mar. 2019, doi: 10.1097/BOR.0000000000000574.

[267] A. R. Nogueira and Y. Shoenfeld, "Microbiome and autoimmune diseases: cause and effect relationship," *Curr. Opin. Rheumatol.*, vol. 31, no. 5, pp. 471–474, Sep. 2019, doi: 10.1097/BOR.0000000000000628.

[268] M. F. Konig, "The microbiome in autoimmune rheumatic disease," *Best Pract. Res. Clin. Rheumatol.*, vol. 34, no. 1, p. 101473, Feb. 2020, doi: 10.1016/j.berh.2019.101473.

[269] "WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020." https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020 (accessed Jul. 22, 2022).

[270] D. P. Oran and E. J. Topol, "The Proportion of SARS-CoV-2 Infections That Are Asymptomatic," *Ann. Intern. Med.*, vol. 174, no. 5, pp. 655–662, May 2021, doi: 10.7326/M20-6976.

[271] Z. Wu and J. M. McGoogan, "Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention," *JAMA*, vol. 323, no. 13, pp. 1239–1242, Apr. 2020, doi: 10.1001/jama.2020.2648.

[272] X. Yang *et al.*, "Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study," *Lancet Respir. Med.*, vol. 8, no. 5, pp. 475–481, May 2020, doi: 10.1016/S2213-2600(20)30079-5.

[273] C. M. Petrilli *et al.*, "Factors associated with hospital admission and critical illness among 5279 people with coronavirus disease 2019 in New York City: prospective cohort study," *bmj*, vol. 369, 2020.

[274] E. K. Stokes, "Coronavirus Disease 2019 Case Surveillance — United States, January 22–May 30, 2020," *MMWR Morb. Mortal. Wkly. Rep.*, vol. 69, 2020, doi: 10.15585/mmwr.mm6924e2.

[275] S. Richardson *et al.*, "Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area," *JAMA*, vol. 323, no. 20, pp. 2052–2059, May 2020, doi: 10.1001/jama.2020.6775.

[276] A. B. Docherty *et al.*, "Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study," *bmj*, vol. 369, 2020.

[277] D. S. W. Ting, L. Carin, V. Dzau, and T. Y. Wong, "Digital technology and COVID-19," *Nat. Med.*, vol. 26, no. 4, Art. no. 4, Apr. 2020, doi: 10.1038/s41591-020-0824-5.

[278] H. Dai *et al.*, "Big Data in Cardiology: State-of-Art and Future Prospects," *Front. Cardiovasc. Med.*, vol. 9, 2022, Accessed: Jul. 22, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fcvm.2022.844296

[279] T. Hernandez-Boussard, K. L. Monda, B. C. Crespo, and D. Riskin, “Real world evidence in cardiovascular medicine: ensuring data validity in electronic health record-based studies,” *J. Am. Med. Inform. Assoc.*, vol. 26, no. 11, pp. 1189–1194, Nov. 2019, doi: 10.1093/jamia/ocz119.

[280] T. Vos *et al.*, “Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016,” *The Lancet*, vol. 390, no. 10100, pp. 1211–1259, Sep. 2017, doi: 10.1016/S0140-6736(17)32154-2.

[281] D. Vigo, G. Thornicroft, and R. Atun, “Estimating the true global burden of mental illness,” *Lancet Psychiatry*, vol. 3, no. 2, pp. 171–178, Feb. 2016, doi: 10.1016/S2215-0366(15)00505-2.

[282] W. R. Beardslee, T. R. G. Gladstone, and E. E. O’Connor, “Transmission and Prevention of Mood Disorders Among Children of Affectively Ill Parents: A Review,” *J. Am. Acad. Child Adolesc. Psychiatry*, vol. 50, no. 11, pp. 1098–1109, Nov. 2011, doi: 10.1016/j.jaac.2011.07.020.

[283] I. of Medicine, N. R. Council, D. of B. and S. S. and Education, B. on C. Families Youth, and, and C. on D. Children Parenting Practices, and the Healthy Development of, *Depression in Parents, Parenting, and Children: Opportunities to Improve Identification, Treatment, and Prevention*. National Academies Press, 2009.

[284] D. E. Bloom *et al.*, “The global economic burden of noncommunicable diseases,” Program on the Global Demography of Aging, 2012.

[285] N. H. Liu *et al.*, “Excess mortality in persons with severe mental disorders: a multilevel intervention framework and priorities for clinical practice, policy and research agendas,” *World Psychiatry*, vol. 16, no. 1, pp. 30–40, 2017.

[286] R. F. Krueger *et al.*, “Progress in achieving quantitative classification of psychopathology,” *World Psychiatry*, vol. 17, no. 3, pp. 282–293, 2018, doi: 10.1002/wps.20566.

[287] B. N. Cuthbert and T. R. Insel, “Toward the future of psychiatric diagnosis: the seven pillars of RDoC,” *BMC Med.*, vol. 11, no. 1, p. 126, May 2013, doi: 10.1186/1741-7015-11-126.

[288] G. Greenberg, *The book of woe: The DSM and the unmaking of psychiatry*. Penguin, 2013.

[289] H. Chen, D. Hailey, N. Wang, and P. Yu, “A Review of Data Quality Assessment Methods for Public Health Information Systems,” *Int. J. Environ. Res. Public. Health*, vol. 11, no. 5, Art. no. 5, May 2014, doi: 10.3390/ijerph110505170.

[290] L. Cai and Y. Zhu, “The Challenges of Data Quality and Data Quality Assessment in the Big Data Era,” *Data Sci. J.*, vol. 14, no. 0, Art. no. 0, May 2015, doi: 10.5334/dsj-2015-002.

[291] S. S. Tripathy, R. K. Saxena, and P. K. Gupta, “Comparison of statistical methods for outlier detection in proficiency testing data on analysis of lead in aqueous solution,” *Am. J. Theor. Appl. Stat.*, vol. 2, no. 6, pp. 233–242, 2013.

[292] E. Schubert, A. Zimek, and H.-P. Kriegel, “Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection,” *Data Min. Knowl. Discov.*, vol. 28, no. 1, pp. 190–237, Jan. 2014, doi: 10.1007/s10618-012-0300-z.

[293] K. S. Kannan and S. S. Raj, “Outlier Labeling Methods for Medical Data,” in *Logistics, Supply Chain and Financial Predictive Analytics: Theory and Practices*, K. Deep, M. Jain, and S. Salhi, Eds. Singapore: Springer, 2019, pp. 67–75. doi: 10.1007/978-981-13-0872-7_6.

[294] J. Yang, S. Rahardja, and P. Fränti, “Outlier detection: how to threshold outlier scores?,” in *Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing*, New York, NY, USA, Dec. 2019, pp. 1–6. doi: 10.1145/3371425.3371427.

[295] A. Burinskas and A. Burinskienė, “Discovering event outliers for drug as commercial products,” *World Acad. Sci. Eng. Technol. Int. J. Pharmacol. Pharm. Sci. Spec. J. Issue ICDCEI 2020 Int. Conf. Drug Classif. Econ. Isues Amst. Neth. August 06-07 2020*, vol. 14, no. 8, 2020, Accessed: Jul. 23, 2022. [Online]. Available: https://vb.vgtu.lt/object/elaba:77864966/

[296] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, “Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median,” *J. Exp. Soc. Psychol.*, vol. 49, no. 4, pp. 764–766, Jul. 2013, doi: 10.1016/j.jesp.2013.03.013.

[297] H. J. Motulsky and R. E. Brown, “Detecting outliers when fitting data with nonlinear regression – a new method based on robust nonlinear regression and the false discovery rate,” *BMC Bioinformatics*, vol. 7, no. 1, p. 123, Mar. 2006, doi:

10.1186/1471-2105-7-123.

[298] Z. Cheng, C. Zou, and J. Dong, “Outlier detection using isolation forest and local outlier factor,” in *Proceedings of the Conference on Research in Adaptive and Convergent Systems*, New York, NY, USA, Sep. 2019, pp. 161–168. doi: 10.1145/3338840.3355641.

[299] O. Alghushairy, R. Alsini, T. Soule, and X. Ma, “A Review of Local Outlier Factor Algorithms for Outlier Detection in Big Data Streams,” *Big Data Cogn. Comput.*, vol. 5, no. 1, Art. no. 1, Mar. 2021, doi: 10.3390/bdcc5010001.

[300] C. Li, L. Guo, H. Gao, and Y. Li, “Similarity-Measured Isolation Forest: Anomaly Detection Method for Machine Monitoring Data,” *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021, doi: 10.1109/TIM.2021.3062684.

[301] M. U. Togbe *et al.*, “Anomaly Detection for Data Streams Based on Isolation Forest Using Scikit-Multiflow,” in *Computational Science and Its Applications – ICCSA 2020*, Cham, 2020, pp. 15–30. doi: 10.1007/978-3-030-58811-3_2.

[302] E. S. Dalmaijer, C. L. Nord, and D. E. Astle, “Statistical power for cluster analysis,” *BMC Bioinformatics*, vol. 23, no. 1, p. 205, May 2022, doi: 10.1186/s12859-022-04675-1.

[303] A. P. Daga, A. Fasana, L. Garibaldi, and S. Marchesiello, “On the use of PCA for Diagnostics via Novelty Detection: interpretation, practical application notes and recommendation for use,” in *PHM Society European Conference*, 2020, vol. 5, no. 1, pp. 13–13.

[304] Ö. Alpu, “Using fast minimum covariance determinant estimators for factor analysis in the presence of outliers,” *Sak. Univ. J. Sci.*, vol. 20, no. 3, pp. 701–709, 2016.

[305] M. del Pilar Angeles and A. E. Gamez, “Comparison of methods Hamming Distance, Jaro, and Monge–Elkan,” *DBKDA 2015*, p. 73, 2015.

[306] M. S. Santos, P. H. Abreu, S. Wilk, and J. Santos, “How distance metrics influence missing data imputation with k-nearest neighbours,” *Pattern Recognit. Lett.*, vol. 136, pp. 111–119, 2020.

[307] B. Ristevski and M. Chen, “Big Data Analytics in Medicine and Healthcare,” *J. Integr. Bioinforma.*, vol. 15, no. 3, Sep. 2018, doi: 10.1515/jib-2017-0030.

[308] M. Kang, E. Ko, and T. B. Mersha, “A roadmap for multi-omics data integration using deep learning,” *Brief. Bioinform.*, vol. 23, no. 1, p. bbab454, Jan. 2022, doi: 10.1093/bib/bbab454.

[309] D. Cirillo and A. Valencia, “Big data analytics for personalized medicine,” *Curr. Opin. Biotechnol.*, vol. 58, pp. 161–167, Aug. 2019, doi: 10.1016/j.copbio.2019.03.004.

[310] “Machine Learning in Healthcare Data Analysis: A Survey.” https://www.iomcworld.org/articles/machine-learning-in-healthcare-data-analysis-a-survey-44184.html (accessed Jul. 25, 2022).

[311] R. D. Kush *et al.*, “FAIR data sharing: The roles of common data elements and harmonization,” *J. Biomed. Inform.*, vol. 107, p. 103421, Jul. 2020, doi: 10.1016/j.jbi.2020.103421.

[312] J. M. Kraus *et al.*, “Big data and precision medicine: challenges and strategies with healthcare data,” *Int. J. Data Sci. Anal.*, vol. 6, no. 3, pp. 241–249, Nov. 2018, doi: 10.1007/s41060-018-0095-0.

[313] E. R. Pfaff *et al.*, “Fast Healthcare Interoperability Resources (FHIR) as a Meta Model to Integrate Common Data Models: Development of a Tool and Quantitative Validation Study,” *JMIR Med. Inform.*, vol. 7, no. 4, p. e15199, Oct. 2019, doi: 10.2196/15199.

[314] “The European medical information framework: A novel ecosystem for sharing healthcare data across Europe - Lovestone - 2020 - Learning Health Systems - Wiley Online Library.” https://onlinelibrary.wiley.com/doi/full/10.1002/lrh2.10214 (accessed Jul. 25, 2022).

[315] F. Frommlet, P. Szulc, F. König, and M. Bogdan, “Selecting predictive biomarkers from genomic data,” *PLOS ONE*, vol. 17, no. 6, p. e0269369, Jun. 2022, doi: 10.1371/journal.pone.0269369.

[316] I. Fortier, D. Doiron, P. Burton, and P. Raina, “Invited Commentary: Consolidating Data Harmonization—How to Obtain Quality and Applicability?,” *Am. J. Epidemiol.*, vol. 174, no. 3, pp. 261–264, Aug. 2011, doi: 10.1093/aje/kwr194.

[317] D. Doiron, P. Raina, F. L’Heureux, and I. Fortier, “Facilitating collaborative research: Implementing a platform supporting data harmonization and pooling,” *Nor. Epidemiol.*, vol. 21, no. 2, 2012.

[318] H. Ulrich, S. Germer, A.-K. Kock-Schoppenhauer, J. Kern, M. Lablans, and J. Ingenerf, “A Smart Mapping Editor for Standardised Data Transformation.,” in *MIE*, 2020, pp. 1185–1186.

[319] S. Mate *et al.*, “Pan-European data harmonization for biobanks in ADOPT BBMRI-ERIC,” *Appl. Clin. Inform.*, vol. 10, no. 04, pp. 679–692, 2019.

[320] Y. Wang, J. Qin, and W. Wang, “Efficient approximate entity matching using jaro-winkler distance,” in *International Conference on Web Information Systems Engineering*, 2017, pp. 231–239.

[321] N. Singla and D. Garg, “String matching algorithms and their applicability in various applications,” *Int. J. Soft Comput. Eng.*, vol. 1, no. 6, pp. 218–222, 2012.

[322] I. Harrow *et al.*, “Ontology mapping for semantically enabled applications,” *Drug Discov. Today*, vol. 24, no. 10, pp. 2068–2075, Oct. 2019, doi: 10.1016/j.drudis.2019.05.020.

[323] I. Qasim *et al.*, “A comprehensive review of type-2 fuzzy Ontology,” *Artif. Intell. Rev.*, vol. 53, no. 2, pp. 1187–1206, Feb. 2020, doi: 10.1007/s10462-019-09693-9.

[324] C. Peng, P. Goswami, and G. Bai, “A literature review of current technologies on health data integration for patient-centered health management,” *Health Informatics J.*, vol. 26, no. 3, pp. 1926–1951, Sep. 2020, doi: 10.1177/1460458219892387.

[325] P. Kubben, M. Dumontier, and A. Dekker, “Fundamentals of clinical data science,” 2019.

[326] “Full article: Present and future of semantic web technologies: a research statement.” https://www.tandfonline.com/doi/full/10.1080/1206212X.2019.1570666 (accessed Jul. 25, 2022).

[327] A. Rhayem, M. B. A. Mhiri, and F. Gargouri, “Semantic Web Technologies for the Internet of Things: Systematic Literature Review,” *Internet Things*, vol. 11, p. 100206, Sep. 2020, doi: 10.1016/j.iot.2020.100206.

[328] “World Wide Web Consortium (W3C).” https://www.w3.org/ (accessed Jul. 25, 2022).

[329] “RDF - Semantic Web Standards.” https://www.w3.org/RDF/ (accessed Jul. 25, 2022).

[330] “OWL - Semantic Web Standards.” https://www.w3.org/OWL/ (accessed Jul. 25, 2022).

[331] “Extensible Markup Language (XML) - Αναζήτηση Google.” https://www.google.com/search?q=Extensible+Markup+Language+(XML)&oq=

Extensible+Markup+Language+(XML)&aqs=chrome..69i57j0i512j0i22i30l8.242j0j7&sourceid=chrome&ie=UTF-8 (accessed Jul. 25, 2022).

[332] "Index - FHIR v4.3.0." https://www.hl7.org/fhir/index.html (accessed Jul. 25, 2022).

[333] "HL7 Clinical Document Architecture, Release 2 - PMC." https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1380194/ (accessed Jul. 25, 2022).

[334] "Anatomical Therapeutic Chemical (ATC) Classification." https://www.who.int/tools/atc-ddd-toolkit/atc-classification (accessed Jul. 25, 2022).

[335] "Home," *LOINC*. https://loinc.org/ (accessed Jul. 25, 2022).

[336] "OMOP Common Data Model – OHDSI." https://www.ohdsi.org/data-standardization/the-common-data-model/ (accessed Jul. 25, 2022).

[337] "OHDSI – Observational Health Data Sciences and Informatics." https://www.ohdsi.org/ (accessed Jul. 25, 2022).

[338] "NLTK :: Natural Language Toolkit." https://www.nltk.org/ (accessed Jul. 25, 2022).

[339] K. Merchant and Y. Pande, "NLP Based Latent Semantic Analysis for Legal Text Summarization," in *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Sep. 2018, pp. 1803–1807. doi: 10.1109/ICACCI.2018.8554831.

[340] S. A. Salloum, R. Khan, and K. Shaalan, "A Survey of Semantic Analysis Approaches," in *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020)*, Cham, 2020, pp. 61–70. doi: 10.1007/978-3-030-44289-7_6.

[341] P. Kherwa and P. Bansal, "Topic modeling: a comprehensive review," *EAI Endorsed Trans. Scalable Inf. Syst.*, vol. 7, no. 24, 2020.

[342] B. Berger, M. S. Waterman, and Y. W. Yu, "Levenshtein distance, sequence comparison and biological database search," *IEEE Trans. Inf. Theory*, vol. 67, no. 6, pp. 3287–3294, 2020.

[343] "CORDIS | European Commission." https://cordis.europa.eu/project/id/731944 (accessed Jul. 25, 2022).

[344] J. P. F. Bai, J. C. Earp, and V. C. Pillai, "Translational Quantitative Systems Pharmacology in Drug Development: from Current Landscape to Good

Practices," *AAPS J.*, vol. 21, no. 4, p. 72, Jun. 2019, doi: 10.1208/s12248-019-0339-5.

[345] "Generating Model Integrated Evidence for Generic Drug Development and Assessment - Zhao - 2019 - Clinical Pharmacology & Therapeutics - Wiley Online Library." https://ascpt.onlinelibrary.wiley.com/doi/full/10.1002/cpt.1282 (accessed Jul. 26, 2022).

[346] A. F. Hernandez, "Preface to theme issue on pragmatic and virtual trials: Progress and challenges," *Contemp. Clin. Trials*, vol. 119, p. 106816, Aug. 2022, doi: 10.1016/j.cct.2022.106816.

[347] C. Tang, S. Vishwakarma, W. Li, R. Adve, S. Julier, and K. Chetty, "Augmenting Experimental Data with Simulations to Improve Activity Classification in Healthcare Monitoring," in *2021 IEEE Radar Conference (RadarConf21)*, May 2021, pp. 1–6. doi: 10.1109/RadarConf2147009.2021.9455314.

[348] V. Leclerc, M. Ducher, and N. Bleyzac, "Bayesian Networks: A New Approach to Predict Therapeutic Range Achievement of Initial Cyclosporine Blood Concentration After Pediatric Hematopoietic Stem Cell Transplantation," *Drugs RD*, vol. 18, no. 1, pp. 67–75, Mar. 2018, doi: 10.1007/s40268-017-0223-7.

[349] T. Van Erven and P. Harremos, "Rényi divergence and Kullback-Leibler divergence," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 3797–3820, 2014.

[350] C. Butucea, A. Dubois, M. Kroll, and A. Saumard, "Local differential privacy: Elbow effect in optimal density estimation and adaptation over Besov ellipsoids," *Bernoulli*, vol. 26, no. 3, pp. 1727–1764, 2020.

[351] P. Benner and T. Mach, "Locally optimal block preconditioned conjugate gradient method for hierarchical matrices," *PAMM*, vol. 11, no. 1, pp. 741–742, 2011.

[352] A. Knyazev, *Locally optimal block preconditioned conjugate gradient*. 2019.

[353] B. Le Bars, P. Humbert, L. Oudre, and A. Kalogeratos, "Learning Laplacian matrix from bandlimited graph signals," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2937–2941.

[354] A. A. Vergani and E. Binaghi, "A soft davies-bouldin separation measure," in *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2018, pp. 1–8.

[355] J. Lin, “On the dirichlet distribution,” *Dep. Math. Stat. Queens Univ.*, 2016.

[356] A. G. Bedeian and K. W. Mossholder, “On the use of the coefficient of variation as a measure of diversity,” *Organ. Res. Methods*, vol. 3, no. 3, pp. 285–297, 2000.

[357] T. Chen *et al.*, “Xgboost: extreme gradient boosting,” *R Package Version 04-2*, vol. 1, no. 4, pp. 1–4, 2015.

[358] “XGBoost | Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.” https://dl.acm.org/doi/abs/10.1145/2939672.2939785 (accessed Jul. 26, 2022).

[359] R. E. Schapire, “Explaining adaboost,” in *Empirical inference*, Springer, 2013, pp. 37–52.

[360] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[361] “The optimal combination of feature selection and data discretization: An empirical study - ScienceDirect.” https://www.sciencedirect.com/science/article/pii/S002002551930711X (accessed Jul. 26, 2022).

[362] A. Dhillon and A. Singh, “Machine learning in healthcare data analysis: a survey,” *J. Biol. Todays World*, vol. 8, no. 6, pp. 1–10, 2019.

[363] C. Toh and J. P. Brody, “Applications of machine learning in healthcare,” *Smart Manuf. Artif. Intell. Meets Internet Things*, p. 65, 2021.

[364] A. E. Ezugwu *et al.*, “A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects,” *Eng. Appl. Artif. Intell.*, vol. 110, p. 104743, 2022.

