UNIVERSITY OF IOANNINA
SCHOOL OF HEALTH SCIENCES
FACULTY OF MEDICINE
SECTION OF SOCIAL MEDICINE & MENTAL HEALTH
DEPARTMENT OF HYGIENE & EPIDEMIOLOGY

Empirical Assessment of the Discrimination Ability of Cardiovascular Disease Risk Prediction Models for Mortality

Georgios C.M. Siontis, MD

PhD THESIS

IOANNINA 2015

UNIVERSITY OF IOANNINA
SCHOOL OF HEALTH SCIENCES
FACULTY OF MEDICINE
SECTION OF SOCIAL MEDICINE & MENTAL HEALTH
DEPARTMENT OF HYGIENE & EPIDEMIOLOGY

Empirical Assessment of the Discrimination Ability of Cardiovascular Disease Risk Prediction Models for Mortality

Georgios C.M. Siontis, MD

PhD THESIS

IOANNINA 2015

Η έγκριση της διδακτορικής διατριβής από το Τμήμα Ιατρικής του Πανεπιστημίου Ιωαννίνων δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα Ν. 5343/32, άρθρο 202, παράγραφος 2 (νομική κατοχύρωση του Ιατρικού Τμήματος).

Ημερομηνία αίτησης του κ. Σιόντη Γεωργίου: 29-3-2010

Ημερομηνία ορισμού Τριμελούς Συμβουλευτικής Επιτροπής: 683^{α}/11-5-2010

Μέλη Τριμελούς Συμβουλευτικής Επιτροπής:

Επιβλέπων

Ιωαννίδης Ιωάννης Καθηγητής Υγιεινής

Μέλη

Γουδέβενος Ιωάννης Καθηγητής Παθολογίας-Καρδιολογίας

Ντζάνη Ευαγγελία Επίκουρη Καθηγήτρια Υγιεινής με έμφαση στην Επιδημιολογία

Ανασύσταση Τριμελούς Συμβουλευτικής Επιτροπής: Αριθμ Συνεδρ. 693^{α}/21-9-2010

Επιβλέπων

Ιωαννίδης Ιωάννης Καθηγητής Υγιεινής

Μέλη

Γουδέβενος Ιωάννης Καθηγητής Παθολογίας-Καρδιολογίας

Τζουλάκη Ιωάννα Λέκτορας Υγιεινής με έμφαση στην Επιδημιολογία

Ανασύσταση Τριμελούς Συμβουλευτικής Επιτροπής: Αριθμ Συνεδρ. 776^{α}/3-4-2015

Επιβλέπων

Ιωαννίδης Ιωάννης Ομότιμος Καθηγητής Επιδημιολογίας

Μέλη

Γουδέβενος Ιωάννης Καθηγητής Παθολογίας-Καρδιολογίας

Τατσιώνη Αθηνά Επίκουρη Καθηγήτρια Γενικής Ιατρικής

Ημερομηνία ορισμού θέματος: 28-5-2010

«Εμπειρική αποτίμηση μελετών διακριτικής ικανότητας για επιβίωση σε καρδιαγγειακά νοσήματα»

ΟΡΙΣΜΟΣ ΕΠΤΑΜΕΛΟΥΣ ΕΞΕΤΑΣΤΙΚΗΣ ΕΠΙΤΡΟΠΗΣ : 777^{α}/19-5-2015

1. Ιωαννίδης Ιωάννης Ομότιμος Καθηγητής Επιδημιολογίας του Τμήματος Ιατρικής του Παν/μίου Ιωαννίνων
2. Windecker Stephan, Professor of Cardiology, Bern University Hospital, Bern, Switzerland
3. Γουδέβενος Ιωάννης Καθηγητής Παθολογίας-Καρδιολογίας του Τμήματος Ιατρικής του Παν/μίου Ιωαννίνων
4. Σαλαντή Γεωργία Επίκουρη Καθηγήτρια Επιδημιολογίας του Τμήματος Ιατρικής του Παν/μίου Ιωαννίνων

5. Τατσιώνη Αθηνά Επίκουρη Καθηγήτρια Γενικής Ιατρικής του Τμήματος Ιατρικής του Παν/μίου Ιωαννίνων
6. Τσιλίδης Κωνσταντίνος, Επίκουρος Καθηγητής Υγιεινής με έμφαση στην Επιδημιολογία του Τμήματος Ιατρικής του Παν/μίου Ιωαννίνων
7. Μαυρίδης Δημήτριος, Λέκτορας του Παιδαγωγικού Τμήματος Δημοτικής Εκπαίδευσης του Παν/μίου Ιωαννίνων

Έγκριση Διδακτορικής Διατριβής με βαθμό **«ΑΡΙΣΤΑ»** στις **3-9-2015**

ΠΡΟΕΔΡΟΣ ΤΟΥ ΤΜΗΜΑΤΟΣ ΙΑΤΡΙΚΗΣ
Πασχόπουλος Μηνάς
Καθηγητής Μαιευτικής-Γυναικολογίας

Η Γραμματέας του Τμήματος
ΚΑΠΙΤΟΠΟΥΛΟΥ ΜΑΡΙΑ

ΙΑΤΡΙΚΗ ΣΧΟΛΗ Π. ΠΑΝΕΠΙΣΤΗΜΙΟΥ ΙΩΑΝΝΙΝΩΝ
1977

“The journey is the reward”
- S. Jobs

To my parents
To my brother
To my wife

Table of contents

Section 1.

Background, aims and outline

1.1 Cardiovascular diseases burden

Cause-specific mortality is arguably one of the most fundamental metrics of population health. The rates and numbers of people who die, where, at what age, and from what, is a crucial input into policy debates, planning interventions, and prioritising research for new health technologies. Cardiovascular diseases (CVD) is the leading global cause of death, accounting for 17.3 million deaths per year, a number that is expected to grow to >23.6 million by 2030.[1] The World Health Organisation (WHO) estimates about 20 million cardiovascular associated deaths in 2015, accounting for 30% of all deaths worldwide. As the burden of this disease affects practically all hospital systems around the globe, in terms of costs and availability of resources, there has also been permanent interest among healthcare providers to examine and improve prevention and prediction. Reducing the burden of CVD has been a public health priority for more than 50 years and will continue to be in foreseeable future.[2] Targeting interventions to reduce the risk of cardiovascular disease by identifying those who are at high risk for development of CVD are now key components in national policies.[3]

1.2 What is prognosis and why is important?

Hippocrates included prognosis as a principal concept of medicine.[4] In medicine, numerous decisions are made by care providers, often in shared decision making, on the basis of an estimated probability that a specific disease or condition is present (diagnostic setting) or a specific event will occur in the future (prognostic setting) in an individual. In the prognostic context, predictions can be used for planning lifestyle or therapeutic decisions on the basis of the risk for developing a particular outcome or state of health within a specific period.[5] Such estimates of risk can also be used to risk-stratify participants in therapeutic intervention trials.[6-9] Prognosis simply means foreseeing, predicting, or estimating the probability or risk of future conditions; familiar examples are weather and economic forecasts. In medicine, prognosis commonly relates to the probability or risk of an individual developing a particular state of health (an outcome) over a specific period of time (**Figure 1.2.1**), based on patient's clinical (baseline characetristics and symptoms) and non-clinical profile (i.e. information derived from imaging tests).