[365] X. Ran, X. Zhou, M. Lei, W. Tepsan, and W. Deng, “A Novel K-Means Clustering Algorithm with a Noise Algorithm for Capturing Urban Hotspots,” *Appl. Sci.*, vol. 11, no. 23, Art. no. 23, Jan. 2021, doi: 10.3390/app112311202.

[366] A. Ghosal, A. Nandy, A. K. Das, S. Goswami, and M. Panday, “A short review on different clustering techniques and their applications,” *Emerg. Technol. Model. Graph.*, pp. 69–83, 2020.

[367] G. Ogbuabor and F. N. Ugwoke, “Clustering algorithm for a healthcare dataset using silhouette score value,” *Int. J. Comput. Sci. Inf. Technol.*, vol. 102, no. 2018, pp. 27–37, 2018.

[368] A. Esteva *et al.*, “A guide to deep learning in healthcare,” *Nat. Med.*, vol. 25, no. 1, pp. 24–29, 2019.

[369] A. Sherstinsky, “Fundamentals of recurrent neural network (RNN) and long

short-term memory (LSTM) network," *Phys. Nonlinear Phenom.*, vol. 404, p. 132306, 2020.

[370] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: analysis, applications, and prospects," *IEEE Trans. Neural Netw. Learn. Syst.*, 2021.

[371] S. K. Zhou *et al.*, "A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises," *Proc. IEEE*, vol. 109, no. 5, pp. 820–838, 2021.

[372] N. Dey, S. Borra, A. S. Ashour, and F. Shi, *Machine learning in bio-signal analysis and diagnostic imaging*. Academic Press, 2018.

[373] Z. S. Ageed *et al.*, "Comprehensive survey of big data mining approaches in cloud systems," *Qubahan Acad. J.*, vol. 1, no. 2, pp. 29–38, 2021.

[374] S. Welten *et al.*, "A Privacy-Preserving Distributed Analytics Platform for Health Care Data," *Methods Inf. Med.*, vol. 61, no. S 1, pp. e1–e11, Jun. 2022, doi: 10.1055/s-0041-1740564.

[375] P. Jain and P. Kar, "Non-convex optimization for machine learning," *Found. Trends® Mach. Learn.*, vol. 10, no. 3–4, pp. 142–363, 2017.

[376] J. Li, Y. Cao, Y. Wang, and H. Xiao, "Online Learning Algorithms for Double-Weighted Least Squares Twin Bounded Support Vector Machines," *Neural Process. Lett.*, vol. 45, no. 1, pp. 319–339, Feb. 2017, doi: 10.1007/s11063-016-9527-9.

[377] "Hybrid kernel identification method based on support vector regression and regularisation network algorithms - Taouali - 2014 - IET Signal Processing - Wiley Online Library." https://ietresearch.onlinelibrary.wiley.com/doi/full/10.1049/iet-spr.2013.0242 (accessed Jul. 26, 2022).

[378] J. Konečný and P. Richtárik, "Semi-Stochastic Gradient Descent Methods," *Front. Appl. Math. Stat.*, vol. 3, 2017, Accessed: Jul. 26, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fams.2017.00009

[379] G.-B. Huang and L. Chen, "Convex incremental extreme learning machine," *Neurocomputing*, vol. 70, no. 16, pp. 3056–3062, Oct. 2007, doi: 10.1016/j.neucom.2007.02.009.

[380] D. P. Bertsekas, "Incremental proximal methods for large scale convex optimization," *Math. Program.*, vol. 129, no. 2, p. 163, Jun. 2011, doi:

10.1007/s10107-011-0472-0.

[381] "Greedy Function Approximation: A Gradient Boosting Machine on JSTOR." https://www.jstor.org/stable/2699986 (accessed Jul. 26, 2022).

[382] "Frontiers | Gradient boosting machines, a tutorial." https://www.frontiersin.org/articles/10.3389/fnbot.2013.00021/full (accessed Jul. 26, 2022).

[383] A. I. Naimi and L. B. Balzer, "Stacked generalization: an introduction to super learning," *Eur. J. Epidemiol.*, vol. 33, no. 5, pp. 459–464, May 2018, doi: 10.1007/s10654-018-0390-z.

[384] "Generating ensembles of heterogeneous classifiers using Stacked Generalization - Sesmero - 2015 - WIREs Data Mining and Knowledge Discovery - Wiley Online Library." https://wires.onlinelibrary.wiley.com/doi/full/10.1002/widm.1143 (accessed Jul. 26, 2022).

[385] K. M. Ting and I. H. Witten, "Issues in Stacked Generalization," May 2011. Accessed: Jul. 26, 2022. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2011arXiv1105.5466T

[386] A. de Haro-García, J. Pérez-Rodríguez, and N. García-Pedrajas, "Combining three strategies for evolutionary instance selection for instance-based learning," *Swarm Evol. Comput.*, vol. 42, pp. 160–172, Oct. 2018, doi: 10.1016/j.swevo.2018.02.022.

[387] N. Shlezinger, J. Whang, Y. C. Eldar, and A. G. Dimakis, "Model-Based Deep Learning." arXiv, Jun. 27, 2021. doi: 10.48550/arXiv.2012.08405.

[388] N. Ketkar, "Stochastic gradient descent," in *Deep learning with Python*, Springer, 2017, pp. 113–132.

[389] V. C. Pezoulas *et al.*, "A federated AI strategy for the classification of patients with Mucosa Associated Lymphoma Tissue (MALT) lymphoma across multiple harmonized cohorts," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2021, pp. 1666–1669.

[390] K. Gokcesu and H. Gokcesu, "Nonconvex Extension of Generalized Huber Loss for Robust Learning and Pseudo-Mode Statistics," *ArXiv Prepr. ArXiv220211141*, 2022.

[391] V. C. Pezoulas, T. P. Exarchos, A. G. Tzioufas, and D. I. Fotiadis, "Multiple additive regression trees with hybrid loss for classification tasks across

heterogeneous clinical data in distributed environments: a case study," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Nov. 2021, pp. 1670–1673. doi: 10.1109/EMBC46164.2021.9629912.

[392] A. Esmaeili and F. Marvasti, "A novel approach to quantized matrix completion using huber loss measure," *IEEE Signal Process. Lett.*, vol. 26, no. 2, pp. 337–341, 2019.

[393] J. Fan, W. Wang, and Y. Zhong, "Robust covariance estimation for approximate factor models," *J. Econom.*, vol. 208, no. 1, pp. 5–22, 2019.

[394] E. Štrumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," *Knowl. Inf. Syst.*, 2013, doi: 10.1007/s10115-013-0679-x.

[395] "protégé." https://protege.stanford.edu/ (accessed Jul. 26, 2022).

[396] L. de Wolff, S. Arends, J. F. van Nimwegen, and H. Bootsma, "Ten years of the ESSDAI: is it fit for purpose," *Clin Exp Rheumatol*, vol. 38, no. Suppl 126, pp. 283–90, 2020.

[397] R. Seror *et al.*, "EULAR Sjögren's Syndrome Patient Reported Index (ESSPRI): development of a consensus patient index for primary Sjögren's syndrome," *Ann. Rheum. Dis.*, vol. 70, no. 6, pp. 968–972, 2011.

[398] "OntoGraf - Protege Wiki." https://protegewiki.stanford.edu/wiki/OntoGraf (accessed Jul. 26, 2022).

[399] W. Zhang, N. Bansback, A. Boonen, A. Young, A. Singh, and A. H. Anis, "Validity of the work productivity and activity impairment questionnaire - general health version in patients with rheumatoid arthritis," *Arthritis Res. Ther.*, vol. 12, no. 5, p. R177, Sep. 2010, doi: 10.1186/ar3141.

[400] K. Tang *et al.*, "The Work Instability Scale for Rheumatoid Arthritis (RA-WIS): Does it work in osteoarthritis?," *Qual. Life Res.*, vol. 19, no. 7, pp. 1057–1068, Sep. 2010, doi: 10.1007/s11136-010-9656-y.

[401] "Welcome to the SHAP documentation — SHAP latest documentation." https://shap.readthedocs.io/en/latest/index.html (accessed Jul. 26, 2022).

[402] "SILICOFCM." https://silicofcm.eu/

[403] M. Horridge, R. Gonçalves, C. Nyulas, and M. Musen, *WebProtégé: A Cloud-Based Ontology Editor*. 2019.

[404] S. Lohmann, V. Link, E. Marbach, and S. Negru, "WebVOWL: Web-based

Visualization of Ontologies," in *Knowledge Engineering and Knowledge Management*, Cham, 2015, pp. 154–158. doi: 10.1007/978-3-319-17966-7_21.

[405] "Home - GEO - NCBI." https://www.ncbi.nlm.nih.gov/geo/ (accessed Jul. 26, 2022).

[406] S. C. Hicks and R. A. Irizarry, "When to use Quantile Normalization?" bioRxiv, p. 012203, Dec. 04, 2014. doi: 10.1101/012203.

[407] S. J. Popper *et al.*, "Gene-expression patterns reveal underlying biological processes in Kawasaki disease," *Genome Biol.*, vol. 8, no. 12, p. R261, Dec. 2007, doi: 10.1186/gb-2007-8-12-r261.

[408] "Identification of a peripheral blood transcriptional biomarker panel associated with operational renal allograft tolerance | PNAS." https://www.pnas.org/doi/abs/10.1073/pnas.0705834104 (accessed Jul. 26, 2022).

[409] A. H. Brachat *et al.*, "Early changes in gene expression and inflammatory proteins in systemic juvenile idiopathic arthritis patients on canakinumab therapy," *Arthritis Res. Ther.*, vol. 19, no. 1, p. 13, Jan. 2017, doi: 10.1186/s13075-016-1212-x.

[410] V. J. Wright *et al.*, "Diagnosis of Kawasaki Disease Using a Minimal Whole-Blood Gene Expression Signature," *JAMA Pediatr.*, vol. 172, no. 10, p. e182293, Oct. 2018, doi: 10.1001/jamapediatrics.2018.2293.

[411] L. T. Hoang *et al.*, "Global gene expression profiling identifies new therapeutic targets in acute Kawasaki disease," *Genome Med.*, vol. 6, no. 11, p. 541, Nov. 2014, doi: 10.1186/s13073-014-0102-6.

[412] "Whole blood transcriptional profiles as a prognostic tool in complete and incomplete Kawasaki Disease | PLOS ONE." https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0197858 (accessed Jul. 26, 2022).

[413] "The genetics of Kawasaki disease - Onouchi - 2018 - International Journal of Rheumatic Diseases - Wiley Online Library." https://onlinelibrary.wiley.com/doi/full/10.1111/1756-185X.13218 (accessed Jul. 26, 2022).

[414] G. R. Brown *et al.*, "Gene: a gene-centered information resource at NCBI," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D36–D42, Jan. 2015, doi: 10.1093/nar/gku1055.

[415] S. Brouard *et al.*, "Identification of a peripheral blood transcriptional biomarker

panel associated with operational renal allograft tolerance," *Proc. Natl. Acad. Sci.*, vol. 104, no. 39, pp. 15448–15453, Sep. 2007, doi: 10.1073/pnas.0705834104.

[416] V. C. Pezoulas *et al.*, "A Multimodal Approach for the Risk Prediction of Intensive Care and Mortality in Patients with COVID-19," *Diagnostics*, vol. 12, no. 1, Art. no. 1, Jan. 2022, doi: 10.3390/diagnostics12010056.

[417] V. C. Pezoulas *et al.*, "ICU admission and mortality classifiers for COVID-19 patients based on subgroups of dynamically associated profiles across multiple timepoints," *Comput. Biol. Med.*, vol. 141, p. 105176, Feb. 2022, doi: 10.1016/j.compbiomed.2021.105176.

[418] T. Haddad *et al.*, "Incorporation of stochastic engineering models as prior information in Bayesian medical device trials," *J. Biopharm. Stat.*, vol. 27, no. 6, pp. 1089–1103, Nov. 2017, doi: 10.1080/10543406.2017.1300907.

[419] T. R. Rieger *et al.*, "Improving the generation and selection of virtual populations in quantitative systems pharmacology models," *Prog. Biophys. Mol. Biol.*, vol. 139, pp. 15–22, 2018, doi: 10.1016/j.pbiomolbio.2018.06.002.

[420] "ARPACK - Arnoldi Package." https://www.caam.rice.edu/software/ARPACK/ (accessed Jul. 26, 2022).

[421] "Algebraic Multigrid — AMGCL 0.0.1 documentation." https://amgcl.readthedocs.io/en/latest/amg_overview.html (accessed Jul. 26, 2022).

[422] "Webdav — Nextcloud latest Developer Manual latest documentation." https://docs.nextcloud.com/server/latest/developer_manual/client_apis/WebDAV/index.html (accessed Jul. 26, 2022).

[423] "XGBoost Documentation — xgboost 1.6.1 documentation." https://xgboost.readthedocs.io/en/stable/ (accessed Jul. 26, 2022).

[424] "Disease Ontology - Institute for Genome Sciences @ University of Maryland." https://disease-ontology.org/ (accessed Jul. 26, 2022).

[425] L. Quartuccio *et al.*, "Biomarkers of lymphoma in Sjögren's syndrome and evaluation of the lymphoma risk in prelymphomatous conditions: Results of a multicenter study," *J. Autoimmun.*, vol. 51, pp. 75–80, Jun. 2014, doi: 10.1016/j.jaut.2013.10.002.

[426] S. De Vita, S. Gandolfo, S. Zandonella Callegher, A. Zabotti, and L. Quartuccio, "The evaluation of disease activity in Sjögren's syndrome based on the degree of MALT involvement: glandular swelling and cryoglobulinaemia compared to

ESSDAI in a cohort study," *Clin. Exp. Rheumatol.*, vol. 36 Suppl 112, no. 3, pp. 150–156, Jun. 2018.

[427] Y. Onouchi *et al.*, "A genome-wide association study identifies three new risk loci for Kawasaki disease," *Nat. Genet.*, vol. 44, no. 5, Art. no. 5, May 2012, doi: 10.1038/ng.2220.

[428] Y.-K. Kim, "Analysis of Circular RNAs in the Coronary Arteries of Patients with Kawasaki Disease," *J. Lipid Atheroscler.*, vol. 8, no. 1, pp. 50–57, May 2019, doi: 10.12997/jla.2019.8.1.50.

[429] X. Lu *et al.*, "Circ_1639 induces cells inflammation responses by sponging miR-122 and regulating TNFRSF13C expression in alcoholic liver disease," *Toxicol. Lett.*, vol. 314, pp. 89–97, Oct. 2019, doi: 10.1016/j.toxlet.2019.07.021.

[430] H. Wan *et al.*, "Probing the Behaviour of Cas1-Cas2 upon Protospacer Binding in CRISPR-Cas Systems using Molecular Dynamics Simulations," *Sci. Rep.*, vol. 9, no. 1, Art. no. 1, Feb. 2019, doi: 10.1038/s41598-019-39616-1.

[431] "Cas1 and the Csy complex are opposing regulators of Cas2/3 nuclease activity | PNAS." https://www.pnas.org/doi/abs/10.1073/pnas.1616395114 (accessed Jul. 26, 2022).

[432] H. Lee, Y. Dhingra, and D. G. Sashital, "The Cas4-Cas1-Cas2 complex mediates precise prespacer processing during CRISPR adaptation," *eLife*, vol. 8, p. e44248, Apr. 2019, doi: 10.7554/eLife.44248.

[433] "Performance of CURB-65, PSI, and APACHE-II for predicting COVID-19 pneumonia severity and mortality - Junnian Chen, Bang Liu, Houwei Du, Hailong Lin, Cunrong Chen, Shanshan Rao, Ranjie Yu, Jingjing Wang, Zhiqiang Xue, Yixian Zhang, Yanghuang Xie, 2021." https://journals.sagepub.com/doi/full/10.1177/20587392211027083 (accessed Sep. 08, 2022).

[434] Y. Han *et al.*, "Lactate dehydrogenase, an independent risk factor of severe COVID-19 patients: a retrospective and observational study," *Aging*, vol. 12, no. 12, pp. 11245–11258, Jun. 2020, doi: 10.18632/aging.103372.

[435] W. Hou, Z. Zhao, A. Chen, H. Li, and T. Q. Duong, "Machining learning predicts the need for escalated care and mortality in COVID-19 patients from clinical variables," *Int. J. Med. Sci.*, vol. 18, no. 8, pp. 1739–1745, Feb. 2021, doi: 10.7150/ijms.51235.

[436] W. Galanter *et al.*, "Predicting clinical outcomes among hospitalized COVID-

19 patients using both local and published models," *BMC Med. Inform. Decis. Mak.*, vol. 21, no. 1, p. 224, Jul. 2021, doi: 10.1186/s12911-021-01576-w.

[437] H. Sun *et al.*, "Risk Factors for Mortality in 244 Older Adults With COVID-19 in Wuhan, China: A Retrospective Study," *J. Am. Geriatr. Soc.*, vol. 68, no. 6, pp. E19–E23, 2020, doi: 10.1111/jgs.16533.

[438] J. Zhang *et al.*, "Risk factors for disease severity, unimprovement, and mortality in COVID-19 patients in Wuhan, China," *Clin. Microbiol. Infect.*, vol. 26, no. 6, pp. 767–772, Jun. 2020, doi: 10.1016/j.cmi.2020.04.012.

[439] A. Di Castelnuovo *et al.*, "Common cardiovascular risk factors and in-hospital mortality in 3,894 patients with COVID-19: survival analysis and machine learning-based findings from the multicentre Italian CORIST Study," *Nutr. Metab. Cardiovasc. Dis.*, vol. 30, no. 11, pp. 1899–1913, Oct. 2020, doi: 10.1016/j.numecd.2020.07.031.

[440] A. J. Singer *et al.*, "Cohort of Four Thousand Four Hundred Four Persons Under Investigation for COVID-19 in a New York Hospital and Predictors of ICU Care and Ventilation," *Ann. Emerg. Med.*, vol. 76, no. 4, pp. 394–404, Oct. 2020, doi: 10.1016/j.annemergmed.2020.05.011.

[441] Z. Zhao *et al.*, "Prediction model and risk scores of ICU admission and mortality in COVID-19," *PLOS ONE*, vol. 15, no. 7, p. e0236618, Jul. 2020, doi: 10.1371/journal.pone.0236618.

[442] "Full article: Clinical and inflammatory features based machine learning model for fatal risk prediction of hospitalized COVID-19 patients: results from a retrospective cohort study." https://www.tandfonline.com/doi/full/10.1080/07853890.2020.1868564 (accessed Sep. 08, 2022).

[443] Y. Chen *et al.*, "A Multimodality Machine Learning Approach to Differentiate Severe and Nonsevere COVID-19: Model Development and Validation," *J. Med. Internet Res.*, vol. 23, no. 4, p. e23948, Apr. 2021, doi: 10.2196/23948.

[444] D. Assaf *et al.*, "Utilization of machine-learning models to accurately predict the risk for critical COVID-19," *Intern. Emerg. Med.*, vol. 15, no. 8, pp. 1435–1443, Nov. 2020, doi: 10.1007/s11739-020-02475-0.

[445] M. Wu *et al.*, "Clinical evaluation of potential usefulness of serum lactate dehydrogenase (LDH) in 2019 novel coronavirus (COVID-19) pneumonia," *Respir. Res.*, vol. 21, no. 1, p. 171, Jul. 2020, doi: 10.1186/s12931-020-01427-8.

[446] G. Bousquet *et al.*, “ADL-dependency, D-Dimers, LDH and absence of anticoagulation are independently associated with one-month mortality in older inpatients with Covid-19,” *Aging*, vol. 12, no. 12, pp. 11306–11313, Jun. 2020, doi: 10.18632/aging.103583.

[447] C. Li *et al.*, “Elevated Lactate Dehydrogenase (LDH) level as an independent risk factor for the severity and mortality of COVID-19,” *Aging*, vol. 12, no. 15, pp. 15670–15681, Aug. 2020, doi: 10.18632/aging.103770.

[448] B. M. Henry *et al.*, “Lactate dehydrogenase levels predict coronavirus disease 2019 (COVID-19) severity and mortality: A pooled analysis,” *Am. J. Emerg. Med.*, vol. 38, no. 9, pp. 1722–1726, Sep. 2020, doi: 10.1016/j.ajem.2020.05.073.

[449] R. Mardani *et al.*, “Laboratory Parameters in Detection of COVID-19 Patients with Positive RT-PCR; a Diagnostic Accuracy Study,” *Arch. Acad. Emerg. Med.*, vol. 8, no. 1, p. e43, Apr. 2020.

[450] I.-E. Galani *et al.*, “Untuned antiviral immunity in COVID-19 revealed by temporal type I/III interferon patterns and flu comparison,” *Nat. Immunol.*, vol. 22, no. 1, Art. no. 1, Jan. 2021, doi: 10.1038/s41590-020-00840-x.

[451] H. Han *et al.*, “Profiling serum cytokines in COVID-19 patients reveals IL-6 and IL-10 are disease severity predictors,” *Emerg. Microbes Infect.*, vol. 9, no. 1, pp. 1123–1130, Jan. 2020, doi: 10.1080/22221751.2020.1770129.

[452] C. Zhang, Z. Wu, J.-W. Li, H. Zhao, and G.-Q. Wang, “Cytokine release syndrome in severe COVID-19: interleukin-6 receptor antagonist tocilizumab may be the key to reduce mortality,” *Int. J. Antimicrob. Agents*, vol. 55, no. 5, p. 105954, May 2020, doi: 10.1016/j.ijantimicag.2020.105954.

[453] S. Li *et al.*, “Clinical and pathological investigation of patients with severe COVID-19,” *JCI Insight*, vol. 5, no. 12, p. e138070, doi: 10.1172/jci.insight.138070.