Outcomes are often specific events, such as death, specific diseases (i.e. cardiovascular disease) or complications of medical interventions, but they may also be quantities, such as disease progression, or quality of life. In medical textbooks, however, prognosis commonly refers to the expected course of an illness. This terminology is too general and has limited utility in practice. Doctors do not predict the course of an illness but the course of an illness in a particular individual. Prognosis may be shaped by a patient's age, sex, history, symptoms, signs, and other test results. Moreover, prognostication in medicine is not limited to those who are ill. Healthcare professionals, especially primary care doctors, regularly predict the future in healthy individuals—for example, using the Apgar score to determine the prognosis of newborns, cardiovascular risk profiles to predict heart disease in the general population, and prenatal testing to assess the risk that a pregnant woman will give birth to a baby with Down's syndrome.

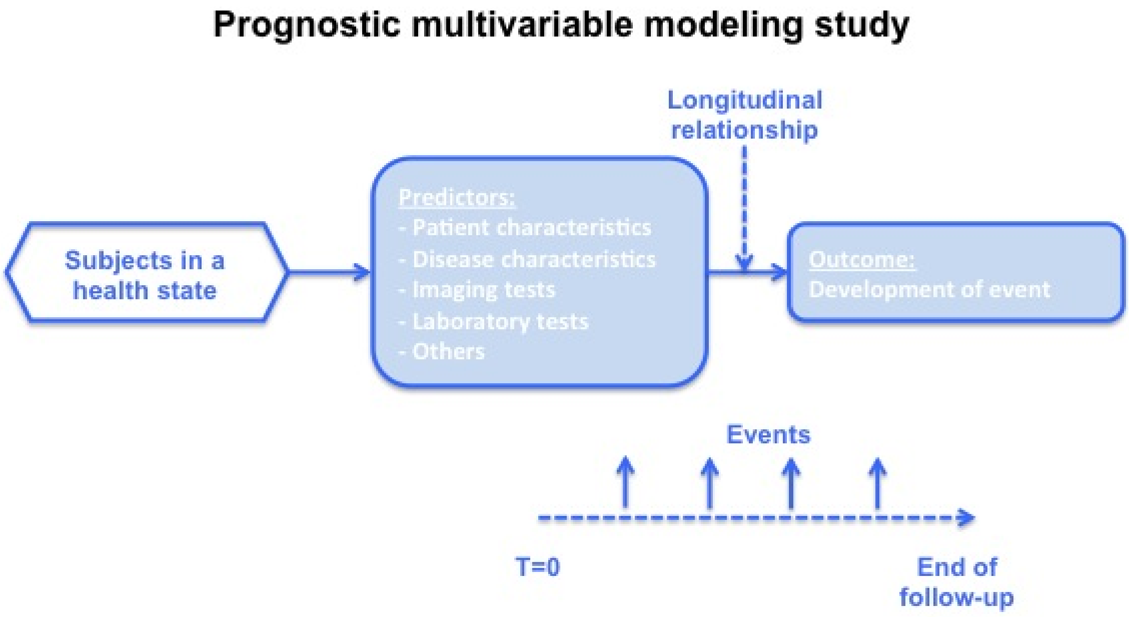
Prognostic multivariable modeling study
Longitudinal relationship
Subjects in a health state
Predictors:
- Patient characteristics
- Disease characteristics
- Imaging tests
- Laboratory tests
- Others
Outcome:
Development of event
Events
T=0
End of follow-up

Figure 1.2.1: Schematic representation of prediction studies. The prediction is about whether an individual will experience a specific outcome within a certain period of time (longitudinal relationship).[10]

Medical prognostication and prognostic models are widely used in various settings and for various reasons. The main reasons are to inform individuals about the future course of their illness (or their risk of developing illness) and to guide doctors and patients in joint decisions on further appropriate management of the patient and treatment. Furthermore, prognostic models are important at different stages in pathways leading to improvements in health (**Figure 1.2.2**). The use of prognostic models ties in with the strong movement towards stratified medicine, where decisions regarding treatment choices are informed by an individual's profile of prognostic factors. Prognostic models aim to assist (but not replace) clinicians with their prediction of a patient's future outcome and to enhance informed decision making with the patient. For example, modifications of the Framingham cardiovascular risk score[11] are used in primary care setting to identify those individuals who are at high risk of developing cardiovascular disease and therefore determine the indication for cholesterol lowering and antihypertensive drugs. Examples from secondary care include use of the HAS-BLED score to assess the individual 1-year bleeding risk of real-world patients with atrial fibriblation[12], Nottingham prognostic index to estimate the long term risk of cancer recurrence or death in breast cancer patients[13], the acute physiology and chronic health evaluation (APACHE) score and simplified acute physiology score (SAPS) to predict hospital mortality in critically ill patients[14-16].

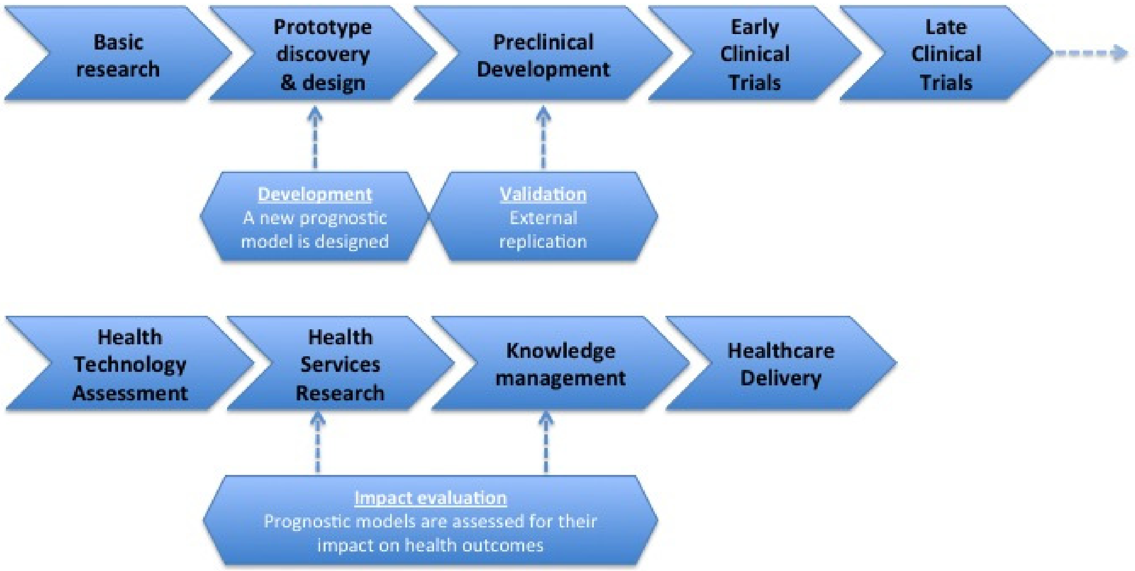

Figure 1.2.2: Position of prognostic models along the translational pathways.