[454] W. Chen, K. I. Zheng, S. Liu, Z. Yan, C. Xu, and Z. Qiao, “Plasma CRP level is positively associated with the severity of COVID-19,” *Ann. Clin. Microbiol. Antimicrob.*, vol. 19, no. 1, p. 18, May 2020, doi: 10.1186/s12941-020-00362-2.

[455] W. Shang *et al.*, “The value of clinical parameters in predicting the severity of COVID-19,” *J. Med. Virol.*, vol. 92, no. 10, pp. 2188–2192, 2020, doi: 10.1002/jmv.26031.

[456] B. R. Sahu, R. K. Kampa, A. Padhi, and A. K. Panda, “C-reactive protein: A promising biomarker for poor prognosis in COVID-19 infection,” *Clin. Chim.*

Acta, vol. 509, pp. 91–94, Oct. 2020, doi: 10.1016/j.cca.2020.06.013.

[457] T. Herold *et al.*, "Elevated levels of IL-6 and CRP predict the need for mechanical ventilation in COVID-19," *J. Allergy Clin. Immunol.*, vol. 146, no. 1, pp. 128-136.e4, Jul. 2020, doi: 10.1016/j.jaci.2020.05.008.

[458] B. Gallo Marin *et al.*, "Predictors of COVID-19 severity: A literature review," *Rev. Med. Virol.*, vol. 31, no. 1, p. e2146, 2021, doi: 10.1002/rmv.2146.

[459] C. Leung, "Risk factors for predicting mortality in elderly patients with COVID-19: A review of clinical data in China," *Mech. Ageing Dev.*, vol. 188, p. 111255, Jun. 2020, doi: 10.1016/j.mad.2020.111255.

[460] S. K. Dhar, V. K, S. Damodar, S. Gujar, and M. Das, "IL-6 and IL-10 as predictors of disease severity in COVID-19 patients: results from meta-analysis and regression," *Heliyon*, vol. 7, no. 2, p. e06155, Feb. 2021, doi: 10.1016/j.heliyon.2021.e06155.

[461] J. Tian *et al.*, "Clinical characteristics and risk factors associated with COVID-19 disease severity in patients with cancer in Wuhan, China: a multicentre, retrospective, cohort study," *Lancet Oncol.*, vol. 21, no. 7, pp. 893–903, Jul. 2020, doi: 10.1016/S1470-2045(20)30309-0.

[462] J. Wu, L. Shi, P. Zhang, Y. Wang, and H. Yang, "Is creatinine an independent risk factor for predicting adverse outcomes in COVID-19 patients?," *Transpl. Infect. Dis.*, vol. 23, no. 2, p. e13539, Apr. 2021, doi: 10.1111/tid.13539.

[463] A.-P. Yang, J. Liu, W. Tao, and H. Li, "The diagnostic and predictive role of NLR, d-NLR and PLR in COVID-19 patients," *Int. Immunopharmacol.*, vol. 84, p. 106504, Jul. 2020, doi: 10.1016/j.intimp.2020.106504.

[464] L. A. Potempa, I. M. Rajab, P. C. Hart, J. Bordon, and R. Fernandez-Botran, "Insights into the Use of C-Reactive Protein as a Diagnostic Index of Disease Severity in COVID-19 Infections," *Am. J. Trop. Med. Hyg.*, vol. 103, no. 2, pp. 561–563, Aug. 2020, doi: 10.4269/ajtmh.20-0473.

[465] O. Pozdnyakova, N. T. Connell, E. M. Battinelli, J. M. Connors, G. Fell, and A. S. Kim, "Clinical Significance of CBC and WBC Morphology in the Diagnosis and Clinical Course of COVID-19 Infection," *Am. J. Clin. Pathol.*, vol. 155, no. 3, pp. 364–375, Mar. 2021, doi: 10.1093/ajcp/aqaa231.

[466] T. A. Khartabil, H. Russcher, A. van der Ven, and Y. B. de Rijke, "A summary of the diagnostic and prognostic value of hemocytometry markers in COVID-19 patients," *Crit. Rev. Clin. Lab. Sci.*, vol. 57, no. 6, pp. 415–431, Aug. 2020, doi:

10.1080/10408363.2020.1774736.

[467] X. Li *et al.*, "Risk factors for severity and mortality in adult COVID-19 inpatients in Wuhan," *J. Allergy Clin. Immunol.*, vol. 146, no. 1, pp. 110–118, Jul. 2020, doi: 10.1016/j.jaci.2020.04.006.

[468] Y. Liu *et al.*, "Neutrophil-to-lymphocyte ratio as an independent risk factor for mortality in hospitalized patients with COVID-19," *J. Infect.*, vol. 81, no. 1, pp. e6–e12, Jul. 2020, doi: 10.1016/j.jinf.2020.04.002.

[469] Y. Wang, L. Shi, Y. Wang, and H. Yang, "An updated meta-analysis of AST and ALT levels and the mortality of COVID-19 patients," *Am. J. Emerg. Med.*, vol. 40, pp. 208–209, Feb. 2021, doi: 10.1016/j.ajem.2020.05.063.

[470] X. Bi *et al.*, "Prediction of severe illness due to COVID-19 based on an analysis of initial Fibrinogen to Albumin Ratio and Platelet count," *Platelets*, vol. 31, no. 5, pp. 674–679, Jul. 2020, doi: 10.1080/09537104.2020.1760230.

[471] L. Y. C. Chen, R. L. Hoiland, S. Stukas, C. L. Wellington, and M. S. Sekhon, "Assessing the importance of interleukin-6 in COVID-19," *Lancet Respir. Med.*, vol. 9, no. 2, p. e13, Feb. 2021, doi: 10.1016/S2213-2600(20)30600-7.

[472] P. C. Robinson, D. Richards, H. L. Tanner, and M. Feldmann, "Accumulating evidence suggests anti-TNF therapy needs to be given trial priority in COVID-19 treatment," *Lancet Rheumatol.*, vol. 2, no. 11, pp. e653–e655, Nov. 2020, doi: 10.1016/S2665-9913(20)30309-X.

[473] B. Tomar, H.-J. Anders, J. Desai, and S. R. Mulay, "Neutrophils and Neutrophil Extracellular Traps Drive Necroinflammation in COVID-19," *Cells*, vol. 9, no. 6, Art. no. 6, Jun. 2020, doi: 10.3390/cells9061383.

[474] S. Wang, L. Fu, K. Huang, J. Han, R. Zhang, and Z. Fu, "Neutrophil-to-lymphocyte ratio on admission is an independent risk factor for the severity and mortality in patients with coronavirus disease 2019," *J. Infect.*, vol. 82, no. 2, pp. e16–e18, Feb. 2021, doi: 10.1016/j.jinf.2020.09.022.

[475] Z. Wang, Z. Du, and F. Zhu, "Glycosylated hemoglobin is associated with systemic inflammation, hypercoagulability, and prognosis of COVID-19 patients," *Diabetes Res. Clin. Pract.*, vol. 164, p. 108214, Jun. 2020, doi: 10.1016/j.diabres.2020.108214.

[476] A. Zinellu, P. Paliogiannis, C. Carru, and A. A. Mangoni, "INR and COVID-19 severity and mortality: A systematic review with meta-analysis and meta-regression," *Adv. Med. Sci.*, vol. 66, no. 2, pp. 372–380, Sep. 2021, doi:

10.1016/j.advms.2021.07.009.

[477] E.-M. Cordeanu *et al.*, “Prognostic Value of Troponin Elevation in COVID-19 Hospitalized Patients,” *J. Clin. Med.*, vol. 9, no. 12, Art. no. 12, Dec. 2020, doi: 10.3390/jcm9124078.

[478] P. Santus *et al.*, “Severity of respiratory failure at admission and in-hospital mortality in patients with COVID-19: a prospective observational multicentre study,” *BMJ Open*, vol. 10, no. 10, p. e043651, Oct. 2020, doi: 10.1136/bmjopen-2020-043651.

[479] T. Mikami *et al.*, “Risk Factors for Mortality in Patients with COVID-19 in New York City,” *J. Gen. Intern. Med.*, vol. 36, no. 1, pp. 17–26, Jan. 2021, doi: 10.1007/s11606-020-05983-z.

[480] F. Mejía *et al.*, “Oxygen saturation as a predictor of mortality in hospitalized adult patients with COVID-19 in a public hospital in Lima, Peru,” *PLOS ONE*, vol. 15, no. 12, p. e0244171, Dec. 2020, doi: 10.1371/journal.pone.0244171.

[481] J. W. Goodall *et al.*, “Risk factors for severe disease in patients admitted with COVID-19 to a hospital in London, England: a retrospective cohort study,” *Epidemiol. Infect.*, vol. 148, p. e251, ed 2020, doi: 10.1017/S0950268820002472.

[482] P. Sharma and A. Kumar, “Metabolic dysfunction associated fatty liver disease increases risk of severe Covid-19,” *Diabetes Metab. Syndr. Clin. Res. Rev.*, vol. 14, no. 5, pp. 825–827, Sep. 2020, doi: 10.1016/j.dsx.2020.06.013.

[483] S. Dash and U. Subudhi, “Multiple power quality event detection and classification using a modified S-transform and WOA tuned SVM classifier,” *Int. J. Power Energy Convers.*, vol. 12, no. 4, pp. 338–363, Jan. 2021, doi: 10.1504/IJPEC.2021.118050.

[484] F.-Y. Cheng *et al.*, “Using Machine Learning to Predict ICU Transfer in Hospitalized COVID-19 Patients,” *J. Clin. Med.*, vol. 9, no. 6, Art. no. 6, Jun. 2020, doi: 10.3390/jcm9061668.

[485] T. Dan *et al.*, “Machine Learning to Predict ICU Admission, ICU Mortality and Survivors’ Length of Stay among COVID-19 Patients: Toward Optimal Allocation of ICU Resources,” in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Dec. 2020, pp. 555–561. doi: 10.1109/BIBM49941.2020.9313292.

[486] R. Aznar-Gimeno *et al.*, “A Clinical Decision Web to Predict ICU Admission or Death for Patients Hospitalised with COVID-19 Using Machine Learning

Algorithms," *Int. J. Environ. Res. Public. Health*, vol. 18, no. 16, Art. no. 16, Jan. 2021, doi: 10.3390/ijerph18168677.

[487] F. T. Fernandes, T. A. de Oliveira, C. E. Teixeira, A. F. de M. Batista, G. Dalla Costa, and A. D. P. Chiavegatto Filho, "A multipurpose machine learning approach to predict COVID-19 negative prognosis in São Paulo, Brazil," *Sci. Rep.*, vol. 11, no. 1, Art. no. 1, Feb. 2021, doi: 10.1038/s41598-021-82885-y.

[488] X. Guan *et al.*, "Clinical and inflammatory features based machine learning model for fatal risk prediction of hospitalized COVID-19 patients: results from a retrospective cohort study," *Ann. Med.*, vol. 53, no. 1, pp. 257–266, Jan. 2021, doi: 10.1080/07853890.2020.1868564.

[489] "Bayesian Augmented Clinical Trials in TB Therapeutic Vaccination - PMC." https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8757686/ (accessed Jul. 26, 2022).

[490] "MLlib | Apache Spark." https://spark.apache.org/mllib/ (accessed Jul. 26, 2022).

Appendix

Supplementary Table 1. Performance evaluation of the virtually generated data in pSS for the UTE, STE and supervised RBF-based ANNs.

Feature	Real	Unsupervised tree ensembles			Supervised tree ensembles			Supervised RBF		
	Mean	Mean	GOF	KL	Mean	GOF	KL	Mean	GOF	KL
Sex	0.95	0.94	0.01	1.13E-10	0.95	0.01	3.25E-11	0.94	0.01	5.43E-11
Dry_mouth	0.95	0.96	0.01	2.16E-10	0.96	0.01	8.50E-11	0.91	0.03	1.23E-09
Dry_eyes	0.92	0.93	0.00	8.56E-12	0.92	0.01	4.53E-11	0.91	0.01	1.73E-10
Abnormal_Shirmer	0.86	0.87	0.01	2.35E-10	0.86	0.00	5.87E-14	0.86	0.00	1.53E-11
ANA	0.86	0.86	0.00	3.13E-13	0.87	0.01	1.98E-10	0.83	0.03	1.06E-09
RF	0.64	0.66	0.02	3.98E-10	0.68	0.04	1.59E-09	0.67	0.03	7.79E-10
Anti_Ro	0.73	0.71	0.02	3.42E-10	0.77	0.03	1.19E-09	0.70	0.03	1.06E-09
Anti_La	0.40	0.36	0.04	2.12E-09	0.39	0.01	1.96E-10	0.37	0.03	1.24E-09
FS_1st_biopsy	2.04	2.00	0.06	3.91E-04	1.99	0.05	3.91E-04	2.03	0.27	3.88E-04
Tarpley	2.34	2.36	0.03	5.11E-09	2.33	0.02	3.81E-09	2.32	0.11	1.50E-07
SGE	0.33	0.34	0.01	3.58E-10	0.35	0.02	9.85E-10	0.33	0.00	4.65E-14
Raynaud	0.26	0.24	0.02	1.35E-09	0.25	0.02	7.44E-10	0.25	0.01	3.21E-10
Ro_La	0.74	0.71	0.03	8.22E-10	0.76	0.02	4.16E-10	0.70	0.04	1.91E-09
RF	0.63	0.61	0.02	3.44E-10	0.63	0.00	6.32E-12	0.64	0.00	2.00E-11
Monoc_gammopathy	0.06	0.07	0.01	2.98E-09	0.07	0.01	3.50E-09	0.07	0.01	2.49E-09
LOW_C4	0.41	0.40	0.01	2.22E-10	0.39	0.03	9.72E-10	0.34	0.07	6.74E-09
Dyspareunia	0.06	0.07	0.01	1.10E-09	0.07	0.01	1.10E-09	0.06	0.01	1.67E-09
Dry_skin	0.04	0.03	0.02	2.73E-08	0.02	0.02	4.17E-08	0.03	0.01	6.21E-09
Dry_upper_resp	0.13	0.08	0.05	2.44E-08	0.10	0.03	9.07E-09	0.11	0.02	3.82E-09
Abnormal_Schirmer	0.89	0.89	0.00	1.25E-11	0.85	0.03	1.31E-09	0.87	0.02	3.43E-10
Abnormal_BUT	0.87	0.84	0.06	6.94E-06	0.84	0.04	6.93E-06	0.77	0.45	7.32E-06
Chronic_Fatigue	0.22	0.21	0.01	1.19E-10	0.19	0.03	2.44E-09	0.24	0.02	1.18E-09
Arthralgias_myalgias	0.66	0.68	0.02	2.46E-10	0.69	0.03	7.08E-10	0.65	0.01	6.75E-11
Arthritis	0.11	0.15	0.04	1.22E-08	0.13	0.02	3.62E-09	0.14	0.03	6.15E-09
Raynaud_phen	0.28	0.27	0.01	3.82E-10	0.27	0.01	1.86E-10	0.26	0.02	9.95E-10
Palpable_purpura	0.12	0.13	0.00	9.48E-11	0.12	0.00	1.77E-11	0.14	0.02	2.39E-09
Vasculitic_ulcer	0.01	0.00	0.01	8.23E-08	0.01	0.00	8.35E-12	0.02	0.01	7.94E-08
Other_rash	0.12	0.10	0.02	4.86E-09	0.11	0.01	1.35E-09	0.13	0.00	1.77E-10
Myositis	0.01	0.01	0.00	6.91E-09	0.01	0.00	2.49E-09	0.02	0.01	7.28E-08
PNS_entrapment	0.04	0.04	0.00	1.02E-10	0.07	0.02	2.08E-08	0.07	0.03	3.15E-08
PNS_vasculitic	0.02	0.02	0.00	8.10E-10	0.02	0.00	1.77E-09	0.04	0.02	3.98E-08
CNS_involvmt	0.01	0.01	0.00	4.53E-09	0.01	0.00	2.49E-08	0.02	0.01	7.94E-08
Psychiatric	0.03	0.02	0.02	5.09E-08	0.01	0.02	1.12E-07	0.02	0.01	2.07E-08
Lymphadenopathy	0.11	0.11	0.00	6.42E-11	0.11	0.00	5.33E-11	0.12	0.01	8.11E-10
Splenomegaly	0.01	0.01	0.00	1.89E-10	0.01	0.00	3.23E-09	0.02	0.01	2.69E-08
Liver_cholangitis	0.03	0.03	0.00	1.00E-11	0.03	0.01	3.17E-09	0.04	0.01	1.62E-08
Liver_hepatitis	0.01	0.01	0.00	1.28E-08	0.02	0.01	6.33E-08	0.03	0.02	1.99E-07
Liver_PBC	0.01	0.01	0.00	3.23E-09	0.00	0.01	2.47E-07	0.04	0.02	9.82E-08
Lung_interstitial_Type	0.04	0.02	0.02	3.01E-08	0.03	0.01	1.85E-08	0.04	0.00	6.49E-10
Lung_bronchocentric	0.03	0.02	0.01	7.93E-09	0.03	0.00	1.30E-09	0.04	0.01	1.23E-08
Lung_pleurisy	0.01	0.02	0.01	3.59E-08	0.02	0.01	2.52E-08	0.04	0.03	1.35E-07
nephrocalcinosis	0.01	0.01	0.00	1.09E-09	0.02	0.01	6.44E-08	0.03	0.02	1.41E-07
Urine_gravity	1014.76	1017.1	0.04	5.65E-01	1017.54	0.04	5.65E-01	1017.3	0.20	5.65E-01
Renal_disease	0.02	0.00	0.02	2.62E-05	0.00	0.02	2.62E-05	0.00	0.02	2.62E-05
Urine_pH	5.91	5.90	0.06	8.74E-09	5.88	0.04	5.86E-09	5.90	0.29	6.80E-07
Kidney_infiltrates	0.01	0.00	0.01	2.53E-05	0.00	0.01	2.53E-05	0.00	0.01	2.53E-05
Kidney_GN_biopsy	0.02	0.01	0.01	2.03E-08	0.01	0.01	2.88E-08	0.03	0.01	1.93E-08
Heart_valvular	0.02	0.04	0.01	2.25E-08	0.04	0.02	4.28E-08	0.05	0.02	5.55E-08
Heart_pericardial	0.00	0.01	0.01	6.22E-08	0.02	0.01	1.30E-07	0.03	0.02	2.47E-07
Heart_CMR	0.00	0.00	0.00	1.19E-07	0.01	0.00	6.24E-09	0.02	0.02	2.22E-07
Esophagus_GER	0.03	0.04	0.01	4.22E-09	0.05	0.02	2.18E-08	0.06	0.03	4.17E-08
WB	0.02	0.01	0.01	3.99E-08	0.02	0.00	2.00E-10	0.03	0.01	1.34E-08
wbc_baseline	5610.23	5603.2	0.03	5.83E-06	5450.39	0.06	5.83E-06	5659.8	0.16	5.83E-06
Neutro_Number	3318.41	3264.6	0.06	6.38E-01	3121.11	0.08	6.38E-01	3366.7	0.20	6.38E-01
Lympho_Number	1695.40	1771.4	0.07	4.47E-01	1695.95	0.06	8.71E-01	1759.6	0.14	8.71E-01
PLT	247874	243238	0.07	-1.1E-16	242138	0.06	-7.30E-17	322854	0.31	-5.04E-17
HGB	13.26	13.38	0.04	6.66E-04	13.38	0.04	6.23E-04	14.09	0.27	9.09E-04
ESR	32.37	35.67	0.09	1.40E-03	33.93	0.09	1.36E-03	37.37	0.17	1.91E-03
CRP	0.14	0.13	0.01	7.29E-10	0.10	0.04	1.30E-08	0.13	0.02	2.61E-09
g_globulins	0.74	0.76	0.04	5.67E-08	0.78	0.04	1.13E-09	0.76	0.07	1.73E-07
Anti_TPO	0.14	0.13	0.01	1.20E-09	0.15	0.01	6.65E-10	0.14	0.00	1.08E-11
C3	111.57	121.09	0.04	1.48E-01	121.57	0.05	4.18E-02	145.27	0.31	1.17E-02
C4	21.66	22.63	0.05	5.06E-03	22.59	0.05	6.01E-03	25.08	0.22	1.21E-02
Cryo	0.72	0.53	0.04	7.40E-02	0.54	0.03	7.44E-02	2.34	0.40	7.50E-02
Lymphoma	0.17	0.16	0.01	1.82E-10	0.16	0.01	1.82E-10	0.15	0.02	1.13E-09

Supplementary Table 2. A summary of the potentially matched terminologies between OCVD and UHEI_LURIC.

Terminologies from TAUH	Terminologies from UHEI_LURIC	Score
Triglycerides in very small VLDL; mmol/l	triglycerides	0.741
Calcium channel blocker agent	calcium	0.605
Cholesterol esters to total lipids ratio in chylomicrons and extremely large VLDL; %	cholesterol ester	0.711
Heart failure	NYHA heart failure classification	0.734
Diastolic KV blood pressure 3rd measurement	diastolic blood pressure measurement 3	0.847
Carotid compliance	oral contraception treatment	0.640
Waist circumference; cm	waist circumference	0.910
Date of laboratory sample	date of …	0.699
Triglycerides in very large HDL; mmol/l	triglycerides	0.744
Myocardial infarction	myocardial infarction	0.968
Free cholesterol in IDL; mmol/l	free cholesterol	0.807
Diabetes; father, I1, 218, 1 = no, 2 = yes, 9 = mother is missing from family	Diabetes	0.701
Mother: coronary angioplasty, 1 = no, 2 = yes	coronary angioplasty	0.698
Length of breastfeeding	no of fathers siblings	0.672
Interleukin-18	interleukin 1	0.877
Total cholesterol to total lipids ratio in small LDL; %	total cholesterol	0.723
Interleukin-15	interleukin 1	0.877
Triglycerides to total lipids ratio in very large VLDL; %	triglycerides	0.711
Diabetes type 2	Diabetes	0.844
Diuretic agent	diuretic	0.792
Diabetes, type 1	Diabetes	0.833
Triglycerides in small LDL; mmol/l	triglycerides	0.759
ACE inhibitors, 0=no, 1=yes	ACE	0.704
calcium channel, voltage-dependent, beta 2	calcium	0.722
Height	height	0.889
Vitamin C mg, F6.2,1275	vitamin C	0.746
Total cholesterol average, F5.2,1154	total cholesterol	0.774
Father: myocardial infarction, 1 = no, 2 = yes	myocardial infarction	0.692
Actual systolic blood pressure, 3rd measurement	systolic blood pressure measurement 3	0.762
Cholesterol esters in small HDL; mmol/l	cholesterol ester	0.784

Terminologies from TAUH	Terminologies from UHEI_LURIC	Score
Phospholipids in medium VLDL; mmol/l	phospholipid	0.604
Interleukin-4	interleukin 4	0.897
Triglysceride,	triglycerides	0.830
Magnesium mg	magnesium	0.852
Sodium intake residual value from regression model standardized by energy in 2007	sodium	0.632
Free testosterone counted by Nanjee-Wheelerin formula, pmol/L:, ftn07=	Testosteron	0.601
Free testosterone counted by Vermeulen´ formula; pmol/L:	Testosteron	0.613
HDL-cholesterol	HDL-cholesterol	1.000
Sodium mg, F7.2,1463	sodium	0.694
Sodium intake residual value from regression model standardized by energy in 1986	sodium	0.632
Triglycerides in large HDL; mmol/l	triglycerides	0.759
Interleukin-6	interleukin 6	0.897
Calcium mg, F8.2, 590	calcium	0.631
Father: coronary angioplasty, I1,, 1 = no, 2 = yes	coronary angioplasty	0.658
Body mass index; weight/	body mass index	0.839
Total cholesterol to total lipids ratio in very large HDL; %	total cholesterol	0.715
Left main body stenosis	abdominal obesity	0.610
Diastolic KV blood pressure, 1st measurement	diastolic blood pressure measurement 1	0.790
Cholesterol esters in chylomicrons and extremely large VLDL; mmol/l	cholesterol ester	0.727
Phospholipids in small HDL; mmol/l	phospholipid	0.610
Triglycerides to total lipids ratio in small VLDL; %	triglycerides	0.718
Potassium mg, F4.2,1335	potassium	0.746
LDL cholesterol, direct; mmol/l	LDL cholesterol	0.828
Iron mg/1000kcal	iron	0.646
Free cholesterol in very large VLDL; mmol/l	free cholesterol	0.762
Apolipoprotein A-1	apolipoprotein A-I	0.926
usage of wine or an equivalent in the past week	age	0.688
Creatinine value upon hospital admission	creatinin	0.696

Terminologies from TAUH	Terminologies from UHEI_LURIC	Score
Potassium mg	potassium	0.852
Triglycerides to total lipids ratio in medium LDL; %	triglycerides	0.718
Total cholesterol in small HDL; mmol/l	total cholesterol	0.767
Other	other med.	0.733
height	height	1.000
total cholesterol score, 0 = total cholesterol at least 5.172 or under medication for hypercholesterolemia, 1 = total cholesterol under 5.172 and not under medication for hypercholesterolemia	total cholesterol	0.696
Cholesterol esters to total lipids ratio in large HDL; %	cholesterol ester	0.742
Lipoprotein	lipoprotein	0.939
Total cholesterol in very small VLDL; mmol/l	total cholesterol	0.747
Myocardial infarction, without ST-evelation	myocardial infarction	0.806
Cause of death	Cause of death	1.000
Glucose	glucose 1h post oGT	0.724
Body mass index weight/	body mass index	0.847
Diabetes treated with insulin, 0 = no, 1 = yes	Diabetes	0.725
Total cholesterol	total cholesterol	0.940
blood pressure score, 0 = adult systolic at least 120 or diastolic at least 80 /, child systolic or diastolic over 90 percentile, 1 = adult systolic under 120 and diastolic under80 /, child systolic and diastolic under 90 percentile	blood pressure	0.687
Mother: myocardial infarction, I1,, 1 = no, 2 = yes	myocardial infarction	0.645
Heart rate on arrival, only for MI dataset patient	heart rate	0.693
Cholesterol esters in large HDL; mmol/l	cholesterol ester	0.784
CAD_SCA & lt; 1h	CAD	0.729
Waist circumference	waist circumference	0.965
Blood pressure measurers code, I2, 768, 10 = Helsinki, 20 = Turku, 30 = Tampere, 40 = Kuopio, 50 = Oulu	blood pressure	0.685

Terminologies from TAUH	Terminologies from UHEI_LURIC	Score
Father: myocardial infarction year of diagnosis yy, I2, 301, 98 = ei ole todettu, 99 = isä puuttuu perheestä	myocardial infarction	0.604
LDL-cholesterol	LDL-cholesterol	1.000
Diabetes type 1	Diabetes	0.844
Free cholesterol to total lipids ratio in very large HDL; %	free cholesterol	0.731
Calcium mg/1000kcal	calcium	0.641
Number of myocardial infarctions	myocardial infarction	0.735
Copper mg	copper	0.796
Comments brought up during the ultrasound test	Endogenous thrombin potential	0.605
Atrial fibrillation, 0 = no, 1 = yes	atrial fibrillation	0.686
Blood pressure measurers code, I2, 959	blood pressure	0.757
Free cholesterol to total lipids ratio in large VLDL; %	free cholesterol	0.737
Arrhythmia	arrhythmia	0.933
Hip circumference; cm	hip circumference	0.901
Zinc mg, F6.3, 214	zinc	0.639
Triglycerides in medium HDL; mmol/l	triglycerides	0.755
Protein g	protein C	0.852
Interleukin-13	interleukin 1	0.877
For menopause, years?	menopause	0.735
Diastolic Blood Pressure	diastolic blood pressure measurement 1	0.809
Total cholesterol in HDL3; mmol/l	total cholesterol	0.788
Hemoglobin ; g/l	hemoglobin	0.821
usage of cider etc in the past week	age	0.695
Free cholesterol to total lipids ratio in small HDL; %	free cholesterol	0.738
Father: myocardial infarction, I1, 191, 1 = no, 2 = yes	myocardial infarction	0.667
Copper mg, F6.2,1369	copper	0.694
Father: myocardial infarction, I1, 300, 1 = no, 2 = yes, 9 = father is missing from family	myocardial infarction	0.617
Free cholesterol in medium VLDL; mmol/l	free cholesterol	0.774
Free cholesterol to total lipids ratio in small VLDL; %	free cholesterol	0.737
Diabetes - Year when first time diagnosed	Diabetes	0.730

Terminologies from TAUH	Terminologies from UHEI_LURIC	Score
Weight of mother in 1980; kg	weight	0.671
Phospholipids in small LDL; mmol/l	phospholipid	0.610
Age of starting regular sports training	age at first use of insulin	0.689
Total cholesterol 1 measurement; mmol/l	total cholesterol	0.763
Cholesterol mg, F7.2, 530	cholesterol	0.770
Date of blood sampling	date of blood sampling	0.970
Internal organs and blood	irregular antibodies	0.622
Body mass index at initial measurements	body mass index	0.764
Free cholesterol to total lipids ratio in medium VLDL; %	free cholesterol	0.735
Heart rate	heart rate	0.933
GRACE-score for 2015-2016 subpopulation	ACE	0.692
Height of father; cm. Average from values collected between 1980 and 1989.	height	0.634
Systolic Blood Pressure	systolic lv pressure	0.843
Chronic obstructory pulmonary disease	mean systolic blood pressure	0.662
Magnesium mg, F8.2, 598	magnesium	0.746
Hemoglobin, F3.0,1201	hemoglobin	0.776
Sodium intake residual value from regression model standardized by energy in 2001	sodium	0.632
Total cholesterol revised average; mmol/l, F5.2, 862	total cholesterol	0.729
Myocardial infarction, strict	myocardial infarction	0.881
Systolic blood pressure, 3rd measurement	systolic blood pressure measurement 3	0.796
The date of the ICD procedure. No FC3!	date of	0.609
Miscarriage or abortion, 1 = no , 2 = yes	miscarriage	0.718
Heart rate on arrival, GRACEdataset patient only	heart rate	0.696
Cholesterol esters to total lipids ratio in medium HDL; %	cholesterol ester	0.741
Creatinine in different unit.	creatinin	0.722
Free cholesterol in large VLDL; mmol/l	free cholesterol	0.777
Smoking during the past week or during the week before attempting to quit, I3, 803,	smoking	0.643
Cholesterol esters in very small VLDL; mmol/l	cholesterol ester	0.766
Hip circumference	hip circumference	0.961
triglyceride	triglycerides	0.974

Terminologies from TAUH	Terminologies from UHEI_LURIC	Score
Vitamin C mg, F6.2, 524	vitamin C	0.746
Free cholesterol to total lipids ratio in small LDL; %	free cholesterol	0.738
Apolipoprotein E-113 genotype	apolipoprotein E	0.818
LDL cholesterol; mmol/l NOTE! Estimated LDL cholesterol. Ldlkol01 values has not been calculated for those participants with trigly01	LDL cholesterol	0.704
Systolic blood pressure 3rd measurement	systolic blood pressure measurement 3	0.803
Cholesterol esters to total lipids ratio in large VLDL; %	cholesterol ester	0.741
Total cholesterol to total lipids ratio in very large VLDL; %	total cholesterol	0.714
Acute Coronary Syndrome	acute coronary syndrome	0.830
Heart rate/min 2nd measurement	heart rate	0.733
Waist/hip circumference ratio	hip circumference	0.803
Copper mg, F6.2, 624	copper	0.694
Heart rate/min 1st measurement	heart rate	0.733
Triglycerides in IDL; mmol/l	triglycerides	0.784
Triglycerides to total lipids ratio in small LDL; %	triglycerides	0.719
Weight of the participant; kg	weight	0.669
Vitamin C mg	vitamin C	0.852
Triglycerides in medium LDL; mmol/l	triglycerides	0.755
Erythrocytes; E12/l	erythrocytes	0.832
Coronary angiopasty	coronary angioplasty	0.949
lipoprotein lipase precursor, Ter*474S	lipoprotein	0.763
diabetes diet , 0 = no, 1 = yes	Diabetes	0.700
Age at DEATH	age	0.611
Free cholesterol in small LDL; mmol/l	free cholesterol	0.781
Free cholesterol to total lipids ratio in very small VLDL; %	free cholesterol	0.729
Triglycerides in large LDL; mmol/l	triglycerides	0.759
Triglyceride	triglycerides	0.921
Total cholesterol to total lipids ratio in IDL; %	total cholesterol	0.735
Total cholesterol to total lipids ratio in large LDL; %	total cholesterol	0.723

Terminologies from TAUH	Terminologies from UHEI_LURIC	Score
Iron mg, F6.2, 618	iron	0.639
Triglycerides in small VLDL; mmol/l	triglycerides	0.755
Body mass indexi	body mass index	0.936
Serum testosterone	Testosteron	0.738
Total cholesterol in very large VLDL; mmol/l	total cholesterol	0.747
Potassium mg, F6.3, 584	potassium	0.746
age in 2011	age	0.758
Weight at birth; g	weight	0.704
Diabetes, type 2	Diabetes	0.833
Cholesterol esters to total lipids ratio in medium LDL; %	cholesterol ester	0.741
Height of the participant; cm	height	0.669
HDL cholesterol 2 meas.; mmol/l	HDL cholesterol	0.828
Weight of mother in 1989; kg	weight	0.671
Total cholesterol in medium VLDL; mmol/l	total cholesterol	0.760
Interleukin 1-beta	interleukin 1	0.863
MCV; fl	MCV	0.810
Creatinine; µmol/l	Creatinin [µmol/L]	0.889
Weight	weight	0.889
Total cholesterol to total lipids ratio in chylomicrons and extremely large VLDL; %	total cholesterol	0.690
lipoprotein, Lp	lipoprotein	0.911
Total cholesterol 2 measurement; mmol/l	total cholesterol	0.763
Fibrinogen, gamma chain	fibrinogen	0.764
Cholesterol esters in medium LDL; mmol/l	cholesterol ester	0.780
Free cholesterol in very large HDL; mmol/l	free cholesterol	0.765
Inducible T-cell co-stimulator precursor	mean diastolic blood pressure	0.630
Zinc mg, F6.2, 636	zinc	0.639
Total cholestrerol	total cholesterol	0.902
blood pressure score, 0 = systolic at least 120 or diastolic at least 80 or under medication for hypertension, 1 = systolic under 120 and diastolic under 80 and not under medication for hypertension	blood pressure	0.690
Lp	LpPLA2 activity	0.711
Triglycerides to total lipids ratio in small HDL; %	triglycerides	0.719

Terminologies from TAUH	Terminologies from UHEI_LURIC	Score
Mother: coronary angioplasty, I1,, 1 = no, 2 = yes	coronary angioplasty	0.683
IDL-cholesterol	HDL-cholesterol	0.956
Triglycerides in very large VLDL; mmol/l	triglycerides	0.741
Triglycerides to total lipids ratio in IDL; %	triglycerides	0.730
1st medically diagnosed permanent injury or disability	systolic blood pressure measurement 1	0.610
cholesterol score, 0 = adult total cholesterol at least 5.17 /, child total cholesterol at least 4.40, 1 = adult total cholesterol under 5.17 /, child total cholesterol under 4.40	cholesterol	0.687
Cholesterol esters in medium HDL; mmol/l	cholesterol ester	0.780
weight	weight	1.000
Cholesterol esters in IDL; mmol/l	cholesterol ester	0.809
HDL cholesterol; mmol/l.	HDL cholesterol	0.875
Weight of mother in 1986; kg	weight	0.671
Date of installation of the ICD	date of	0.684
Phospholipids in medium LDL; mmol/l	phospholipid	0.607
Date of testing	date of	0.752
Systolic blood pressure, 2nd measurement	systolic blood pressure measurement 2	0.796
Free cholesterol in small VLDL; mmol/l	free cholesterol	0.777
date of birth	date of	0.846
Total cholesterol to total lipids ratio in large VLDL; %	total cholesterol	0.721
Triglycerides to total lipids ratio in chylomicrons and extremely large VLDL; %	triglycerides	0.692
Iron mg, F6.3, 198	iron	0.639
Sodium mg, F8.2, 722	sodium	0.694
Smoking, 0=no, 1=yes. This variable should be used for contemporary smoking.	smoking	0.645
Fasting	fasting leptin	0.762
Diabetes mellitus, tyype 1	Diabetes	0.769
Free cholesterol in chylomicrons and extremely large VLDL; mmol/l	free cholesterol	0.723
Weight of father in 1980; kg	weight	0.671
Total cholesterol to total lipids ratio in medium VLDL; %	total cholesterol	0.720
HDL cholesterol; mmol/l	HDL cholesterol	0.884

Terminologies from TAUH	Terminologies from UHEI_LURIC	Score
Cholesterol esters in very large HDL; mmol/l	cholesterol ester	0.768
Zinc mg, F6.2,1381	zinc	0.639
Total cholesterol to total lipids ratio in medium LDL; %	total cholesterol	0.721
Free testosterone counted by Sartorius formula; pmol/L:, ftz07=	Testosteron	0.606
body-mass index	body mass index	0.896
Weight of father in 1986; kg	weight	0.671
CAD that really contributed	CAD	0.704
Myocardial infarction, 0 = no, 1 = yes	myocardial infarction	0.826
Interleukin-8	interleukin 8	0.897
Total cholesterol to total lipids ratio in large HDL; %	total cholesterol	0.723
HDL mediated cholesterol efflux capacity; %	Cholesterol efflux capacity	0.696
Spironolactone for use when returning home or switching to follow-up treatment	iron	0.684
Triglyserides	triglycerides	0.807
Cholesterol esters to total lipids ratio in small HDL;%	cholesterol ester	0.744
Magnesium mg,F6.2,1346	magnesium	0.751
smoking score, 0 = currently smoking, 1 = never smoked or has quit or is on a break	smoking	0.695
Free cholesterol to total lipids ratio in chylomicrons and extremely large VLDL; %	free cholesterol	0.707
Chest pain in rest	heparin test	0.752
Total cholesterol in large HDL; mmol/l	total cholesterol	0.767
Free cholesterol to total lipids ratio in large HDL; %	free cholesterol	0.738
Weight of father in 1983; kg	weight	0.671
Free cholesterol to total lipids ratio in medium LDL; %	free cholesterol	0.737
Interleukin-17	interleukin 1	0.877
Cholesterol esters in small VLDL; mmol/l	cholesterol ester	0.780
Mean glucose measurements	oral glucose tolerance test	0.657
Zinc mg	zinc	0.726
Heart rate/min 3rd measurement	heart rate	0.733
Weight of father in 1989; kg	weight	0.671
Apolipoprotein A-I; g/l	apolipoprotein A-I	0.895

Terminologies from TAUH	Terminologies from UHEI_LURIC	Score
Cholesterol mg, F6.2,1281	cholesterol	0.770
Diastolic KV blood pressure 2nd measurement	diastolic blood pressure measurement 2	0.847
Diabetes; mother, I1, 217, 1 = no, 2 = yes, 9 = mother is missing from family	Diabetes	0.701
Blood pressure meters code, I2, 766	blood pressure	0.767
Triglycerides in LDL; mmol/l	triglycerides	0.784
Systolic blood pressure, 1st measurement	systolic blood pressure measurement 1	0.813
Mitral valve disease	valve disease	0.768
Apolipoprotein E-219 promoottorigenotype	apolipoprotein E	0.771
Total cholesterol; mmol/l	total cholesterol	0.840
Fibrinogen, gamma chain , 0 = genotyping did not succeed, 1=T, 2=TC, 3=C, rs number: rs1800792	fibrinogen	0.665
Interleukin 1-beta, numeric	interleukin 1	0.789
Phospholipids in IDL; mmol/l	phospholipid	0.633
Apolipoprotein E, 0 = genotyping did not succeed, 1=G, 2=AG, 3=A	apolipoprotein E	0.724
Cholesterol mg	cholesterol	0.874
Diastolic KIV blood pressure 3rd measurement	diastolic blood pressure measurement 3	0.841
Mother: myocardial infarction, 1 = no, 2 = yes	myocardial infarction	0.660
Interleukin 1 beta , 1=G, 2=AG, 3=A	interleukin 1	0.755
Iron mg, F6.2,1363	iron	0.639
Cholesterol esters in small LDL; mmol/l	cholesterol ester	0.784
Age 1986	age	0.639
Cholesterol esters in medium VLDL; mmol/l	cholesterol ester	0.777
id number in study	study number	0.620
age in 1980	age	0.758
Heart Rate	heart rate	0.867
Total cholesterol in small LDL; mmol/l	total cholesterol	0.767
Body mass index; kg/m2	body mass index	0.857
Cholesterol esters to total lipids ratio in IDL; %	cholesterol ester	0.754
Beta blockers, 0=no, 1=yes	beta blocker	0.704
Phospholipids in medium HDL; mmol/l	phospholipid	0.607
Date of ultrasound test, according to blood pressure listing	date of	0.652

Terminologies from TAUH	Terminologies from UHEI_LURIC	Score
Body Mass Index	body mass index	0.867
usage of wine or an equivalent in the past week, I2,	age	0.686
usage of medium	age	0.733
Apo	ApoA-IV	0.810
Total cholesterol; mmol/l, NOTE! tutkno80=728: abnormally high total cholesterol value =10.2 mmol/l. Delete the value in question if needed as an outlier.	total cholesterol	0.661
Cholesterol esters in very large VLDL; mmol/l	cholesterol ester	0.766
Waist circumference average measurement; cm	waist circumference	0.789
Diabetes mellitus, tyype 2	Diabetes	0.769
Father: myocardial infarction, I1,, 1 = no, 2 = yes	myocardial infarction	0.677
Insulin	insulin treatment	0.737
Calcium mg, F7.2,1339	calcium	0.631
Actual systolic blood pressure, 1st measurement	systolic blood pressure measurement 1	0.767
Diabetes	Diabetes	1.000
Creatinine upn arrival or mean value from hospitalization or before hospitalization	creatinin	0.662
Free cholesterol to total lipids ratio in medium HDL; %	free cholesterol	0.737
Diastolic KIV blood pressure 2nd measurement	diastolic blood pressure measurement 2	0.841
Total cholesterol in very large HDL; mmol/l	total cholesterol	0.750
Date of angiography	date of	0.724
Diastolic KV blood pressure 1st measurement	diastolic blood pressure measurement 1	0.796
Total cholesterol 2nd measurement, F5.2,1129	total cholesterol	0.747
Iron mg	iron	0.726
Triglycerides to total lipids ratio in large LDL; %	triglycerides	0.719
Free testosterone counted by Nanjee-Wheelerin formula, pmol/L:, ftn01=	Testosteron	0.601
Total cholesterol in large LDL; mmol/l	total cholesterol	0.767