Another reason for prognostication and use of prognostic models is to select relevant patients for therapeutic research. For example, researchers used a previously validated prognostic model to select women with an increased risk of developing cancer for a randomised trial of tamoxifen to prevent breast cancer.[17] Another randomised trial on the efficacy of radiotherapy after breast conserving resection used a prognostic model to select patients with a low risk of cancer recurrence.[18] Prognostic models are also used to compare differences in performance between hospitals. For example, the clinical risk index for babies (CRIB) was originally developed to compare performance and mortality among neonatal intensive care units.[19]

Many national and international guidelines endorse the use of such risk prediction models to guide individualised decision-making for lifestyle recommendations and medical treatments as part of primary or secondary prevention of specific medical conditions. Well-known examples are the Pooled Cohort Equations (PCE) of the American Heart Association, the SCORE model in various European countries, and QRisk in the UK. Risk scores to predict cardiovascular disease risk are abundant, as shown by a recent comprehensive review of the literature, which identified 796 cardiovascular disease risk models published between 1990 - 2012.[20] Moreover, the findings of the comprehensive review[21] and others showed that most cardiovascular disease prediction models are never validated for predictive accuracy in individuals outside the population they were developed for.[22] Moreover, most cardiovascular disease prediction models are developed from single-country cohort or registry studies, which are, generally, from North American or European countries. However, cardiovascular disease burden is also rapidly increasing in low-income and middle-income countries, including those in Asia, Africa, and Latin America.[23] Therefore, now is the time for either cardiovascular disease prediction models to be developed from and validated in datasets from these countries, or for existing cardiovascular disease prediction models to be tailored or recalibrated to these populations.

Previously developed and newly introduced risk prediction models, such as the Framingham risk score[24], the Reynolds risk score[25,26], and QRISK[27-31] have been used to identify people who are at high risk (≥20%)[3] of developing (10 year) cardiovascular disease and could benefit from intervention targeting in primary prevention. In the United States, about 10.2 million people have chest pain complaints each year[1], and more than 1.1 million diagnostic procedures of catheter based coronary angiography are performed on inpatients each year.[32] In a recent report based on the national cardiovascular data registry of the American College of Cardiology[33], only 41% of patients undergoing elective procedures of catheter based coronary angiographies are diagnosed with obstructive coronary artery disease. These findings highlight the need for better risk stratification and further diagnostic investigation in such patients presenting with chest pain; whereas similar examples are available for various settings.[34-36]

Although there are clear similarities in the design and analysis of prognostic and aetiological studies, predicting outcomes is not synonymous with explaining their cause.[37] In aetiological research, the mission is to explain whether an outcome can reliably be attributed to a particular risk factor, with adjustment for other causal factors (confounders) using a multivariable approach. In prognostic research the mission is to use single or multiple variables to predict, as accurately as possible, the risk of future outcomes. Although a prognostic model may be used to provide insights into causality or pathophysiology of the studied outcome, that is neither an aim nor a requirement. All variables potentially associated with the outcome, not necessarily causally, can be considered in a prognostic study. Every causal factor is a predictor—albeit sometimes a weak one—but not every predictor is a cause. Nice examples of predictive but non-causal factors used in everyday practice are skin colour in the Apgar score and tumour markers as predictors of cancer progression or recurrence. Both are surrogates for obvious causal factors that are more difficult to measure.

Furthermore, to guide prognostication in individuals, analysis and reporting of prognostic studies should focus on absolute risk estimates of outcomes given combinations of predictor values. Relative risk estimates (i.e. odds ratio, risk ratio, or hazard ratio) have no direct meaning or relevance to prognostication in practice. In prediction research, relative risks are used only to obtain an absolute probability of the outcome for an individual. In contrast, aetiological and therapeutic studies commonly focus on relative risks—for example, the risk of an outcome in presence of a causal factor relative to the risk in its absence. Also, other metrics, such as the calibration and discrimination of a multivariable model are highly relevant to prognostic research but meaningless in aetiological research.

Prognostic models are important at different stages in pathways leading to improvements in health. The use of prognostic models ties in with the strong movement towards stratified medicine, where decisions regarding treatment choices are informed by an individual's profile of prognostic factors. Prognostic models aim to assist (not replace) clinicians with their prediction of a patient's future outcome and to enhance informed decision making with the patient. The results from randomised therapeutic trials can be used to estimate how a specific treatment would modify a patient's estimated prognosis. Under the common assumption that a particular treatment has a constant relative benefit across all risk groups, the absolute treatment benefit depends on a person's predicted risk of the outcome without treatment.[7] Expensive therapies or those with harmful potential side effects may thus be reserved for those at higher risk, as estimated by a prognostic model.

Some prognostic models are used in clinical practice without being identified as such, such as the Apgar score for assessing the wellbeing of newborn babies.[38] Other examples of well-used prognostic models include the Nottingham Prognostic Index[39], the Oerebro Musculoskeletal Pain Screening Questionnaire to help clinicians identify patients with low back pain at risk of poor recovery[40], and the Manchester Triage System to assign priority based on clinical need among patients visiting an emergency department.[41] A prognostic model can thus be seen as an intervention that requires preclinical development, validation, and subsequent evaluation of its impact on health outcomes and cost effectiveness of care. Prognostic models are also important to help improve the design and analysis of randomised therapeutic trials[42,43], and to adjust for case mix variation in health services research[38], such as in understanding variations in patients' outcome across hospitals.[44]

1.3 Multivariable prognostic research

Given the variability among patients and in the aetiology, presentation, and treatment of diseases and other health states, a single predictor or variable rarely gives an adequate individual estimate of prognosis. To improve the targeting of interventions to patients based on their predicted individual risk of subsequent outcomes, decision makers can use multiple prognostic factors combined within a prognostic model. Other names for a prognostic model include prognostic (or prediction) index or rule, risk (or clinical) prediction model, and predictive model. For an individual with a given state of health (startpoint), a prognostic model converts the combination of predictor values to an estimate of the risk of experiencing a specific outcome within a specific period of time. Ideally this produces an estimate of the absolute risk (absolute probability) of experiencing the endpoint, but it may instead provide a relative risk or risk score.[5,45] Using prognostic models to make predictions for individual patients is more accurate and so is often preferred to risk grouping, although risk groups may inform treatment choices and enable stratification for risk severity in clinical trials. Nowadays, the majority of the prognostic models are easily accessible as web tools, providing additional details in individual level. Doctors—implicitly or explicitly—use a combination of multiple predictors to estimate an absolute risk or probability that an outcome will occur in an individual - patient's level prognosis. Prognostic studies therefore need to use a multivariable approach in design and analysis to determine the important predictors of the studied outcomes and to provide outcome probabilities for different combinations of predictors, or to provide tools to estimate such probabilities.[46-55] A multivariable prediction model is a mathematical equation that relates multiple predictors for a particular individual to the probability of or risk for the presence (diagnosis) or future occurrence (prognosis) of a particular outcome.[54,56,57]

A multivariable approach also enables researchers to investigate whether specific prognostic factors or markers that are, more invasive or costly to measure, have worthwhile added predictive value beyond cheap or simply obtained predictors—for example, from patient history or physical examination.

Nonetheless, many prognostic studies still consider a single rather than multiple predictors.[58] Predictors are also referred to as covariates, risk indicators, prognostic factors, determinants, test results, or—more statistically—independent variables. They may range from demographic characteristics (for example, age and sex), medical history–taking, and physical examination results to results from imaging, electrophysiology, blood and urine measurements, pathologic examinations, and disease stages or characteristics, or results from genomics, proteomics, transcriptomics, pharmacogenomics, metabolomics, and other new biological measurement platforms that continuously emerge. In 2013, a wide scale systematic search was performed in Medline and Embase to identify studies that described the development, validation or incremental value of a multivariable prognostic model predicting CVD in the general population.[59] 314 studies were included, describing the development of 373 prognostic models, 519 external validations and 278 incremental value assessments. Most prevalent predictors were age and smoking (n=323 and n=332 respectively, **Figure 1.3.1**), with frequently (n=234) separate models for males and females; whereas substantial heterogeneity in predictor and outcome definitions was seen between models.