Terminologies from TAUH	Terminologies from UHEI_LURIC	Score
Triglycerides to total lipids ratio in large VLDL; %	triglycerides	0.718
Cholesterol esters to total lipids ratio in very large HDL; %	cholesterol ester	0.734
HDL cholesterol. Same procedure as in 2001; mmol/l. , NOTE! In 2007 samples, limit of detection for HDL is 0.5.	HDL cholesterol	0.712
Magnesium mg, F7.3, 191	magnesium	0.746
Free cholesterol to total lipids ratio in very large VLDL; %	free cholesterol	0.729
Total cholesterol in chylomicrons and extremely large VLDL; mmol/l	total cholesterol	0.707
MACE or PAD	ACE	0.758
Apolipoprotein B	apolipoprotein B	0.958
Phospholipids in large HDL; mmol/l	phospholipid	0.610
Digitalis, 0=no, 1=yes	digitalis	0.751
Systolic blood pressure 1st measurement	systolic blood pressure measurement 1	0.821
HDL cholesterol average.; mmol/l	HDL cholesterol	0.823
usage of spirits/liqueurs in the past week	age	0.690
Cholesterol esters to total lipids ratio in very small VLDL; %	cholesterol ester	0.733
Date of ultrasound testing	date of	0.696
, 0 = Total cholesterol at least 5.172 or medication for hypercholesterolemia , 1 = Total cholesterol less than 5.172 and no medication for hypercholesterolemia	total cholesterol	0.639
Calcium mg, F8.3, 183	calcium	0.631
Free cholesterol in small HDL; mmol/l	free cholesterol	0.781
Free cholesterol in medium HDL; mmol/l	free cholesterol	0.777
notice	cotinine	0.722
HDL-kolesteroli	HDL-cholesterol	0.911
Total cholesterol to total lipids ratio in small HDL; %	total cholesterol	0.723
smoking score, 0 = adult currently smoking /, child smoked at least one cigarette, 1 = adult never smoked or has quit or is on a break / , child never smoked	smoking	0.682

Terminologies from TAUH	Terminologies from UHEI_LURIC	Score
Cholesterol esters to total lipids ratio in small VLDL; %	cholesterol ester	0.741
Free cholesterol in very small VLDL; mmol/l	free cholesterol	0.762
Smoking at age 18-24. Has the participant smoked daily at least in some point of his/her youth age	smoking	0.639
Triglycerides in medium VLDL; mmol/l	triglycerides	0.752
Triglycerides to total lipids ratio in medium VLDL; %	triglycerides	0.716
Total cholesterol in large VLDL; mmol/l	total cholesterol	0.763
Phospholipids in small VLDL; mmol/l	phospholipid	0.607
Myocardial infarction, with ST-evelation	myocardial infarction	0.817
Free testosterone counted by Sartorius formula; pmol/L:, ftz01=	Testosteron	0.606
Cholesterol esters to total lipids ratio in medium VLDL; %	cholesterol ester	0.739
Apolipoprotein B; g/l	apolipoprotein B	0.884
Total cholesterol in medium LDL; mmol/l	total cholesterol	0.763
sodium score, 0 = sodium proportion of whole energy intake at least 1.15 mg/ Kcal, 1 = sodium proportion of whole energy intake under 1.15 mg/ Kcal	sodium	0.680
Weight of mother in 1983; kg	weight	0.671
Cholesterol esters to total lipids ratio in very large VLDL; %	cholesterol ester	0.733
age in 2007	age	0.758
Free cholesterol in large HDL; mmol/l	free cholesterol	0.781
Atrial fibrillation	atrial fibrillation	0.835
Total cholesterol 1st measurement, F5.2,1124	total cholesterol	0.747
Cardiomyopathy, 0=no, 1=yes	cardiomyopathy	0.803
Blood pressure meters code, F2.0, 957	blood pressure	0.760
LDL cholesterol, direct	LDL cholesterol	0.884
Creatinine. Same procedure as in 2001; µmol/l	creatinin	0.689
proprotein convertase subtilisin/kexin type 9, L46R	protein C	0.682
Nitric oxide synthase 1	nitric oxide	0.798
zinc finger protein 652	zinc	0.725

Terminologies from TAUH	Terminologies from UHEI_LURIC	Score
Systolic blood pressure 2nd measurement	systolic blood pressure measurement 2	0.803
smoking score, 0 = current smoking, 1 = never smoked or has quit or is on a break	smoking	0.695
usage of spirits/liqueurs in the past week, I2,,	age	0.688
Smoking at age 12-18. Has the participant smoked daily at least in some point of his/her youth age	smoking	0.639
GRACE-score for 2015-2016 subpopulation with missing values imputed	ACE	0.681
Triglycerides to total lipids ratio in very large HDL; %	triglycerides	0.712
Total cholesterol in VLDL; mmol/l	total cholesterol	0.788
Triglycerides to total lipids ratio in large HDL; %	triglycerides	0.719
Triglycerides in VLDL; mmol/l	triglycerides	0.779
Actual systolic blood pressure, 2nd measurement	systolic blood pressure measurement 2	0.762
Myocardial infarction, unclassifiable	myocardial infarction	0.831
Total cholesterol average; mmol/l	total cholesterol	0.788
Mother: myocardial infarction, I1, 190, 1 = no, 2 = yes	myocardial infarction	0.635
Free cholesterol to total lipids ratio in large LDL; %	free cholesterol	0.738
Interleukin-9	interleukin 9	0.897
Heart rate mean value, from the status form	heart rate	0.703
Hip circumference average measurement; cm	hip circumference	0.777
HDL cholesterol 1 meas.; mmol/l	HDL cholesterol	0.828
Lpa = lipoprotein	lipoprotein	0.807
Body mass index at final measurements	body mass index	0.771
Heart rate upon arrival	heart rate	0.764
apolipoprotein E precursor	apolipoprotein E	0.872
Homocysteine; µmol/l	homocysteine	0.822
Cholesterol esters in large VLDL; mmol/l	cholesterol ester	0.780
Age	age	0.778
Calcium mg	calcium	0.736
Beta Blocker used when returning home or switching to follow-up treatment	beta blocker	0.657
Interleukin-16	interleukin 1	0.877

Terminologies from TAUH	Terminologies from UHEI_LURIC	Score
Total cholesterol in medium HDL; mmol/l	total cholesterol	0.763
Phospholipids in large LDL; mmol/l	phospholipid	0.610
Creatinine; mg/dl	Creatinin [mg/dL]	0.882
Interleukin-10	interleukin 10	0.905
Total cholesterol in HDL2; mmol/l	total cholesterol	0.788
Total cholesterol. correction factor 1.000; mmol/l	total cholesterol	0.733
Triglycerides to total lipids ratio in very small VLDL; %	triglycerides	0.711
Diabetes diagnosed by doctor, I1, 134, 1 = no, 2 = yes	Diabetes	0.716
usage of cider etc in the past week, I2,,	age	0.691
cotinine in 1980; ng/ml	cotinine	0.783
The date of the exercise test	date of	0.628
Interleukin-2	interleukin 2	0.897
Father: coronary angioplasty, 1 = no, 2 = yes	coronary angioplasty	0.673
Cholesterol esters in large LDL; mmol/l	cholesterol ester	0.784
Triglycerides in HDL; mmol/l	triglycerides	0.784
Total cholesterol to total lipids ratio in very small VLDL; %	total cholesterol	0.714
Triglycerides in small HDL; mmol/l	triglycerides	0.759
Diastolic KIV blood pressure 1st measurement	diastolic blood pressure measurement 1	0.790
MCH; pg	MCH	0.810
Triglycerides in chylomicrons and extremely large VLDL;mmol/l	triglycerides	0.707
cholesterol medication when returning home or switching to follow-up treatment	cholesterol	0.714
Free cholesterol in large LDL; mmol/l	free cholesterol	0.781
Myocardial infarction - Year when first time diagnosed	myocardial infarction	0.774
Total cholesterol in HDL; mmol/l	total cholesterol	0.793
GRACE score for MI-ECG subpopulation	ACE	0.694
Total cholesterol to total lipids ratio in medium HDL; %	total cholesterol	0.721
other	other med.	0.833
Type of physical exercise	Cholesterol efflux capacity	0.603

Terminologies from TAUH	Terminologies from UHEI_LURIC	Score
Heart rate mean value when sitting, from the status form; Heart rate/min	heart rate	0.675
ACE inhibitor or ATR blocker used when returning home or switching to follow-up care	ACE	0.679
Free cholesterol in medium LDL; mmol/l	free cholesterol	0.777
Sodium mg	sodium	0.796
Triglycerides to total lipids ratio in medium HDL; %	triglycerides	0.718
Hematocrit; Osuus	hematocrit	0.810
usage of strong	age	0.733
Total cholesterol in LDL; mmol/l	total cholesterol	0.793
Cholesterol esters to total lipids ratio in large LDL; %	cholesterol ester	0.742
Total cholesterol in IDL; mmol/l	total cholesterol	0.793
Free cholesterol to total lipids ratio in IDL; %	free cholesterol	0.750
Apolipoprotein E	apolipoprotein E	0.958
Phospholipids in large VLDL; mmol/l	phospholipid	0.607
Calcium mg. DO NOT USE.	calcium	0.623
Smoking during the past week or during the week before attempting to quit, I3, 666,	smoking	0.643
Copper mg, F5.3, 204	copper	0.694
Cholesterol esters to total lipids ratio in small LDL; %	cholesterol ester	0.742
Total cholesterol to total lipids ratio in small VLDL; %	total cholesterol	0.721
Height at birth; cm,	height	0.694
Triglycerides in large VLDL; mmol/l	triglycerides	0.755
Height of mother; cm. Average from values collected between 1980 and 1989.	height	0.634
Total cholesterol in small VLDL; mmol/l	total cholesterol	0.763

Supplementary Table 3. A summary of the potentially matched terminologies between TAUH and UVA using the proposed approach.

Terminologies from OCVD	Terminologies from OMD	Score
Other chronic disease	chronic disease	0.859
How many hours spends on moderately heavy yard and housework in a month, on average	Pain disappear, remain when standing or walking slowly	0.622

Terminologies from OCVD	Terminologies from OMD	Score
HDL cholesterol. Same procedure as in 2001; mmol/l. , NOTE! In 2007 samples, limit of detection for HDL is 0.5.	HDL cholesterol	0.712
Triglycerides in very small VLDL; mmol/l	triglycerides	0.741
Depression	Depression found	0.875
medication for hypercholesterolemia, 0 = other cholesterol lowering medication, 1 = statin, 2 = statin and other cholesterol lowering medication	medication	0.690
Reason for termination, difficulty in breathing	Mean cellular hemoglobin concentration of erythrocytes	0.663
Total cholesterol to total lipids ratio in medium LDL; %	Total cholesterol	0.768
Heart failure	heart failure	0.822
Weight of father in 1986; kg	Weight	0.738
Total cholesterol in chylomicrons and extremely large VLDL; mmol/l	Total cholesterol	0.753
Apolipoprotein B	Lipoprotein	0.761
Age-adjusted expected maximum HR	Age	0.698
Age of youngest child in the family, I2, 93	Age	0.690
Total cholesterol to total lipids ratio in large HDL; %	Total cholesterol	0.770
Age of starting habitual smoking	Age	0.698
C-reactive protein gene	C-reactive protein	0.928
HDL cholesterol average.; mmol/l	HDL cholesterol	0.823
Marital status since year ?, I4,,	marital status	0.774
Date of laboratory sample	date	0.623
Glucose; mmol/l	glucose	0.752
Weight of mother in 1980; kg	Weight	0.738
C-reactive protein,	C-reactive protein	0.982
The death was later performed in cardiac surgery	When complaints for the first time occurred in the life	0.620
Left side, technical level of angiography	almost never, things develop according to my ideas	0.657
Triglyserides	triglycerides	0.807
Age of starting regular sports training	Age	0.692
Size, LAD DG I	size	0.655
Glucose g, F5.2,1579	glucose	0.719
Total cholesterol 1 measurement; mmol/l	Total cholesterol	0.812

Terminologies from OCVD	Terminologies from OMD	Score
Triglycerides in very large HDL; mmol/l	triglycerides	0.744
, 0 = Total cholesterol at least 5.172 or medication for hypercholesterolemia , 1 = Total cholesterol less than 5.172 and no medication for hypercholesterolemia	Total cholesterol	0.643
Date of ultrasound testing	date	0.622
C-reactive protein	C-reactive protein	1.000
Date of blood sampling	date	0.629
Myocardial infarction	myocardial infarction	0.857
Heart rate	heart rate	0.933
Mothers weight before possible current pregnancy	Mean cellular hemoglobin concentration of erythrocytes	0.653
Age of youngest child in the family, I2, 49	Age	0.690
Height of father; cm. Average from values collected between 1980 and 1989.	height	0.634
Does the participant attempt to select low salt products	When complaints for the first time occurred in the life	0.627
Size, RCA RPL	size	0.660
Size, LCx LOM II	size	0.646
Systolic Blood Pressure	Target systolic blood pressure	0.659
Age when first trying alcohol, I2, 810, 99 = has never tried	Age	0.683
How many hours spends on light yard and housework in a month, on average	Pain disappear, remain when standing or walking slowly	0.656
Index of physical activity 1986.	physical activity	0.756
Fathers daily consumption of whole milk	When complaints for the first time occurred in the life	0.623
How many hours spends daily working on the computer	I rarely count that happens to me something good	0.655
Total cholesterol revised average; mmol/l, F5.2, 862	Total cholesterol	0.776
Age when first trying smoking, I2, 780, 99 = has never smoked	Age	0.683
Total cholesterol in large HDL; mmol/l	Total cholesterol in mg/dl	0.853
Total cholesterol to total lipids ratio in small HDL; %	Total cholesterol	0.770
Leukocytes; E9/l	leukocytes	0.821
Myocardial infarction, strict	Myocardial infarction, first time	0.754

Terminologies from OCVD	Terminologies from OMD	Score
Age of oldest child in the family, I2, 51	Age	0.691
Total cholesterol to total lipids ratio in small LDL; %	Total cholesterol	0.770
Medication, I1, 836, 1 = has consumed drugs, 2 = has not consumed drugs	medication	0.676
Triglycerides in medium VLDL; mmol/l	triglycerides	0.752
Triglycerides to total lipids ratio in medium VLDL; %	triglycerides	0.716
Total cholesterol in large VLDL; mmol/l	Total cholesterol	0.812
Longest RR interval	Long nitrates	0.678
Triglycerides to total lipids ratio in very large VLDL; %	triglycerides	0.711
Age at death of uncle or aunt died of myocardial infarction	Age	0.684
Weight of father in 1983; kg	Weight	0.738
Age at time of action	Age	0.714
Size, RCA	size	0.694
Diabetes type 2	diabetes type	0.726
How often performs physical exercise during spare time	when in a hurry or during physical exertion	0.664
Triglycerides in small LDL; mmol/l	triglycerides	0.759
Heart rate on arrival, GRACEdataset patient only	heart rate	0.696
Mean glucose measurements	Oral glucose tolerance test	0.644
? Age when celiac disease observed	Age	0.696
Metabolic syndrome according to the IDF definition	among my friends I feel comfortable	0.614
Index of physical activity 2011.	physical activity	0.756
Hyperlipidemia	hyperlipidemia	0.898
Heart rate/min 3rd measurement	heart rate	0.733
Height	height	0.889
Apolipoprotein B; g/l	Lipoprotein	0.712
Weight of father in 1989; kg	Weight	0.738
triglyceride	triglycerides	0.974
Apolipoprotein A-I; g/l	Lipoprotein	0.698
Age of starting habitual use of alcohol, I2, 812, 98 = does not use alcohol habitually, 99 = has never tried	Age	0.676
Total cholesterol in medium LDL; mmol/l	Total cholesterol	0.812

Terminologies from OCVD	Terminologies from OMD	Score
Childs movement or activity when playing compared to other children	almost never, things develop according to my ideas	0.707
Fathers date of birth ddmm, I4, 14	date	0.613
Weight of mother in 1983; kg	Weight	0.738
Amount of cigars/cigarillos smoked per day at present	Difficulty concentrating / decision problems	0.645
Apolipoprotein E-113 genotype	Lipoprotein	0.668
Total cholesterol average, F5.2,1154	Total cholesterol	0.824
age in 2007	Age	0.616
LDL cholesterol; mmol/l NOTE! Estimated LDL cholesterol. Ldlkol01 values has not been calculated for those participants with trigly01	LDL cholesterol	0.704
Total cholesterol to total lipids ratio in very large VLDL; %	Total cholesterol	0.760
Triglycerides in LDL; mmol/l	triglycerides	0.784
Acute Coronary Syndrome	with acute coronary syndrome	0.738
Heart rate/min 2nd measurement	heart rate	0.733
Age of starting habitual use of alcohol	Age	0.692
Alcohol	Alcohol consumption in last 12 months	0.730
I have a hard time finding the right words to describe my feelings	before how many months last determined by the doctor	0.696
Size, RCA RV	size	0.667
How many hours spends daily watching TV, on average	almost never, things develop according to my ideas	0.671
Atrial fibrillation	atrial fibrillation	0.792
Total cholesterol 1st measurement, F5.2,1124	Total cholesterol	0.795
Heart rate/min 1st measurement	heart rate	0.733
Systolic arterial pressure when standing	when in a hurry or during physical exertion	0.609
Triglycerides in IDL; mmol/l	triglycerides	0.784
Apolipoprotein E-219 promoottorigenotype	Lipoprotein	0.636
LDL cholesterol, direct	LDL cholesterol	0.884
Total cholesterol; mmol/l	Total cholesterol in mmol/l	0.950
Triglycerides to total lipids ratio in small LDL; %	triglycerides	0.719
Weight of the participant; kg	Weight	0.736
Triglysceride,	triglycerides	0.830
Age of starting regular sports training, I2,1021	Age	0.688

Terminologies from OCVD	Terminologies from OMD	Score
Triglycerides in medium LDL; mmol/l	triglycerides	0.755
Age when first trying alcohol	Age	0.701
Erythrocytes; E12/l	erythrocytes	0.832
Triglycerides to total lipids ratio in very large HDL; %	triglycerides	0.712
Total cholesterol in VLDL; mmol/l	Total cholesterol in mmol/l	0.939
lipoprotein lipase precursor, Ter*474S	Lipoprotein	0.724
Age at DEATH	Age	0.750
Index of physical activity 1980.	physical activity	0.756
Triglycerides to total lipids ratio in large HDL; %	triglycerides	0.719
Marital status since year ?	marital status	0.803
Mothers daily consumption of whole milk	When complaints for the first time occurred in the life	0.618
Index of physical activity 1989.	physical activity	0.756
Childs date of birth ddmm, I4, 5.	Date of birth	0.637
Triglycerides in large LDL; mmol/l	triglycerides	0.759
glucose score, 0 = glucose at least 5.55 or diabetes, 1 = glucose at least 5.55 and not diabetes	glucose	0.691
Triglycerides in VLDL; mmol/l	triglycerides	0.779
Triglyceride	triglycerides	0.921
Myocardial infarction, unclassifiable	Myocardial infarction, Number	0.691
Total cholesterol average; mmol/l	Total cholesterol	0.838
HDL-cholesterol	HDL cholesterol	0.956
Marital status of the spouse, I1,, 1 = unmarried, 2 = married, 3 = engaged, 4 = cohabiting, 5 = divorced/legally separated, 6 = widow	marital status	0.675
Total cholesterol to total lipids ratio in IDL; %	Total cholesterol	0.782
Triglycerides in large HDL; mmol/l	triglycerides	0.759
Index of physical activity 1983.	physical activity	0.756
Apolipoprotein E, 0 = genotyping did not succeed, 1=G, 2=AG, 3=A	Lipoprotein	0.605
Total cholesterol to total lipids ratio in large LDL; %	Total cholesterol	0.770
Type 1 diabetes	Type I diabetes	0.956
Heart rate mean value, from the status form	heart rate	0.703
Triglycerides in small VLDL; mmol/l	triglycerides	0.755
Total cholesterol to total lipids ratio in very large HDL; %	Total cholesterol	0.761

Terminologies from OCVD	Terminologies from OMD	Score
milk products and milk glasses	current setting and compliance	0.650
Serum testosterone	testosterone	0.722
Margarine and oils	more diagnoses	0.623
Total cholesterol in very large VLDL; mmol/l	Total cholesterol	0.795
age in 2011	Age	0.616
Time of ultrasound test	master stressful situations	0.628
Weight at birth; g	Weight	0.778
Teacher makes me feel I am not good enough	more alcohol drinking justifiable as health	0.628
Previous heart failure or now decompensation.	When complaints for the first time occurred in the life	0.632
Triglycerides to total lipids ratio in small VLDL; %	triglycerides	0.718
LDL cholesterol, direct; mmol/l	LDL cholesterol	0.828
Oral medication as a treatment for diabetes, 0=no, 1=yes	medication	0.626
Height of the participant; cm	height	0.669
SWM Double errors; number of errors that may be classified as both between and within errors	before how many months last determined by the doctor	0.636
Size, LAD SEPT I	size	0.646
HDL cholesterol 1 meas.; mmol/l	HDL cholesterol	0.828
Marital status of the participant, I1, , 1 = unmarried, 2 = married, 3 = engaged, 4 = cohabiting, 5 = divorced/legally separated, 6 = widow	marital status	0.674
Apolipoprotein A-1	Lipoprotein	0.738
Room temperature in degrees celcius	current setting and compliance	0.636
It’s easy for me to describe my feelings	among my friends I feel comfortable	0.652
HDL cholesterol 2 meas.; mmol/l	HDL cholesterol	0.828
Lpa = lipoprotein	Lipoprotein	0.822
Weight of mother in 1989; kg	Weight	0.738
Total cholesterol in medium VLDL; mmol/l	Total cholesterol	0.808
Heart rate upon arrival	heart rate	0.764
Gender, 1 = female, 2 = male	gender	0.671
apolipoprotein E precursor	Lipoprotein	0.681
Marital status of the participant, 1 = unmarried , 2 = married , 3 = in a registered relationship , 4 = cohabiting , 5 = divorced/legally separated , 6 = widow	marital status	0.670