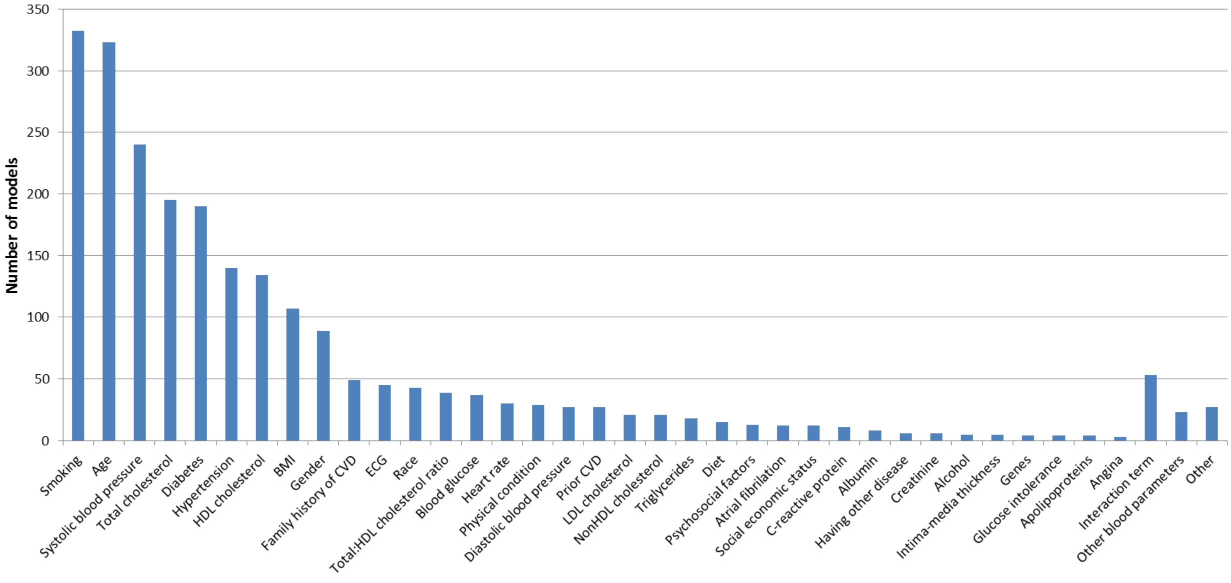

Figure 1.3.1: Single predictors included in models for CVD prediction.

In a prognostic model, multiple predictors are combined to estimate the probability of a particular event (i.e. mortality, disease recurrence, complication, or therapy response) occurring in a certain period in the future. This period may range from hours (i.e. predicting postoperative complications[60]) to weeks or months (i.e. predicting 30-day mortality after cardiac surgery[61]) or years (i.e. predicting the 5-year risk for developing type 2 diabetes[62]). Prognostic models are developed and are to be used in individuals at risk for developing that outcome. They may be models for either ill or healthy individuals. For example, prognostic models include models to predict recurrence, complications, or death in a certain period after being diagnosed with a particular disease. But they may also include models for predicting the occurrence of an outcome in a certain period in individuals without a specific disease: for example, models to predict the risk for developing type 2 diabetes[63] or cardiovascular events in middle-aged nondiseased individuals[64], or the risk for preeclampsia in pregnant women[65]. We thus use prognostic in the broad sense, referring to the prediction of an outcome in the future in individuals at risk for that outcome, rather than the narrower definition of predicting the course of patients who have a particular disease with or without treatment.[5]

The multivariable character of prognostic research makes it difficult to estimate the required sample size. There are no straightforward methods for this. When the number of predictors is much larger than the number of outcome events, there is a risk of overestimating the predictive performance of the model. Ideally, prognostic studies that aim to build a multivariable prediction model, require at least several hundred outcome events. Various studies have suggested that for each candidate predictor studied at least 10 events are required,[52,54,66,67] although a more recent study showed that this number could be lower in certain circumstances.[68]

1.4 Use of prognostic models

Medical prognostication and prognostic models are used in various settings and for various reasons. The main reasons are to inform individuals about the future course of their illness (or their risk of developing illness/specific outcome) and to guide doctors and patients in joint decisions on further treatment, if any. For example, several modifications of the Framingham cardiovascular risk score[11] are widely used in primary care to determine the indication for cholesterol lowering and antihypertensive drugs. Examples from secondary care include use of the CHA2DS2-VASc score for esrtimation the risk of stroke in patients with atrial fibrillation[69], and the GRACE score for estimation the risk of 6-month postdischarge death in patients hospitalised due to acute coronary syndromes[70].

Another reason for prognostication and use of prognostic models is to select relevant patients for therapeutic research. For example, researchers used a previously validated prognostic model to select women with an increased risk of developing cancer for a randomised trial of tamoxifen to prevent breast cancer.[17] Another randomised trial on the efficacy of radiotherapy after breast conserving resection used a prognostic model to select patients with a low risk of cancer recurrence.[18]

Prognostic models are also used to compare differences in performance between hospitals. For example, the clinical risk index for babies (CRIB) was originally developed to compare performance and mortality among neonatal intensive care units.[19] More recently Jarman et al developed a model to predict the hospital standardised mortality ratio to explain differences between English hospitals.[71] Prognostic models are also important to help improve the design and analysis of randomised therapeutic trials[42,43], and to adjust for case mix variation in health services research[38], such as in understanding variations in patients' outcome across hospitals[44].

The results from randomised therapeutic trials can be used to estimate how a specific treatment would modify a patient's estimated prognosis. Under the common assumption that a particular treatment has a constant relative benefit across all risk groups, the absolute treatment benefit depends on a person's predicted risk of the outcome without treatment.[7] Expensive therapies or those with harmful potential side effects may thus be reserved for those at higher risk, as estimated by a prognostic model.

1.5 Developing prediction models

Development of a multivariable prognostic model, includes identification of the important predictors, assigning relative weights to each predictor, and estimating the model's predictive performance through calibration and discrimination and its potential for optimism using internal validation techniques, and, if necessary, adjusting the model for overfitting. The goal is to construct an accurate and discriminating prediction model from multiple variables. Models may be a complicated function of the predictors, as in weather forecasting, but in clinical applications considerations of practicality and face validity usually suggest a simple, interpretable model.

Examples of logistic regression models are shown. A logistic regression model of the final ADVANCE model for prediction of CVD is illustrated below.[74] Positive regression coefficients indicate an increased risk of CVD.

$$\begin{aligned}\log it(P(CVD)) &= 0.062Age + 0.083Diabetes\ \ duration + 0.007Pulse\ \ pressure + 0.242Hypertension \\ &+ 0.193\log(urinary\ \ albu\min/creatinin) + 0.099HbA1c + 0.126Non - HDL + 0.383Retinopathy \\ &+ 0.601Atrial\ \ \ fibrillati\,on - 0.474Gender\end{aligned}$$

the numbers are the estimated regression coefficients for the predictors, which indicate their mutually adjusted relative contribution to the outcome risk.

Surprisingly, there is no widely agreed approach to build a multivariable prognostic model from a set of candidate predictors. The best design to answer prognostic questions is a cohort study. A prospective study is preferable as it enables optimal measurement of predictors and outcomes. Studies using cohorts already assembled for other reasons allow longer follow-up times but usually at the expense of poorer data. Unfortunately, the prognostic literature is dominated by retrospective studies. Case-control studies are sometimes used for prognostic analysis, but they do not automatically allow estimation of absolute risks because cases and controls are often sampled from a source population of unknown size. Since investigators are free to choose the ratio of cases and controls, the absolute outcome risks can be manipulated.[75] An exception is a case-control study nested in a cohort of known size.[76]

Data from randomised trials of treatment can also be used to study prognosis. When the treatment is ineffective (relative risk=1.0), the intervention and comparison group can simply be combined to study baseline prognosis. If the treatment is effective the groups can be combined, but the treatment variable should then be included as a separate predictor in the multivariable model. Here treatments are studied on their independent predictive effect and not on their therapeutic or preventive effects. However, prognostic models obtained from randomised trial data may have restricted generalisability because of strict eligibility criteria for the trial, low recruitment levels, or large numbers refusing consent.

Before starting to develop a multivariable prediction model, numerous decisions must be made that affect the model and therefore the conclusions of the research. These are provided below:

- ***Selecting clinically relevant candidate predictors for possible inclusion in the model.***

Studies often measure more predictors than can sensibly be used in a model, and pruning is required. Predictors already reported as prognostic would normally be candidates. Candidate predictors can be obtained from patient demographics, clinical history, physical examination, disease characteristics, test results, and previous treatment. Prognostic studies may focus on a cohort of patients who have not (yet) received prognosis modifying treatments—that is, to study the natural course or baseline prognosis of patients with that condition. They can also examine predictors of prognosis in patients who have received treatments. Studied predictors should be clearly defined, standardised, and reproducible to enhance generalisability and application of study results to practice.[77] Predictors requiring subjective interpretation, such as imaging test results, are of particular concern in this context because there is a risk of studying the predictive ability of the observer rather than that of the predictors. Also, predictors should be measured using methods applicable—or potentially applicable—to daily practice. Specialised measurement techniques may yield optimistic predictions.

Predictors that are highly correlated with others contribute little independent information and may be excluded beforehand.[78] However, predictors that are not significant in univariable analysis should not be excluded as candidates.[54,79,80]

- ***Evaluating the quality of the data and judging what to do with missing values.***

There are no secure rules for evaluating the quality of data. Judgment is required. In principle, data used for developing a prognostic model should be fit for purpose. Measurements of candidate predictors and outcomes should be comparable across clinicians or study centres. Predictors known to have considerable measurement error may be unsuitable because this dilutes their prognostic information.

Modern statistical techniques (such as multiple imputation) can handle data sets with missing values.[81,82] However, all approaches make critical but essentially untestable assumptions about how the data went missing. The likely influence on the results increases with the amount of data that are missing. Missing data are seldom completely random. They are usually related, directly or indirectly, to other subject or disease characteristics, including the outcome under study. Thus exclusion of all individuals with a missing value leads not only to loss of statistical power but often to incorrect estimates of the predictive power of the model and specific predictors.[82] A complete case analysis may be sensible when few observations (say <5%) are missing.[78] If a candidate predictor has a lot of missing data it may be excluded because the problem is likely to recur.

- ***Data handling decisions.***

For the appropriate building of a new prediction model, new variables may need to be created (for example, diastolic and systolic blood pressure may be combined to give mean arterial pressure). For ordered categorical variables, such as stage of disease, collapsing of categories or a judicious choice of coding may be required. Including continuous predictors as they are provided is recommended.[83] Keeping variables continuous is preferable since much more predictive information is retained.[84,85]

- ***Choosing a strategy for selecting the important variables in the final model.***

Suprisingly, no consensus exists on the best method for selecting the appropriate variables in the final prediction model. Two main strategies have been proposed, each with variants. In the full model approach all the candidate variables are included in the model. This model is claimed to avoid overfitting and selection bias and provide correct standard errors and P values.[78] However, as many important preliminary choices must be made and it is often impractical to include all candidates, the full model is not always easy to define.

The backward elimination approach starts with all the candidate variables. A nominal significance level, often 5%, is chosen in advance. A sequence of hypothesis tests is applied to determine whether a given variable should be removed from the model. Backward elimination is preferable to forward selection (whereby the model is built up from the best candidate predictor). The choice of significance level has a major effect on the number of variables selected. A 1% level almost always results in a model with fewer variables than a 5% level. Significance levels of 10% or 15% can result in inclusion of some unimportant variables, as can the full model approach. A variant is the Akaike information criterion,[86] a measure of model fit that includes a penalty against large models and hence attempts to reduce overfitting. For a single predictor, the criterion equates to selection at 15.7% significance.[86]

Selection of predictors by significance testing, particularly at conventional significance levels, is known to produce selection bias and optimism as a result of overfitting, meaning that the model is (too) closely adapted to the data.[78,79,86] Selection bias means that a regression coefficient is overestimated, because the corresponding predictor is more likely to be significant if its estimated effect is larger (perhaps by chance) rather than smaller. Overfitting leads to worse prediction in independent data; it is more likely to occur in small data sets or with weakly predictive variables. Note, however, that selected predictor variables with very small P values (say, <0.001) are much less prone to selection bias and overfitting than weak predictors with P values near the nominal significance level. Commonly, prognostic data sets include a few strong predictors and several weaker ones.

- ***Deciding how to model continuous variables.***

Handling continuous predictors in multivariable prognostic modelling is important. It is unwise to assume linearity as it can lead to misinterpretation of the influence of a predictor and to inaccurate predictions in new patients.[84] Simple predictor transformations intended to detect and model non-linearity can be systematically identified using, for example, fractional polynomials, a generalisation of conventional polynomials (linear, quadratic, etc).[87,88] Power transformations of a predictor beyond squares and cubes, including reciprocals, logarithms, and square roots are allowed. These transformations contain a single term, but to enhance flexibility can be extended to two term models (eg, terms in log x and x^2). Fractional polynomial functions can successfully model non-linear relationships found in prognostic studies. The multivariable fractional polynomial procedure is an extension to multivariable models including at least one continuous predictor,[87] and combines backward elimination of weaker predictors with transformation of continuous predictors. Restricted cubic splines are an alternative approach to modelling continuous predictors. Their main advantage is their flexibility for representing a wide range of perhaps complex curve shapes. Drawbacks are the frequent occurrence of wiggles in fitted curves that may be unreal and open to misinterpretation[89-91] and the absence of a simple description of the fitted curve.

- ***Selecting measure(s) of model performance.***

The performance of a logistic regression model may be assessed in terms of calibration and discrimination. Calibration can be investigated by plotting the observed proportions of events against the predicted risks for groups defined by ranges of individual predicted risks; a common approach is to use 10 risk groups of equal size. Ideally, if the observed proportions of events and predicted probabilities agree over the whole range of probabilities, the plot shows a 45° line (that is, the slope is 1). This plot can be accompanied by the Hosmer-Lemeshow test,[91] although the test has limited power to assess poor calibration. The overall observed and predicted event probabilities are by definition equal for the sample used to develop the model. This is not guaranteed when the model's performance is evaluated on a different sample in a validation study.