Terminologies from OCVD	Terminologies from OMD	Score
Triglycerides to total lipids ratio in medium LDL; %	triglycerides	0.718
Total cholesterol in small HDL; mmol/l	Total cholesterol	0.816
Other	Other	0.875
height	height	1.000
Homocysteine; µmol/l	homocysteine	0.822
total cholesterol score, 0 = total cholesterol at least 5.172 or under medication for hypercholesterolemia, 1 = total cholesterol under 5.172 and not under medication for hypercholesterolemia	Total cholesterol	0.654
Age	Age	1.000
Age 1986	Age	0.792
Age at baseline	Age	0.733
Weight	Weight	1.000
Total cholesterol to total lipids ratio in chylomicrons and extremely large VLDL; %	Total cholesterol	0.735
lipoprotein, Lp	Lipoprotein	0.859
Total cholesterol 2 measurement; mmol/l	Total cholesterol	0.812
Size, LCx LOM I	size	0.650
Lipoprotein	Lipoprotein	1.000
Total cholesterol in medium HDL; mmol/l	Total cholesterol	0.812
Total cholesterol in very small VLDL; mmol/l	Total cholesterol	0.795
Glucose	glucose	0.905
Femur neck bone density at final measurements	Number of almost daily Drinks	0.612
Total cholesterol in HDL2; mmol/l	Total cholesterol in mmol/l	0.939
age in 1980	Age	0.616
Total cholestrerol	Total cholesterol	0.962
glucose score, 0 = glucose at least 5.55 or under medication for diabetes, 1 = glucose at least 5.55 and not under medication for diabetes	glucose	0.684
Heart Rate	heart rate	0.867
Index of physical activity. The higher the value, the more active the participant is.	physical activity	0.645
Total cholesterol in small LDL; mmol/l	Total cholesterol	0.816
Total cholesterol. correction factor 1.000; mmol/l	Total cholesterol	0.780
Right side, technical level of angiography	almost never, things develop according to my ideas	0.679

Terminologies from OCVD	Terminologies from OMD	Score
Triglycerides to total lipids ratio in small HDL; %	triglycerides	0.719
C-reactive protein from 2001	C-reactive protein	0.881
Triglycerides to total lipids ratio in very small VLDL; %	triglycerides	0.711
Total cholesterol	Total cholesterol	1.000
Glucose g, F5.2,1511	glucose	0.719
Cholesterol esters in large LDL; mmol/l	Total cholesterol in mg/dl	0.611
Triglycerides in very large VLDL; mmol/l	triglycerides	0.741
date from the search Feb 2007; additional coronary angiographies	date	0.688
Triglycerides in HDL; mmol/l	triglycerides	0.784
Triglycerides to total lipids ratio in IDL; %	triglycerides	0.730
Heart rate on arrival, only for MI dataset patient	heart rate	0.693
Cholesterol esters in large HDL; mmol/l	Total cholesterol in mg/dl	0.611
Total cholesterol to total lipids ratio in very small VLDL; %	Total cholesterol	0.760
Index of physical activity 1992.	physical activity	0.756
Triglycerides in small HDL; mmol/l	triglycerides	0.759
Has the participant used nutrient products or food supplements	When complaints for the first time occurred in the life	0.645
Does the participant attempt to select products with added nutrients	When complaints for the first time occurred in the life	0.650
Triglycerides in chylomicrons and extremely large VLDL;mmol/l	triglycerides	0.707
Clopidogrel is used when you return home or switch to further treatment	When complaints for the first time occurred in the life	0.661
weight	Weight	0.889
I often lose my courage at school	Mean cellular hemoglobin content of erythrocytes	0.606
cholesterol medication when returning home or switching to follow-up treatment	When complaints for the first time occurred in the life	0.671
Size, LCx	size	0.694
Total cholesterol; mmol/l, NOTE! tutkno80=728: abnormally high total cholesterol value =10.2 mmol/l. Delete the value in question if needed as an outlier.	Total cholesterol	0.703
Age when first trying smoking, I2, 647, 99 = has never smoked	Age	0.683

Terminologies from OCVD	Terminologies from OMD	Score
Hours spent daily playing outside during the summer	I rarely count that happens to me something good	0.642
Absences from school or work or days stayed indoors due to illness	before how many months last determined by the doctor	0.667
LDL-cholesterol	LDL cholesterol	0.956
Coronary artery induced chest pain	Diagnostic status hypertension	0.622
Potato and vegetables	against elevated blood lipids	0.623
HDL cholesterol; mmol/l.	HDL cholesterol	0.875
Diabetes type 1	diabetes type	0.726
Total cholesterol in HDL; mmol/l	Total cholesterol in mmol/l	0.948
Total cholesterol to total lipids ratio in medium HDL; %	Total cholesterol	0.768
Index of physical activity 2001.	physical activity	0.756
other	other	0.905
Insulin	insulin	0.680
Weight of mother in 1986; kg	Weight	0.738
Heart rate mean value when sitting, from the status form; Heart rate/min	heart rate	0.675
Age when ovariectomy for both ovary?	Age	0.694
Index of physical activity 2007.	physical activity	0.756
Size, LCx PD	size	0.667
Date of installation of the ICD	date	0.616
Lauric acid C12 g	uric acid	0.732
Marital status since the year	marital status	0.792
Triglycerides to total lipids ratio in medium HDL; %	triglycerides	0.718
Hematocrit; Osuus	hematocrit	0.810
Diabetes	diabetes	0.837
Number of study years	Number of days	0.710
recorded baseline mortality	metabolic syndrome	0.618
Date of testing	date	0.650
Total cholesterol in LDL; mmol/l	Total cholesterol in mmol/l	0.948
glucose score, 0 = glucose at least 5.6 or diabetes, 1 = glucose at least 5.6 and not diabetes	glucose	0.691
date of birth	Date of birth	0.949
Total cholesterol to total lipids ratio in large VLDL; %	Total cholesterol	0.768
Total cholesterol in IDL; mmol/l	Total cholesterol in mmol/l	0.948

Terminologies from OCVD	Terminologies from OMD	Score
Stenosis, RCA dist	Carotid stenosis	0.639
Metabolic syndrome according to the EGIR definition	among my friends I feel comfortable	0.611
Size, LIM	size	0.694
I don't know what really is happening in the innermost me	Pain disappear, remain when standing or walking slowly	0.647
Total cholesterol in very large HDL; mmol/l	Total cholesterol	0.798
Date of angiography	date	0.636
ASA used when returning home or switching to follow-up care	When complaints for the first time occurred in the life	0.642
Apolipoprotein E	Lipoprotein	0.761
Marital status of the participant, 1 = unmarried, 2 = married, 3 = cohabiting, 4 = divorced/legally separated, 5 = widow	marital status	0.679
Triglycerides to total lipids ratio in chylomicrons and extremely large VLDL; %	triglycerides	0.692
Delay of STEMI patients from ECG to arterial puncture	Difficulty concentrating / decision problems	0.618
Amount of cigars/cigarillos smoked per day	Difficulty concentrating / decision problems	0.633
Room temperature in degrees celsius	current setting and compliance	0.608
Triglycerides in medium HDL; mmol/l	triglycerides	0.755
Size, RCA PD	size	0.667
C-reactive protein from 1980	C-reactive protein	0.881
Total cholesterol 2nd measurement, F5.2,1129	Total cholesterol	0.795
Diastolic Blood Pressure	Target diastolic blood pressure	0.672
Iron mg	morning	0.680
Total cholesterol to total lipids ratio in small VLDL; %	Total cholesterol	0.768
Height at birth; cm,	height	0.694
Fasting	Fasting blood sugar	0.773
Triglycerides to total lipids ratio in large LDL; %	triglycerides	0.719
Triglycerides in large VLDL; mmol/l	triglycerides	0.755
Weight of father in 1980; kg	Weight	0.738
Total cholesterol to total lipids ratio in medium VLDL; %	Total cholesterol	0.766
Height of mother; cm. Average from values collected between 1980 and 1989.	height	0.634

Terminologies from OCVD	Terminologies from OMD	Score
Total cholesterol in small VLDL; mmol/l	Total cholesterol	0.812
Total cholesterol in HDL3; mmol/l	Total cholesterol in mmol/l	0.939
Total cholesterol in large LDL; mmol/l	Total cholesterol in mg/dl	0.853
Triglycerides to total lipids ratio in large VLDL; %	triglycerides	0.718
HDL cholesterol; mmol/l	HDL cholesterol	0.884
Size, LAD DG II	size	0.650

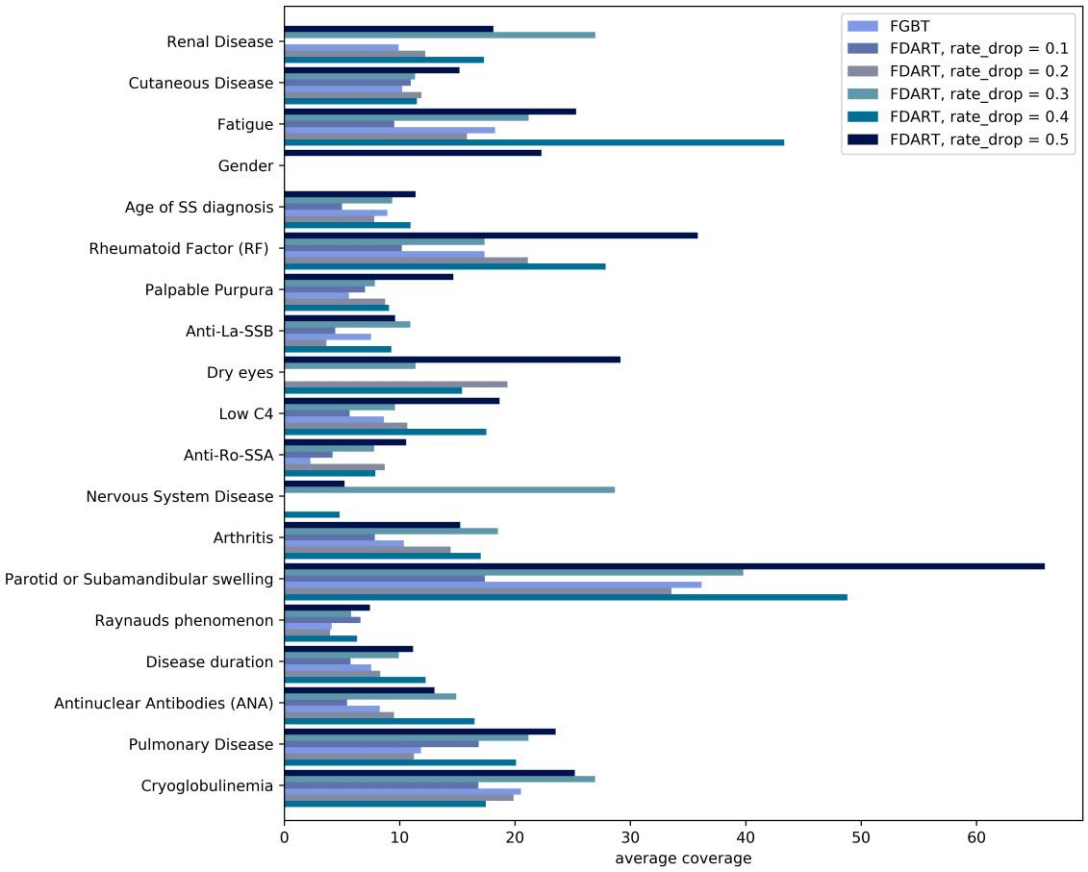
Supplementary Table 4. A summary of the input features (and their corresponding abbreviations) including those having either good or fair quality within any time point between time-points 1 and 4 and in the baseline.

Time-series clinical data		
No.	**Feature**	**Abbreviation**
1	SBP	Systolic Blood Pressure
2	DBP	Diastolic Blood Pressure
3	temperature	-
4	dyspnea	-
5	tachypnea	-
6	SatO2	Oxygen saturation
7	cardiac_frequency	-
8	WBC	White Blood Cell Count
9	Neut_percent	Percentage of neutrophils
10	Neut_abs_number	Absolute number of neutrophils
11	Lymph_percent	Percentage of lymphocytes
12	Lymph_abs_number	Absolute number of lymphocytes
13	Hb	Hemoglobin
14	Hct	Hematocrit
15	MCV	Mean Corpuscular Volume (red blood cells)
16	PLT	Number of platelets
17	Glu	
18	Urea	-
19	Creatinine	-
20	Tbil	Total Bilirubin
21	AST	Aspartate Aminotransferase
22	ALT	Alanine Aminotransferase
23	ALP	Alkaline Phosphatase Level
24	LDH	Lactate Dehydrogenase

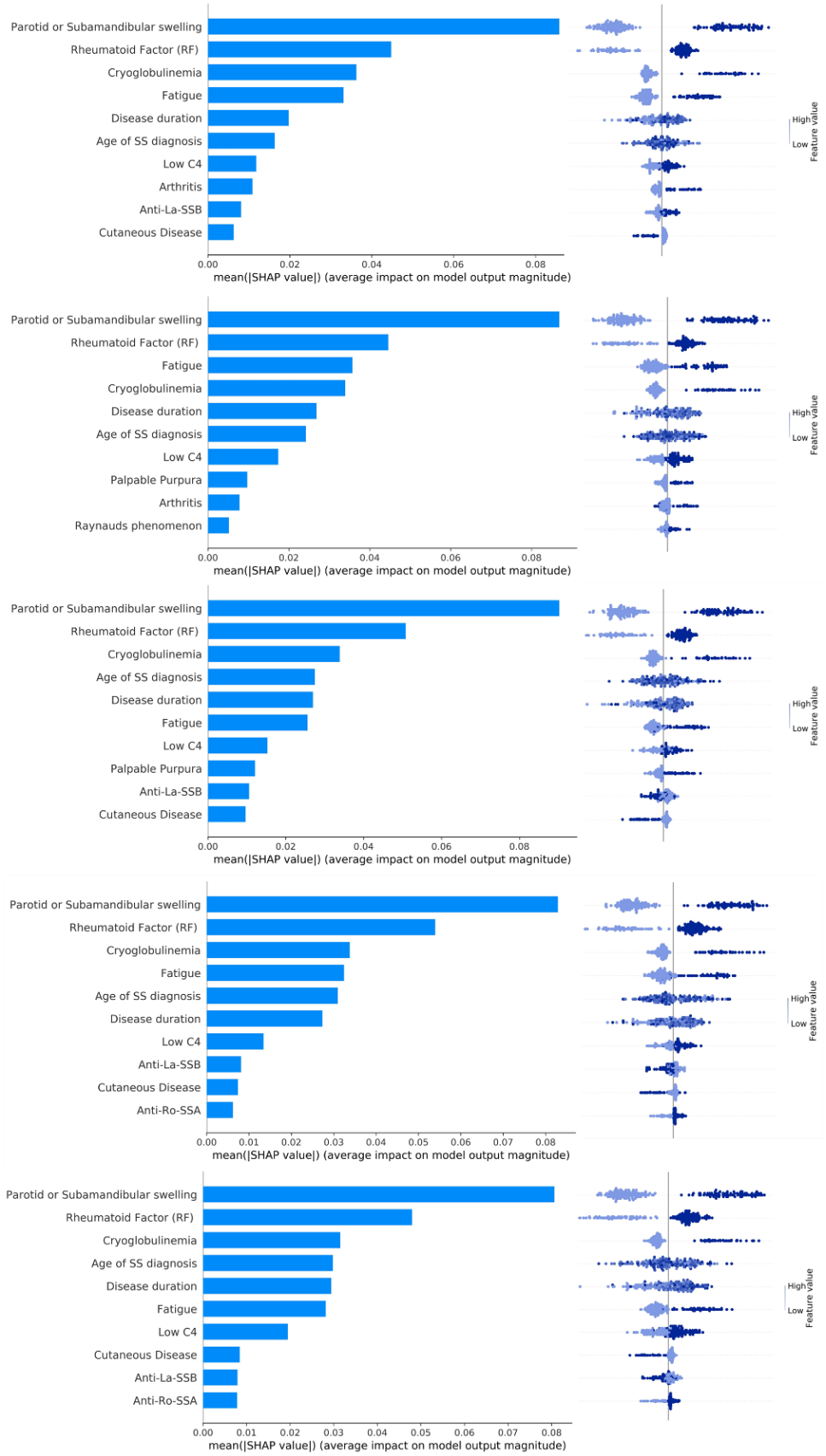
Time-series clinical data		
No.	**Feature**	**Abbreviation**
25	CK	Creatinine Kinase
26	Na	Natrium
27	K	Kalium
28	CRP	C-reactive protein
29	med_chloroquine	Chloroquine administration
30	med_hydroxychloroquine	Hydroxychloroquine administration
31	med_azithromycin	Azithromycin administration
32	med_colchiquine	Colchicine administration
33	med_CAP_antibiotics	Community-acquired pneumonia antibiotics administration
34	med_O2_supply	Oxygen supply
35	O2_supply_type	Oxygen supply type
36	O2_supply_lit	Oxygen supply in liters
37	secondary_O2_supply_lit	Secondary Oxygen supply in liters
38	FiO2	Fraction of inspired oxygen
39	PO2_FiO2_ratio	Ratio of arterial oxygen partial pressure (PaO2 in mmHg) to fractional inspired oxygen (FiO2 expressed as a fraction, not a percentage)
40	PO2	Partial pressure of oxygen
41	med_tocilizumab	Tocilizumab administration
42	med_low_dose_steroids	Systemic corticosteroid administration (low dose)
43	Biguanides	Biguanides administration
44	TZDs	Thiazolidinediones administration
45	GLP1_RAs	Glucagon-like peptide 1 receptor agonists administration
46	DPP4i	Dipeptidyl Peptidase-4 Inhibitor administration
47	SGLT2i	Sodium-glucose cotransporter 2 inhibitors administration
48	Insulin_long_acting	Long-acting insulin administration
49	Insulin_long_acting_dose	Long-acting insulin administration dose
50	Insulin_short_acting	Short-acting insulin administration
51	Insulin_short_acting_dose	Short-acting insulin administration dose
52	d_dimers	D-Dimer test
53	hs_TPN	Troponin
54	ABD	A (ABO1), B (ABO2) and D (RH1) antigens on red blood cells
55	AMS	Amylase
56	APTT	Activated Partial Thromboplastin Time

Time-series clinical data		
No.	**Feature**	**Abbreviation**
57	CTchest	Computer tomography of the chest
58	CXR	Chest X-Ray
59	Cl	Chloride
60	Dbil	Direct bilirubin
61	FER	Ferritin
62	HCO3	Bicarbonates
63	LAC	Lupus Anticoagulant Testing
64	Mg	Magnesium
65	Mono	Mononucleosis
66	PCO2	Partial Pressure of Carbon Dioxide (arterial)
67	Ph	pH (arterial)
68	Sulfonylureas	-
69	TCA	Tricyclic Antidepressant
70	fibrogen	Fibrinogen
71	gGT	Gamma-glutamyl Transferase
72	INR	International Normalized Ratio
Demographics		
No.	**Feature**	**Abbreviation**
1	Age	-
2	Weight	-
3	Height	-
4	risk_factor_obesity	-
5	risk_factor_age50	Age above 50 years (yes/no)
6	risk_factor_DM	Diabetes mellitus (yes/no)
7	risk_factor_Hypertension	Hypertension (yes/no)
8	risk_factor_COPD	Chronic Obstructive Pulmonary Disease (yes/no)
9	risk_factor_Dyslipidemia	Dyslipidemia status (yes/no)
10	risk_factor_Smoking	Smoking status (yes/no)
11	risk_factor_CKD	Chronic Kidney Disease (yes/no)
12	risk_factor_immunosupression	Immunosupression (yes/no)
Baseline clinical data (symptoms)		
No.	**Feature**	**Abbreviation**
1	symptom_fever	Fever (yes/no)
2	symptom_dry_cough	Dey cough (yes/no)
3	symptom_fatigue	Fatigue(yes/no)