Various statistics can summarise discrimination between individuals with and without the outcome event. The area under the receiver operating curve,[54,92] or the equivalent c (concordance) index, is the chance that given two patients, one who will develop an event and the other who will not, the model will assign a higher probability of an event to the former. The c index for a prognostic model is typically between about 0.6 and 0.85 (higher values are seen primarly in diagnostic settings).[93] Another measure is R^2, which for logistic regression assesses the explained variation in risk and is the square of the correlation between the observed outcome (0 or 1) and the predicted risk.[94]

As discussed above, the prognostic value of treatments can also be studied, especially when randomised trials are used. However, caution is needed in including treatments as prognostic factors when data are observational. Indications for treatment and treatment administration are often not standardised in observational studies and confounding by indication could lead to bias and large variation in the (type of) administered treatments.[95] Moreover, in many circumstances the predictive effect of treatments is small compared with that of other important prognostic variables such as age, sex, and disease stage. Finally, of course, studies should include only predictors that will be available at the time when the model is intended to be used.[96] If the aim is to predict a patient's prognosis at the time of diagnosis, for example, predictors that will not be known until actual treatment has started are of little value.

Preferably, prognostic studies should focus on outcomes that are relevant to patients, such as occurrence or remission of disease, death, complications, tumour growth, pain, treatment response, or quality of life. Surrogate or intermediate outcomes, such as hospital stay or physiological measurements, are unhelpful unless they have a clear causal relation to relevant patient outcomes, such as CD4 counts instead of development of AIDS or death in HIV studies. The period over which the outcome is studied and the methods of measurement should be clearly defined. Finally, outcomes should be measured without knowledge of the predictors under study to prevent bias, particularly if measurement requires observer interpretation. Blinding is not necessary when the outcome is all cause mortality. But if the outcome is cause specific mortality, knowledge of the predictors might influence assessment of outcomes (and vice versa in retrospective studies where predictors are documented after the outcome was assessed).

Clinical prediction models are usually developed in a single large cohort using multivariate models to determine which variables are associated with disease or disease outcomes.[97] These variables are often weighted to produce a score that is predictive of the outcome. To maximize utility, prediction models should be developed using data from patients who are representative of those for whom the rule will eventually be used, include all of the variables that might be predictive, and target an appropriate clinical outcome.[97,98] Variables that might be included in a prediction model include demographics, symptoms, physical findings, laboratory test results, imaging, pathology findings, and other variables. Measurement of these variables must be clear and reproducible,[52,97] or others will not be able to reliably calculate the risk score. For example, a risk prediction score that relies on subjectively measured muscle strength or degree of confusion is unlikely to be reproducible.

Although there is no clear consensus on the best method of model building, the importance of having an adequate sample size and high quality data is widely agreed. Model building from small data sets requires particular care. A model's performance is likely to be overestimated when it is developed and assessed on the same dataset. The problem is greatest with small sample sizes, many candidate predictors, and weakly influential predictors. The amount of optimism in the model can be assessed and corrected by internal validation techniques.

Developing a model is a complex process, so readers of a report of a new prognostic model need to know sufficient details of the data handling and modelling methods.[99] All candidate predictors and those included in the final model and their explicit coding should be carefully reported. All regression coefficients should be reported (including the intercept) to allow readers to calculate risk predictions for their own patients.

The predictive performance or accuracy of a model may be adversely affected by poor methodological choices or weaknesses in the data. But even with a high quality model there may simply be too much unexplained variation to generate accurate predictions. A critical requirement of a multivariable model is thus transportability, or external validity—that is, confirmation that the model performs as expected in new datasets of similar patients.[50]

1.6 Validating prediction models

Before any prognostic model might be adopted in practice it is necessary to show that it provides predictions that are valid outside the specific context of the sample that was used for model development (external validation) and ideally has real clinical impact. To show that a prognostic model is valuable, it is not sufficient to show that it successfully predicts outcome in the initial development data. We need evidence that the model performs well for other groups of patients, out of the development set. Quantifying the predictive ability of a model on the same data from which the model was developed will tend to give an optimistic estimate of performance, owing to overfitting (too few outcome events relative to the number of candidate predictors) and the use of predictor selection strategies.[79,100,101] Studies developing new prediction models should therefore always include some form of internal validation to quantify any optimism in the predictive performance (for example, calibration and discrimination) of the developed model and adjust the model for overfitting. Internal validation techniques use only the original study sample and include such methods as bootstrapping or crossvalidation. Internal validation is a necessary part of model development.

Various statistical or clinical factors may lead a prognostic model to perform poorly when applied to other patients.[50,102] Prediction models are generally accurate in the cohort in which the risk model was developed, but they are often not accurate in other populations. There are a number of reasons for this failure. Prediction models are generally developed as a fitted statistical model, assuring that the rule will perform optimally in the cohort in which it was developed. However, accuracy almost always decreases when the rule is applied in other clinical settings, even if the patients are similar. The model's predictions may not be reproducible because of deficiencies in the design or modelling methods used in the study to derive the model, if the model was overfitted, or if an important predictor is absent from the model (which may be hard to know).[56] Poor performance in new patients can also arise from differences between the setting of patients in the new and derivation samples, including differences in healthcare systems, methods of measurement, and patient characteristics. Prediction models are often applied in clinical settings that differ significantly from the cohort in which the model was initially developed. For example, differences in age, sex, prevalence of disease, or severity of disease may make the prediction model much less predictive in a different group of patients.[98] Thus, prediction models intended for broad clinical use should be shown to be accurate in a wide variety of clinical settings.

When a validation study shows disappointing results, researchers often reject the original prediction model and develop a new one from their own data. [48,103] However, the redeveloped model also often has several limitations, and multiple models for the same outcome create an impracticable situation where the user has to decide which model to use. Clearly, many more models are developed than are implemented or used in clinical practice. Moreover, if a new clinical prediction model is developed from every new population sample, previous predictive information already captured in previous studies and models is lost.[48,103] This goes against the intention that scientific inferences should be based on evidence from as many sources and individuals as possible; a principle that is well recognised and used in intervention studies (eg, cumulative meta-analyses of randomised trials). An alternative solution to redevelopment, is to adjust or update existing prediction models with the external validation set data at hand.[48,103]

The main ways to assess or validate the performance of a prognostic model on a new dataset are to compare observed and predicted event rates for groups of patients (calibration) and to quantify the model's ability to distinguish between patients who do or do not experience the event of interest (discrimination).[54,104] A model's performance can be assessed using new data from the same source as the derivation sample, but a true evaluation of generalisability (also called transportability) requires evaluation on data from elsewhere. After developing a prediction model, it is strongly recommended to evaluate the performance of the model in other participant data than was used for the model development. External validation[50,105] (**Figure 1.6.1**) requires that for each individual in the new participant data set, outcome predictions are made using the original model (that is, the published model or regression formula) and compared with the observed outcomes.