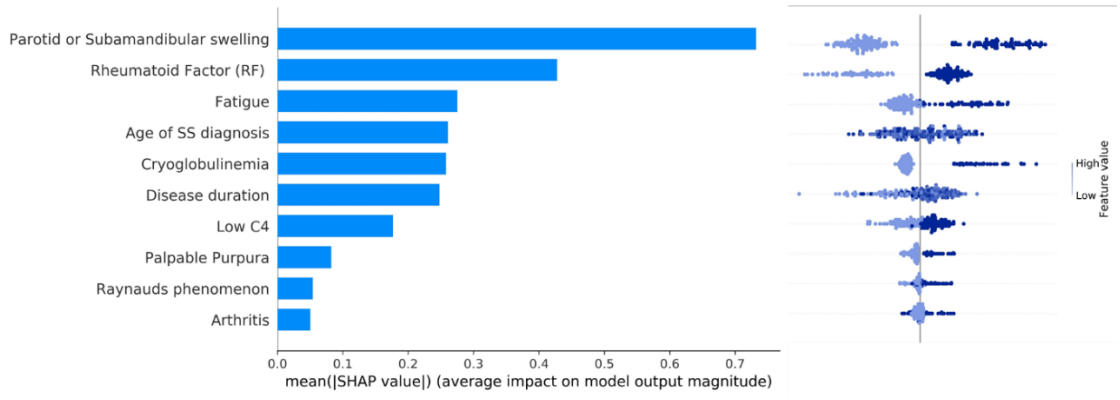
Time-series clinical data		
No.	**Feature**	**Abbreviation**
4	symptom_anorexia	Anorexia (yes/no)
5	symptom_myalgias	Myalgias (yes/no)
6	symptom_dyspnea	Dyspnea (yes/no)
7	symptom_sputum_production	Sputum production (yes/no)
8	symptom_anosmia	Lack of smell (yes/no)
9	symptom_dysgeusia	Lack of taste (yes/no)
10	symptom_GI_tract	Gastrointestinal tract abnormalities (yes/no)
11	symptom_headache	Headache (yes/no)
12	symptom_sore_throat	Sore throat (yes/no)
13	symptom_rhinorrhea	Runny nose (yes/no)
14	duration_of_symptoms	Time duration of the symptoms (in days)
Treatments		
No.	**Feature**	**Abbreviation**
1	Statin	Statins (yes/no)
2	ACEi	Angiotensin-converting-enzyme inhibitors (yes/no)
3	Sartan	Angiotensin-II-receptor antagonists (yes/no)
4	CaBlocker	Calcium channel blockers (yes/no)
5	Diuretic	Diuretics (yes/no)
6	Betablocker	Beta blockers (yes/no)
7	Biguanides	Biguanides (yes/no)
8	Sulfonylureas	Sulfonylureas (yes/no)
9	TZDs	Thiazolidinediones (yes/no)
10	GLP1_RAs	Contemporary classification of glucagon-like peptide 1 receptor agonists (yes/no)
11	DPP4i	Dipeptidyl peptidase-4 inhibitors (yes/no)
12	SGLT2i	Sodium-glucose Cotransporter 2 Inhibitors (yes/no)
13	Insulin_long_acting	Long-acting insulin (yes/no)
14	Insulin_long_acting_dose	Long-acting insulin dose
15	Insulin_short_acting	Short-acting insulin (yes/no)
16	Insulin_short_acting_dose	Short-acting insulin dose
17	PPI	Proton pump inhibitors (yes/no)
18	Corticosteroids	-
19	other_immunomodulators	Other immunomodulators (yes/no)
20	AntiXa	Anti-Xa heparin (yes/no)
21	Coumarin	-



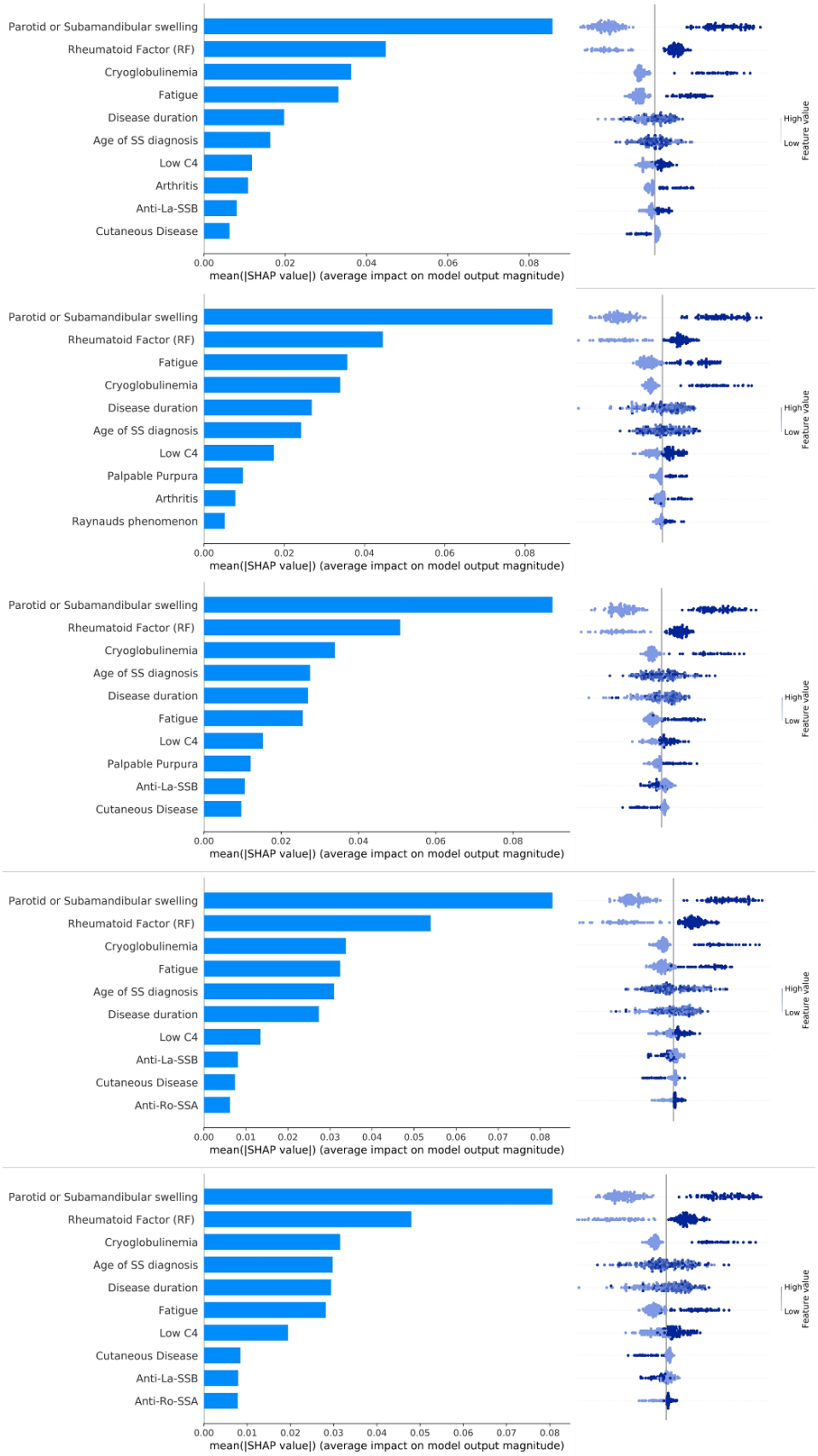
Supplementary Figure 1. The average coverage for each federated tree ensemble algorithm in federated scenario 1 which quantifies the average number of observations that passed through this feature (node) during the node splitting process.



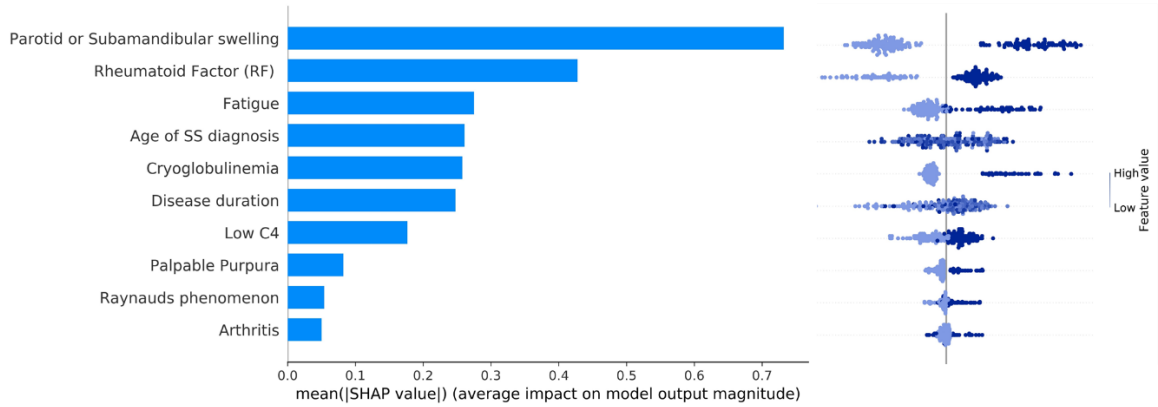
Supplementary Figure 2. An illustration of the SHAP plot in federated scenario 2 for the FDART schemas.



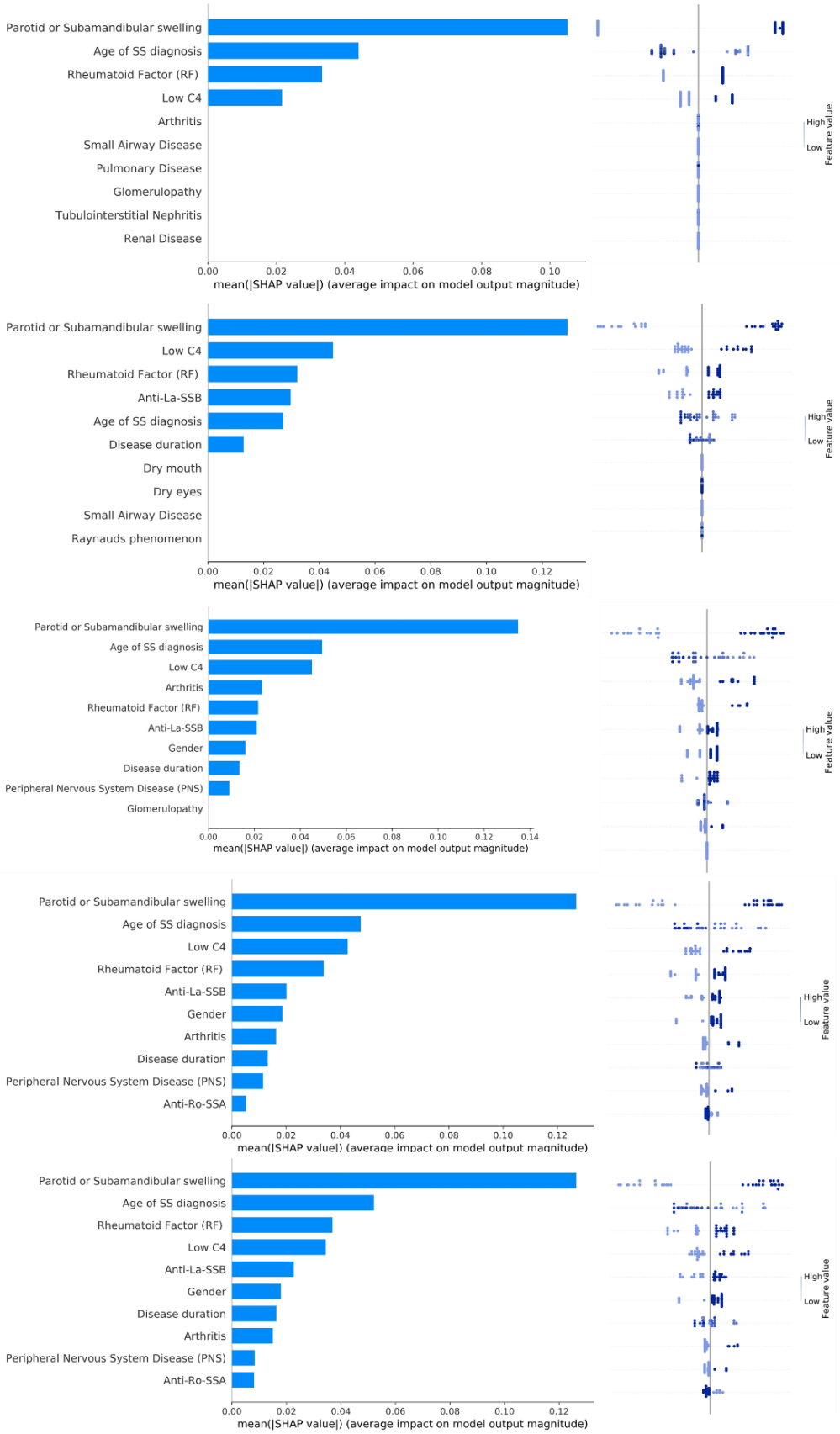
Supplementary Figure 3. An illustration of the SHAP plot in federated scenario 2 for the FDART schemas.



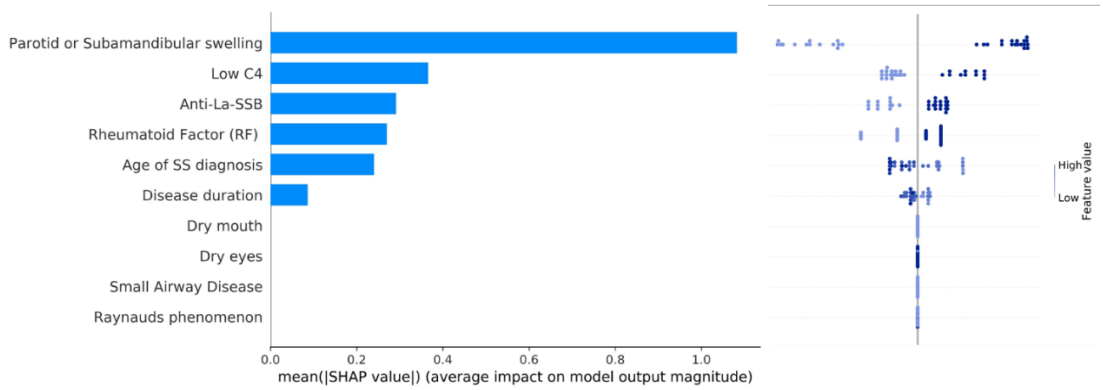
Supplementary Figure 4. An illustration of the SHAP plot in federated scenario 3 for the FDART schemas.



Supplementary Figure 5. An illustration of the SHAP plot in federated scenario 3 for the FDART schemas.



Supplementary Figure 6. An illustration of the SHAP plot in federated scenario 4 for the FDART schemas.



Supplementary Figure 7. An illustration of the SHAP plot in federated scenario 4 for the FDART schemas.

Short CV and related publications

Vasileios C. Pezoulas received the Diploma degree in Electrical and Computer Engineering from the Technical University of Crete, Chania, Greece, in 2015 and the MSc degree in Electrical and Computer Engineering from the same institute, in 2017. During his master studies he received a scholarship of excellence from the Pancretan Endowment Fund for conducting research at the Technical University of Crete in the field of biomedical engineering. From 2017 he is a PhD student in biomedical engineering at the Unit of Medical Technology and Intelligent Information Systems, University of Ioannina, Ioannina, Greece, with a great experience in the analysis and processing of big medical and other related data. Since 2019 he holds a scholarship from the Hellenic Foundation for Research and Innovation (HFRI) for conducting research as a PhD student in biomedical engineering and particularly in the field of data science and AI modeling in healthcare. His research interests include machine learning, artificial intelligence, data sharing, data curation, data harmonization, and synthetic data generation, among others.

Publications

Book(s)

1. **V. C. Pezoulas**, T. P. Exarchos and D. I. Fotiadis, "Medical data sharing, harmonization and analytics," Academic Press, Elsevier, 2022.

International conferences

1. **V. C. Pezoulas**, T. P. Exarchos, A. G. Tzioufas, and D. I. Fotiadis, "Multiple additive regression trees with hybrid loss for classification tasks across heterogeneous clinical data in distributed environments: a case study," In Proceedings of the 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2021. doi: 10.1109/EMBC46164.2021.9629912.
2. **V. C. Pezoulas**, F. Kalatzis, T. P. Exarchos, L. Chatzis, S. Gandolfo, A. Goules, A. G. Tzioufas, S. De Vita, and D. I. Fotiadis, "A federated AI strategy for the classification of patients with Mucosa Associated Lymphoma Tissue (MALT) lymphoma across multiple harmonized cohorts," In Proceedings of the 43rd Annual

International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2021. doi: 10.1109/EMBC46164.2021.9630014.

3. **V. C. Pezoulas**, A. Sakellarios, M. Kleber, J. A. Bosch, S. W. van der Laan, F. Lamers, T. Lehtimäki, W. März, and D. I. Fotiadis, "A hybrid data harmonization workflow using word embeddings for the interlinking of heterogeneous cross-domain clinical data structures," In Proceedings of the IEEE EMBS International Conference on Biomedical and Health Informatics (BHI) 2021. doi: 10.1109/BHI50953.2021.9508484.
4. **V. C. Pezoulas**, G. I. Grigoriadis, N. S. Tachos, F. Barlocco, I. Olivotto, and D. I. Fotiadis, "Generation of virtual patient data for in-silico cardiomyopathies drug development using tree ensembles: a comparative study," In Proceedings of the 42nd IEEE EMBC, pp. 5343-5346, 2020.
5. **V. C. Pezoulas**, G. I. Grigoriadis, N. S. Tachos, F. Barlocco, I. Olivotto, and D. I. Fotiadis, "Variational Gaussian Mixture Models with robust Dirichlet concentration priors for virtual population generation in hypertrophic cardiomyopathy: a comparison study," In Proceedings of the 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2021. doi: 10.1109/EMBC46164.2021.9629653.
6. **V. C. Pezoulas**, F. Kalatzis, T. P. Exarchos, A. Goules, S. Gandolfo, E. Zampeli, F. Skopouli, S. De Vita, A. G. Tzioufas, and D. I. Fotiadis, D. I., "Dealing with Open Issues and Unmet Needs in Healthcare Through Ontology Matching and Federated Learning," In Proceedings of the 8th European Medical and Biological Engineering Conference (EMBEC), Springer, pp. 306- 313, 2020. doi: 10.1007/978-3-030-64610-3_36.
7. **V. C. Pezoulas**, T. P. Exarchos, K. D. Kourou, A. G. Tzioufas, S. De Vita, and D. I. Fotiadis, "Utilizing Incremental Learning for the Prediction of Disease Outcomes Across Distributed Clinical Data: A Framework and a Case Study," In Proceedings of the Mediterranean Conference on Medical and Biological Engineering and Computing, pp. 823-831, 2019. doi: 10.1007/978-3-030-31635-8_98.
8. **V. C. Pezoulas**, T. P. Exarchos, A. G. Tzioufas, S. De Vita, and D. I. Fotiadis, "Predicting lymphoma outcomes and risk factors in patients with primary Sjögren's Syndrome using gradient boosting tree ensembles," In Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2165-2168, 2019. doi: 10.1109/EMBC.2019.8857557.

9. **V. C. Pezoulas**, K. D. Kourou, F. Kalatzis, T. P. Exarchos, A. Venetsanopoulou, E. Zampeli, S. Gandolfo, F. Skopouli, S. De Vita, A. G. Tzioufas, and D. I. Fotiadis, "AB0166 – Enhancing the quality of clinical data through data curation in primary Sjögren's Syndrome," Annals of the Rheumatic Diseases, vol. 78, pp. 1541-1542, 2019. doi: 10.1136/annrheumdis-2019- eular.4139.
10. **V. C. Pezoulas**, K. D. Kourou, T. P. Exarchos, V. Andronikou, T. Varvarigou, A. G. Tzioufas, S. De Vita, and D. I. Fotiadis, "A de-centralized framework for data sharing, ontology matching and distributed analytics," In Proceedings of the 1st International Workshop on Semantic Web Technologies for Health Data Management, 2018.
11. **V. C. Pezoulas**, T. P. Exarchos, V. Andronikou, T. Varvarigou, A. Tzioufas, S. De Vita, Dimitrios I. Fotiadis, "Towards the establishment of a biomedical ontology for the primary Sjögren's Syndrome," In Proceedings of the 40th International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 17-21, 2018. doi: 10.1109/EMBC.2018.8513349.
12. **V. C. Pezoulas**, N. Tachos, N., and D. I. Fotiadis. Generation of Virtual Patients for in Silico Cardiomyopathies Drug Development. In 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE), pp. 671-674, 2019. doi: 10.1109/BIBE.2019.00126.
13. **V. Pezoulas**, T. Exarchos, A. Venetsanopoulou, E. Zampeli, S. Gandolfo, S. De Vita, F. N. Skopouli, A. Tzioufas, and D. I. Fotiadis, "AB0166 Enhancing the quality of clinical data through data curation in primary Sjögren's Syndrome," Annals of the Rheumatic Diseases, vol. 78, pp. 1541-1542, 2019.
14. A. Goules, L. Chatzis, **V. Pezoulas**, C. Baldini, F. Skopouli, A. Venetsanopoulou, P Voulgari, S. De Vita, M. Voulgarelis, H. M. Moutsopoulos, D. Fotiadis, and A. Tzioufas, "OP0291 – Severity of Labial Minor salivary gland focus score and future lymphoma development in Sjögren's Syndrome," Annals of the Rheumatic Diseases, vol. 80, Suppl 1, pp. 178-179, 2021. doi: 10.1136/annrheumdis-2021-eular.2539.
15. L. Chatzis, **V. Pezoulas**, A. Goules, I. Stergiou, C. Mavragani, H. M. Moutsopoulos, M. Voulgarelis, D. Fotiadis, and A. Tzioufas, "POS0290 – Predicting risk factors of MALT lymphoma in Sjögren's Syndrome," Annals of the Rheumatic Diseases, vol. 80, Suppl 1, pp. 370-371, 2021. doi: 10.1136/annrheumdis-2021-eular.2260.