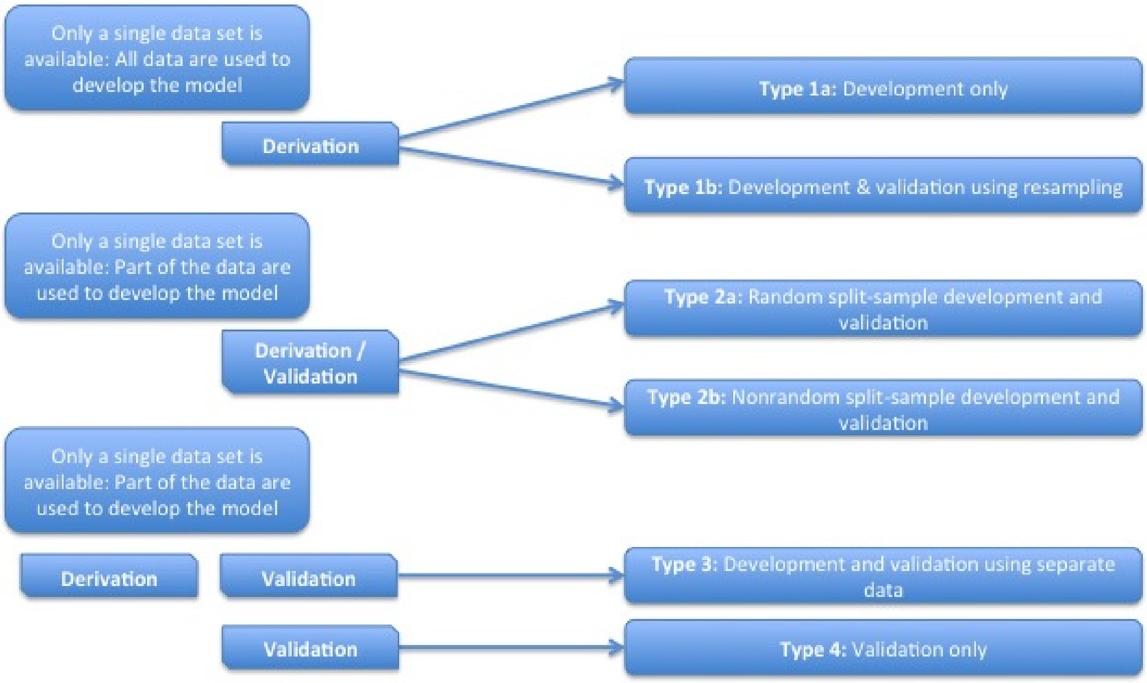
Only a single data set is available: All data are used to develop the model
Derivation
Type 1a: Development only
Type 1b: Development & validation using resampling
Only a single data set is available: Part of the data are used to develop the model
Derivation / Validation
Type 2a: Random split-sample development and validation
Type 2b: Nonrandom split-sample development and validation
Only a single data set is available: Part of the data are used to develop the model
Derivation
Validation
Type 3: Development and validation using separate data
Validation
Type 4: Validation only

Figure 1.6.1. Types of prediction model studies (TRIPOD statement)[10].

- ***Internal validation***

A common approach is to split the dataset randomly into two parts (often 2:1), develop the model using the first portion (often called the "training" set), and assess its predictive accuracy on the second portion ("testing test"). This approach will tend to give optimistic results because the two datasets are very similar. Non-random splitting (for example, by centre) may be preferable as it reduces the similarity of the two sets of patients.[50,56] If the available data are limited, the model can be developed on the whole dataset and techniques of data re-use, such as cross validation and bootstrapping, applied to assess performance.[56] Internal validation is helpful, but it cannot provide information about the model's performance elsewhere.

- ***Temporal and geographical validation***

An alternative is to evaluate the performance of a model on subsequent patients from the same centre(s).[102,106] Temporal validation is no different in principle from splitting a single dataset by time. There will clearly be many similarities between the two sets of patients and between the clinical and laboratory techniques used in evaluating them. However, temporal validation is a prospective evaluation of a model, independent of the original data and development process. Temporal validation can be considered external in time and thus intermediate between internal validation and external validation. A temporal validation may allow for more variation if not only owing to changes in healthcare over timed when it involves a prospective study specifically designed for the validation purpose, which starts after the model has been developed.

Temporal validation cannot examine the transportability or generalisability of the predictive performance of the model to other institutes or countries; which can be done through the geographical validation. Geographical validation studies commonly apply different in/exclusion criteria, and predictor and outcome definitions and measurements, as compared with the development study. As with temporal validation, geographical validation can again be done by non-random splitting of an existing study dataset by centre or country in for example, multicentre studies, or by validating a previously developed model in another centre or country that was not involved in the original development study. The latter geographical validation study involves a more stringent ‘proof of concept (prediction)’ owing to the probably greater differences in case mix, predictors and outcome measurements. Moreover, geographical validation may also be done retrospectively; that is, using existing datasets from other institutes or countries, or prospectively, by including new individuals in a specifically predesigned validation study.

- ***External validation***

Neither internal nor temporal/geographical validation examines the generalisability of the model, for which it is necessary to use new data collected from an appropriate (similar) patient population in a different centre. The data can be retrospective data and so external validation is possible for prediction models that need long follow-up to gather enough outcome events. Clearly, the second dataset must include data on all the variables in the model. Fundamental design issues for external validation, such as sample selection and sample size, have received limited attention.[107]

Model validation is not simply repeating the analytical steps applied in the development study in other individuals to see whether the same predictors and weights are found. Model validation is also not refitting the final developed model in the new individuals and checking whether the model performance that is, discrimination, calibration and classification, is different as was found in the development study. Model validation is taking the original model or simplified score, with its predictors and assigned weights (eg, regression coefficients), as estimated from the development study; measuring the predictor and outcome values in the new individuals; applying the original model to these data; and quantifying the model's predictive performance. External validation may use participant data collected by the same investigators, typically using the same predictor and outcome definitions and measurements, but sampled from a later period (temporal or narrow validation); by other investigators in another hospital or country (though disappointingly rare[108]), sometimes using different definitions and measurements (geographic or broad validation); in similar participants, but from an intentionally different setting (for example, a model developed in secondary care and assessed in similar participants, but selected from primary care); or even in other types of participants (for example, model developed in adults and assessed in children, or developed for predicting fatal events and assessed for predicting nonfatal events).[50,97,102,105,109,110] In case of poor performance (for example, systematic miscalibration), when evaluated in an external validation data set, the model can be updated or adjusted (for example, recalibrating or adding a new predictor) on the basis of the validation data set.[98,103,105]

Randomly splitting a single data set into model development and model validation data sets is frequently done to develop and validate a prediction model; this is often, yet erroneously, believed to be a form of external validation.

However, this approach is a weak and inefficient form of internal validation, because not all available data are used to develop the model.[100,111] If the available development data set is sufficiently large, splitting by time and developing a model using data from one period and evaluating its performance using the data from the other period (temporal validation) is a stronger approach. With a single data set, temporal splitting and model validation can be considered intermediate between internal and external validation.

External validation

- *Objective:* To apply a previously developed model to new individuals whose data were not used in the model development, and quantify the model's predictive performance.
- *Study individuals:* An adequate sample of "different but related individuals" compared to the development study sample. Related refers to "individuals at risk of developing the same event" for prognostic models.
- *Temporal validation:* new individuals are from the same institution as in the development sample, but in a different (preferably later) time period.
- *Geographical external validation*, new individuals are from different institutions or countries as in the development sample.
- *Domain validation*, new individuals are very different from the individuals from which the model was developed.
- *Procedure:* External validation of any type consists of taking the original model, with its predictors and assigned weights (eg, regression coefficients), as estimated from the development study; obtaining the measured predictor and outcome values in the new individuals; applying the original model to these data; and quantifying the model's predictive performance.
- *Performance measures:* Discrimination, calibration, re-classification.