16. G. I. Grigoriadis, **V. C. Pezoulas**, M. Roumpi, G. Gkois, N. S. Tachos, M. Prodanovic, D. Prodanovic, B. Stojanovic, S. M. Mijailovich, N. Philipovic, and D. I. Fotiadis, "Towards the development of a unified virtual population model in hypertrophic cardiomyopathy," Accepted for presentation at the IEEE EMBS International Conference on Biomedical and Health Informatics (BHI) 2021. doi: 10.1109/BHI50953.2021.9508598.
17. G. I. Grigoriadis, M. Roumpi, D. Zaridis, **V. C. Pezoulas**, A. Rammos, N. S. Tachos, K. N. Naka, and D. I. Fotiadis, "Comparison of three U-Net family architectures for left ventricular myocardial wall automatic segmentation," In Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 2932-2935, 2021. doi: 10.1109/BHI50953.2021.9508598.
18. M. D. Mantzaris, V. T. Potsika, P. Siogkas, V. I. Kigka, **V. C. Pezoulas**, I. G. Pappas, T. P. Exarchos, I. B. Koncar, J. Pelisek, E. Andreakos, and D. I. Fotiadis, "A multimodal advanced approach for the stratification of carotid artery disease," In Proceedings of the IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE), pp. 706-709, 2019. doi: 10.1109/BIBE.2019.00133.
19. L. Chatzis, **V. Pezoulas**, F. Ferro, V. Donati, A. Venetsanopoulou, E. Zampeli, M. Mavromati, P. Voulgari, C. Mavragani, D. Fotiadis, F. Skopouli, S. De Vita, G. Vassilis, C. Baldini, H. M. Moutsopoulos, A. Goules, and A. Tzioufas., "FRI0161 - Phenotypic differences between Sjögren's Syndrome patients with low and high-grade inflammation based on salivary gland focus score," Annals of the Rheumatic Diseases, vol. 79, pp. 664, 2020. doi: 10.1136/annrheumdis-2020-eular.4335.
20. O. Argyropoulou, **V. Pezoulas**, L. Quartuccio, F. Ferro, S. Gandolfo, V. Donati, A. Venetsanopoulou, L. Chatzis, E. Zampeli, M. Mavromati, P. Voulgari, C. Mavragani, C. Baldini, F. Skopouli, D. Fotiadis, M. Galli, S. De Vita, H. M. Moutsopoulos, A. Goules, and A.Tzioufas, "THU0294 - The differences in the clinical spectrum of cryoglobulinemic vasculitis between Sjögren's Syndrome and HCV hepatitis," Annals of the Rheumatic Diseases, vol. 79, pp. 375, 2020. doi: 10.1136/annrheumdis-2020-eular.4233.
21. L. Chatzis, **V. Pezoulas**, F. Ferro, V. Donati, A. Venetsanopoulou, E. Zampeli, M. Mavromati, P. Voulgari, C. Mavragani, D. Fotiadis, F. Skopouli, S. De Vita, C. Baldini, H. M. Moutsopoulos, A. Tzioufas, and A. Goules, "OP0096 - The differences between Sjögren's Syndrome patients with combined seronegativity

and Anti-Ro/SSA seropositivity," Annals of the Rheumatic Diseases, vol. 79, pp. 63, 2020. doi: 10.1136/annrheumdis-2020-eular.4203.

22. A. Goules, O. Argyropoulou, **V. Pezoulas**, F. Ferro, S. Gandolfo, V. Donati, M. Binutti, S. Z. Callegher, L. Chatzis, A. Venetsanopoulou, E. Zampeli, M. Mavromati, P. Voulgari, C. Mavragani, C. Baldini, F. Skopouli, D. Fotiadis, S. De Vita, H. M. Moutsopoulos, and A. Tzioufas, "FRI0149 – The clinical features of Sjögren's Syndrome patients with early and late disease onset," Annals of the Rheumatic Diseases, vol. 79, pp. 658-659, 2020. doi: 10.1136/annrheumdis-2020-eular.4094.
23. P. Panagopoulos, V. E. Georgakopoulou, **V. Pezoulas**, A. Goules, D. I. Fotiadis, and T. Vassilakopoulos, "POS0894 Comparison of pulmonary and small airways function between idiopathic and inflammatory myopathies patients with and without interstitial lung disease," Annals of the Rheumatic Diseases, vol. 81, pp. 744, 2022.
24. T. Androutsakos, T. Voulgaris, A. D. Bakasis, M. L. Koutsompina, L. Chatzis, O. Argyropoulou, **V. Pezoulas**, D. I. Fotiadis, A. Goules, G. Papatheodoridis, and A. Tzioufas, "AB0565 Prevalence of liver fibrosis assessed by transient elastography in patients with Sjögren's Syndrome," Annals of the Rheumatic Diseases, vol. 81, pp. 1409, 2022.
25. K. Bitzogli, E. Jahaj, A. D. Bakasis, E. Kapsogeorgou, A. Goules, I. Stergiou, **V. Pezoulas**, P. Skendros, K. Ritis, D. I. Fotiadis, A. Kotanidou, A. Tzioufas, and P. Vlachoyiannopoulos, "POS1240 High prevalence of serum autoantibodies in severely ill COVID-19 patients hospitalized in the intensive care unit," Annals of the Rheumatic Diseases, vol. 81, pp. 954, 2022.
26. L. Chatzis, **V. Pezoulas**, A. Goules, I. Stergiou, C. Mavragani, G. Tsouroufli, D. Fotiadis, H. M. Moutsopoulos, M. Voulgarelis, and A. Tzioufas, "OP0294 - Sjögren's Syndrome associated lymphomas: Clinical description and 10-year survival," Annals of the Rheumatic Diseases, vol. 80, no. 1, pp. 180-181, 2021. doi: 10.1136/annrheumdis-2021-eular.2400.
27. A. I. Sakellarios, **V. C. Pezoulas**, C. Bourantas, K. K. Naka, L. K. Michalis, P. W. Serruys, G. Stone, H. M. Garcia-Garcia, and D. I. Fotiadis, "Prediction of atherosclerotic disease progression combining computational modelling with machine learning," In Proceedings of the 42nd Annual International Conference of

the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 2760-2763, 2020. doi: 10.1109/EMBC44109.2020.9176435.

Peer-reviewed journals

1. **V. C. Pezoulas**, Goules, A., Kalatzis, F., Chatzis, L., Kourou, K. D., Venetsanopoulou, A., T. P. Exarchos, S. Gandolfo, K. Votis, E. Zampeli, J. Burmeister, T. May, M. M. Pérez, I. Lishchuk, T. Chondrogiannis, V. Andronikou, T. Varvarigou, N. Filipovic, M. Tsiknakis, C. Baldini, M. Bombardieri, H. Bootsma, S. J. Bowman, M. S. Soyfoo, D. Parisis, C. Delporte, V. Devauchelle-Pensec, J. O. Pers, T. Dörner, E. Bartoloni, R. Gerli, R. Giacomelli, R. Jonsson, W. F. Ng, R. Priori, M. Ramos-Casals, K. Sivils, F. Skopouli, W. Torsten, J. A. G. van Roon J, X. Mariette, S. De Vita, A. G. Tzioufas, and D. I. Fotiadis, "Addressing the clinical unmet needs in primary Sjögren's Syndrome through the sharing, harmonization and federated analysis of 21 European cohorts," Computational and structural biotechnology journal, vol. 20, 471-482, 2022. doi: 10.1016/j.csbj.2022.01.002.
2. **V. C. Pezoulas**, C. Papaloukas, M. Veyssiere, A. Goules, A. G. Tzioufas, V. Soumelis, and D. I. Fotiadis, "A computational workflow for the detection of candidate diagnostic biomarkers of Kawasaki disease using time-series gene expression data," Computational and Structural Biotechnology Journal, vol. 19, pp. 3058-3068, 2021. doi: 10.1016/j.csbj.2021.05.036.
3. **V. C. Pezoulas**, G. I. Grigoriadis, G. Gkois, N. S. Tachos, T. Smole, Z. Bosnić, M. Pičulin, I. Olivotto, F. Barlocco, M. Robnik-Šikonja, D. G. Jakovljevic, A. Goules, A. G. Tzioufas, and D. I. Fotiadis, "A computational pipeline for data augmentation towards the improvement of disease classification and risk stratification models: a case study in two clinical domains," Computers in Biology and Medicine, pp. 104520, 2021.
4. **V. C. Pezoulas**, K. D. Kourou, E. Mylona, C. Papaloukas, A. Liontos, D. Biros, D., O. I. Milionis, C. Kyriakopoulos, K. Kostikas, H. Milionis, and D. I. Fotiadis, "ICU admission and mortality classifiers for COVID-19 patients based on subgroups of dynamically associated profiles across multiple timepoints," Computers in biology and medicine, 105176, 2021. doi: 10.1016/j.compbiomed.2021.105176.

5. **V. C. Pezoulas**, Kourou, K. D., Papaloukas, C., Triantafyllia, V., Lampropoulou, V., Siouti, E., M. Papadaki, M. Salagianni, E. Koukaki, N. Rovina, A. Koutsoukou, E. Andreakos, and D. I. Fotiadis, "A Multimodal Approach for the Risk Prediction of Intensive Care and Mortality in Patients with COVID-19," Diagnostics, vol. 12, no 1, pp. 56, 2022. doi: 10.3390/diagnostics12010056.
6. **V. C. Pezoulas**, K. D. Kourou, F. Kalatzis, T. P. Exarchos, E. Zampeli, S. Gandolfo, A. Goules, C. Baldini, F. Skopouli, S. De Vita, A. G. Tzioufas, and D. I. Fotiadis, "Overcoming the barriers that obscure the interlinking and analysis of clinical data through harmonization and incremental learning," IEEE Open Journal of Engineering in Medicine and Biology (OJEMB), vol. 1, pp. 83-90, 2020. doi: 10.1109/OJEMB.2020.2981258.
7. **V. C. Pezoulas**, K. D. Kourou, F. Kalatzis, T. P. Exarchos, A. Venetsanopoulou, E. Zampeli, S. Gandolfo, F. Skopouli, S. De Vita, A. G. Tzioufas, and D. I. Fotiadis, "Medical data quality assessment: On the development of an automated framework for medical data curation," Computers in biology and medicine, vol. 107, pp. 270-283, 2019. doi: 10.1016/j.compbiomed.2019.03.001.
8. **V. C. Pezoulas**, K. D. Kourou, F. Kalatzis, T. P. Exarchos, A. Venetsanopoulou, E. Zampeli, S. Gandolfo, F. Skopouli, S. De Vita, A. G. Tzioufas, and D. I. Fotiadis, "Enhancing medical data quality through data curation: a case study in primary Sjögren's syndrome," Clinical and experimental rheumatology, vol. 37, no 3, pp. 90-96, 2019. PMID: 31287405.
9. **V. C. Pezoulas**, N. S. Tachos, G. Gkois, I. Olivotto, F. Barlocco, and D. I. Fotiadis, "Bayesian Inference-Based Gaussian Mixture Models With Optimal Components Estimation Towards Large-Scale Synthetic Data Generation for In Silico Clinical Trials," IEEE Open Journal of Engineering in Medicine and Biology (OJEMB), vol. 3, pp. 108-114, 2022.
10. **V. C. Pezoulas**, O. Hazapis, N. Lagopati, T. P. Exarchos, A. V. Goules, A. G. Tzioufas, D. I. Fotiadis, I. G. Stratis, A. N. Yannacopoulos, and V. G. Gorgoulis, "Machine Learning Approaches on High Throughput NGS Data to Unveil Mechanisms of Function in Biology and Disease," Cancer Genomics & Proteomics, vol. 18, no 5, pp. 605-626, 2021.
11. A. G. Tzioufas, A. D., Bakasis, A. V. Goules, K. Bitzogli, I. I. Cinoku, L. G. Chatzis, O. D. Argyropoulou, A. I. Venetsanopoulou, M. Mavrommati, I. E. Stergiou, **V. Pezoulas**, P. V. Voulgari, C. Katsimpari, S. Katechis, S. Gazi, G.

Katsifis, C. I. Sfontouris, A. I. Georgountzos, S-N Liossis, C. Papagoras, D. I. Fotiadis, F. N. Skopouli, P. G. Vlachoyiannopoulos, and H. M. Moutsopoulos, “A prospective multicenter study assessing humoral immunogenicity and safety of the mRNA SARS-CoV-2 vaccines in Greek patients with systemic autoimmune and autoinflammatory rheumatic diseases,” Journal of autoimmunity, vol. 125, pp. 102743, 2021.

12. L. Chatzis, A. V. Goules, **V. Pezoulas**, C. Baldini, S. Gandolfo, F. N. Skopouli, T. P. Exarchos, E. K. Kapsogeorgou, V. Donati, P. V. Voulgari, C. P. Mavragani, V. Gorgoulis, S. De Vita, D. I. Fotiadis, M. Voulgarelis, H. M. Moutsopoulos, and A. G. Tzioufas, A. G, “A biomarker for lymphoma development in Sjogren's syndrome: Salivary gland focus score,” Journal of autoimmunity, vol. 121, pp. 102648, 2021.
13. T. Smole, B. Žunkovič, M. Pičulin, E. Kokalj, M. Robnik-Šikonja, M. Kukar, D. I. Fotiadis, **V. C. Pezoulas**, N. S. Tachos, F. Barlocco, F. Mazzarotto, D. Popović, L. Maier, L. Velicki, G. A. MacGowan, I. Olivotto, N. Filipović, D. G. Jakovljević, and Z. Bosnić, “A machine learning-based risk stratification model for ventricular tachycardia and heart failure in hypertrophic cardiomyopathy,” Computers in biology and medicine, pp. 104648, 2021.
14. A. V. Goules, O. D. Argyropoulou, **V. C. Pezoulas**, L. Chatzis, E. Critselis, S. Gandolfo, F. Ferro, M. Binutti, V. Donati, S. Z. Callegher, A. Venetsanopoulou, E. Zampeli, M. Mavrommati, P. V. Voulgari, T. Exarchos, C. P. Mavragani, C. Baldini, F. N. Skopouli, D. I. Fotiadis, S. De Vita, H. M. Moutsopoulos, A. G. Tzioufas, “Primary Sjögren’s Syndrome of Early and Late Onset: Distinct Clinical Phenotypes and Lymphoma Development,” Frontiers in immunology, vol. 11, pp. 2707. doi: doi: 10.3389/fimmu.2020.594096.
15. L. Chatzis, **V. C. Pezoulas**, F. Ferro, S. Gandolfo, V. Donati, M. Binutti, S. Z. Callegher, A. Venetsanopoulou, E. Zampeli, M. Mavrommati, O. D. Argyropoulou, G. Michalopoulos, P. V. Voulgari, T. Exarchos, C. Baldini, F. N. Skopouli, D. I. Fotiadis, S. De Vita, H. M. Moutsopoulos, A. G. Tzioufas, A. V. Goules, “Sjögren’s Syndrome: The Clinical Spectrum of Male Patients,” Journal of clinical medicine, vol. 9, no. 8, pp. 2620, 2020. doi: 10.3390/jcm9082620.
16. K. D. Kourou, **V. C. Pezoulas**, E. I. Georga, T. Exarchos, C. Papaloukas, M. Voulgarelis, A. Goules, A. Nezos, A. G. Tzioufas, H. M. Moutsopoulos, C. Mavragani, and D. I. Fotiadis, “Predicting Lymphoma Development by Exploiting

Genetic Variants and Clinical Findings in a Machine Learning-Based Methodology With Ensemble Classifiers in a Cohort of Sjögren's Syndrome Patients," IEEE Open Journal of Engineering in Medicine and Biology (OJEMB), vol. 1, pp. 49-56, 2020. doi: 10.1109/OJEMB.2020.2965191.

17. A. V. Goules, T. P. Exarchos, **V. C. Pezoulas**, K. D. Kourou, A. I. Venetsanopoulou, S. De Vita, D. I. Fotiadis, and A. G. Tzioufas, "Sjögren's syndrome towards precision medicine: the challenge of harmonisation and integration of cohorts," Clinical and Experimental Rheumatology, vol. 37, no. Suppl 118, pp. S175-84, 2019. PMID: 31464663.
18. K. D. Kourou, **V. C. Pezoulas**, E. I. Georga, T. P. Exarchos, P. Tsanakas, M. Tsiknakis, T. Varvarigou, S. De Vita, A. Tzioufas, and D. I. Fotiadis, "Cohort harmonization and integrative analysis from a biomedical engineering perspective," IEEE Reviews in Biomedical Engineering journal (RBME), vol. 12, pp. 303-318, 2018. doi: 10.1109/RBME.2018.2855055.
19. M. Pičulin, T. Smole, B. Žunkovič, E. Kokalj, M. Robnik-Šikonja, M. Kukar, D. I. Fotiadis, **V. C. Pezoulas**, N. S. Tachos, F. Barlocco, F. Mazzarotto, D. Popović, L. S. Maier, L. Velicki, I. Olivotto, G. A. MacGowan, D. G. Jakovljević, N. Filipović, Z. Bosnić, "Disease Progression of Hypertrophic Cardiomyopathy: Modeling Using Machine Learning," JMIR Med Inform, vol. 10, no 2, pp. e30483, 2022.
20. O. Hazapis, N. Lagopati, **V. C. Pezoulas**, G. I. Papayiannis, D. I. Fotiadis, D. Skaltsas, V. Vergetis, A. Tsirigos, I. G. Stratis, A. N. Yannacopoulos, and V. G. Gorgoulis, "Machine Learning: A Tool to Shape the Future of Medicine," In: Roy, S.S., Taguchi, YH. (eds) Handbook of Machine Learning Applications for Genomics. Studies in Big Data, vol 103. Springer, Singapore. https://doi.org/10.1007/978-981-16-9158-4_12.
21. T. Androutsakos, T. A. Voulgaris, A. D. Bakasis, M. L. Koutsompina, L. Chatzis, O. D. Argyropoulou, **V. Pezoulas**, D. I. Fotiadis, G. Papatheodoridis, A. G. Tzioufas and A. V. Goules, "Liver Fibrosis in Primary Sjögren's Syndrome," Front. Immunol. 13:889021, 2022. doi: 10.3389/fimmu.2022.889021.
22. A. Zarachi, **V. Pezoulas**, I. Komnos, A. Lianou, O. Milionis, E. Klouras, K. Katsikatsos, D. Fotiadis, I. Kastanioudakis, C. Milionis, and A. Liontos, "Clinical Symptoms in Hospitalized and Self-Quarantined Patients with SARS-CoV-2 Infection in Northwestern Greece-Association with Olfactory and Gustatory Dysfunction," Maedica-a Journal of Clinical Medicine, vol. 17, no 2, 2022.

23. A. Zarachi, **V. Pezoulas**, A. Lianou, A. Tsikou, I. Tsiakas, K. Dinaki, D. Fotiadis, and A. Liontos, “Dizziness in the Emergency Department: Insights and Epidemiological Data–a Population Based Study,” Maedica A Journal of Clinical Medicine, vol. 17, no 1, 2022.
24. A. Zarachi, **V. Pezoulas**, O. Milionis, A. Lianou, E. Klouras, I. Komnos, D. Fotiadis, I. Kastanioudakis, C. Milionis, and A. Liontos, “The Impact of Age and Gender and Their Association with Chemosensory Dysfunction, in Hospitalized and Self-Quarantine Patients with Covid-19 Infection, in Epirus, Greece,” Maedica-a Journal of Clinical Medicine, vol. 17, no 1, 2022.
25. F. M. Aarestrup, A. Albeyatti, W. J. Armitage, C. Auffray, L. Augello, R. Balling, N. Benhabiles, G. Bertolini, J. G. Bjaalie, M. Black, N. Blomberg, P. Bogaert, M. Bubak, B. Claerhout, L. Clarke, B. De Meulder, G. D’Errico, A. Di Meglio, N. Forgo, C. Gans-Combe, A. E. Gray, I. Gut, A. Gyllenberg, G. Hemmrich-Stanisak, L. Hjorth, Y. Ioannidis, S. Jarmalaite, A. Kel, F. Kherif, J. O. Korbel, C. Larue, M. Laszlo, A. Maas, L. Magalhaes, I. Manneh-Vangramberen, E. Morley-Fletcher, C. Ohmann, P. Oksvold, N. P. Oxtoby, I. Perseil, **V. Pezoulas**, O. Riess, H. Riper, J. Roca, P. Rosenstiel, P. Sabatier, F. Sanz, M. Tayeb, G. Thomassen, J. Van Bussel, M. Van den Bulcke, and H. Van Oyen, “Towards a European health research and innovation cloud (HRIC),” Genome Med vol. 12, pp. 18, 2020. https://doi.org/10.1186/s13073-020-0713-z.
26. O. D. Argyropoulou, **V. Pezoulas**, L. Chatzis, E. Critselis, S. Gandolfo, F. Ferro, L. Quartuccio, V. Donati, E. Treppo, C.R. Bassoli, A. Venetsanopoulou, E. Zampeli, M. Mavrommati, P. V. Voulgari, T.E. Exarchos, C. P. Mavragani, C. Baldini, F. N. Skopouli, M. Galli, D. I. Fotiadis, S. De Vita, H. M. Moutsopoulos, A. G. Tzioufas, A. V. Goules, “Cryoglobulinemic vasculitis in primary Sjögren's Syndrome: Clinical presentation, association with lymphoma and comparison with Hepatitis C-related disease,” Seminars in Arthritis and Rheumatism, vol. 50, no 5, pp. 846-853, 2020.
27. L. Chatzis, S. Gandolfo, F. Ferro, M. Binutti, V. Donati, S. Z. Callegher, **V. Pezoulas,** et al. "Comparison of Clinical Phenotype, Serological Characteristics and Histologic Features Between Males and Females Patients with Primary Sjogren's Syndrome (pSS),” In Arthritis & Rheumatology, vol. 71. 111 River St, Hoboken, 07030-5774, NJ USA: Wiley, 2019.

28. A. Goules, M. Voulgarelis, L. Chatzis, **V. Pezoulas**, F. Ferro, S. Gandolfo, V. Donati, et al. “Data Driven Prediction Lymphoma Model and 10-year Overall Survival Rates of a Large Harmonized Cohort of Patients with Primary Sjogren's Syndrome Associated Lymphomas,” In Arthritis and Rheumatology, vol. 71. 111 River St, Hoboken, 07030-5774, NJ USA: Wiley, 2019.
29. A. V. Goules, T. P. Exarchos, **V. C. Pezoulas**, K. D. Kourou, A. I. Venetsanopoulou, S. De Vita, D. I. Fotiadis, and A. G. Tzioufas, “Sjögren's syndrome towards precision medicine: the challenge of harmonisation and integration of cohorts,” Clinical and Experimental Rheumatology, vol. 118, no 3, pp. 175-184, 2019.