-

- ***Domain validation***

A specific, and more rigid form of geographical validation or transportability test, is the validation of a developed model in very different individuals than those from whom it is developed, sometimes referred to as domain or setting validation.[93,98] Examples are validating a prediction developed in secondary care individuals suspected of having venous thromboembolism in a primary care setting,[112] validating a model developed in healthy individuals to predict the risk of cardiovascular events within 10 years (such as the Framingham risk score) in individuals diagnosed with diabetes mellitus type[52,113] or validating a model developed in adults to children.[114] Note that like geographical validation, domain validation may also be carried out retrospectively; that is, using existing datasets, or prospectively, by including new individuals in a specifically predesigned validation study.

Validation studies are necessary because performance in the original data may well be optimistic,[102] but temporal and (especially) external validation studies are scarce.[22,46,115,116] Proper validation requires that we use the fully specified existing prognostic model (that is, both the selected variables and their coefficients) to predict outcomes for the patients in the second dataset and then compare these predictions with the patients' actual outcomes. This analysis uses each individual's event probability calculated from their risk score from the first model.[56]

Both calibration and discrimination should be evaluated.[56] Calibration can be assessed by plotting the observed proportions of events against the predicted probabilities for groups defined by ranges of predicted risk, as discussed in the previous article.[56] This plot can be accompanied by the Hosmer-Lemeshow test, although the test has limited statistical power to assess poor calibration and is oversensitive for very large samples. For grouped data, as in the examples below, a χ^2 test can be used to compare observed and predicted numbers of events. It may also be helpful to compare observed and predicted outcomes in groups defined by key patient variables, such as diagnostic or demographic subgroups.

Discrimination may be summarised by the *c* index (area under the receiver-operator curve) or R^2.[56]

It may be helpful to prespecify acceptable performance of a model in terms of calibration and discrimination. If this performance is achieved, the model may be suitable for clinical use. It is, however, unclear how to determine what is acceptable, especially as prognostic assessments will still be necessary and even moderately performing models are likely to do better than clinicians' own assessments.[117,118] Simplicity of models and reliability of measurements are important criteria in developing clinically useful prognostic models.[5] Experience shows that more complex models tend to give overoptimistic predictions, especially when extensive variable selection has been performed, but there are notable exceptions.

As the aim of most prognostic studies is to create clinically valuable risk scores, the definition of risk groups should ideally be driven mainly by clinical rather than statistical criteria. If a clinician would leave untreated a patient with at least a 90% chance of surviving five years, would apply aggressive therapy if the prognosis was 30% survival or less, and would use standard therapy in intermediate cases, then three prognostic groups seem sensible. Validation of the model would investigate whether the observed proportions of events were similar in groups of patients from other settings and whether separation in outcome across those groups was maintained.

Few prognostic models are routinely used in clinical practice, probably because most have not been externally validated.[46] To be considered useful, a risk score should be clinically credible, accurate (well calibrated with good discriminative ability), have generality (be external validated), and, ideally, be shown to be clinically effective—that is, provide useful additional information to clinicians that improves therapeutic decision making and thus patient outcome.[46] It is crucial to quantify the performance of a prognostic model on a new series of patients, ideally in a different location, before applying the model in daily practice to guide patient care. Although still rare, temporal and external validation studies do seem to be becoming more common.

1.7 Updating prediction models

Researchers probably encounter a poorer performance of a prediction model when tested in new individuals compared with that found in the development study. The likelihood of finding a lower predictive accuracy will increase if a more stringent form of validation is used: this is more likely in a geographical or domain validation than in a temporal validation. When a lower predictive accuracy is found, "validation investigators" tend to simply reject that model and develop or fit a new one, sometimes by completely repeating the entire selection of predictors. This leads to a loss of previous scientific information captured in the previous (development) study, which is counterintuitive to the notion that inferences and guidelines to enhance evidence-based medicine should be based on as much information as possible. In addition, doctors are faced with the impracticable situations of having to decide which model to use in their patients, when many have been developed for the same outcome. A much better alternative to redeveloping new models in each new patient sample is to update existing prediction models and adjust or recalibrate them to the local circumstances or setting of the validation sample at hand. As a result, the adjusted, or updated, models combine the information captured in the original model with information from new individuals.[48,50,93,119,120] Hence, the updated models are adjusted to the characteristics of new individuals and probably have improved transportability to other individuals.

Updating a model is often desirable.[48,93,103] In particular, some systematic miscalibration is common for predictions obtained from prognostic models in settings that differ from that of the development sample. Updating methods include recalibrating the model to the new setting or investigating the addition of new prognostic factors, including biomarkers, to an existing model.[48] Ideally there should be an ongoing process of model validation and updating.[48,93,103] The contribution of genomic, proteomic, or metabolomics measures and new imaging tests over and above established prediction models is a key issue in current prognostic research.[119,121] For example, a simple model for patients with traumatic brain injury that included just three strong prognostic factors was extended with computed tomography results in a second stage, and laboratory test results in a third stage.[122] The more extended model yielded more refined predictions and better discrimination.

The importance of assessing the impact of new markers on the accuracy of a model is widely agreed, but how best to quantify any changes in prediction is an active topic of methodological research.[123-125] The recent trend when comparing models is to consider the extent of reclassification of individual patients between risk groups rather than using global measures of discrimination such as the area under a receiver operating characteristic (ROC) curve.[124,126] These different statistics are mathematically related, however.[127,128] The addition of new markers may yield only marginal benefit.[129] Because standard models generally include important predictors, the independent effects of new prognostic factors need to be quite strong before a clinically useful improvement is achieved.[130] Furthermore the measurement of new markers carries cost implications.[119]

Several methods for updating prediction models have been proposed and evaluated (**Table 1.7.1**).[48,103] Most often, differences are seen in the outcome or event frequency between the development and new validation sample. These result in poor calibration of the model in the latter, due to predicted probabilities being systematically too high or too low. By adjusting the baseline risk or hazard (if known) of the original prediction model to the individuals in the validation sample, calibration can easily be improved.[98,103] This requires the adjustment of only one parameter of the original model (Method 1 presented in **Table 1.7.1**). Additional updating methods vary from overall adjustment of all predictor weights simultaneously, adjustment of a particular predictor weight, to the addition of a completely new predictor or marker to the existing model (**Table 1.7.1**). Note that simple updating methods (Methods 1 and 2 in **Table 1.7.1**) at best improve calibration; discrimination remains unchanged as the relative ranking of the model’s predicted probabilities stays the same after the updating. To improve discrimination, methods 3-6 are needed.

Application of the above methods leads to updated models which are adjusted to the circumstances of the validation sample. However, just like a newly developed model, updated models should still be tested on their transportability and impact before they can be applied in routine practice.[93] Individual participant data from the new sample are needed for model updating, using standard methods (**Table 1.7.1**) and these may not be available in some settings. In this case, it still may be possible to perform a simple adjustment to the prediction model should the frequency of the outcome and mean levels of the predictors in the new population be available.[98,131]

Table 1.7.1: Updating methods for prediction models.

Method	Updating method	Reason for updating
0	No adjustment (the original prediction model)	-
1	Adjustment of the intercept (baseline risk)	Difference in the outcome frequency (prevalence or incidence) between development and validation sample
2	*Method 1* + adjustment of all predictor regression coefficients by one overall adjustment factor	Regression coefficients of the original model are overfitted (or underfitted)
3	*Method 2* + extra adjustment of regression coefficients for predictors with different strength in the validation sample as compared with the development sample	As in method 2, and the strength (regression coefficient) of one or more predictors may be different in the validation sample

4	*Method 2* + stepwise selection of additional predictors	As in method 2, and one or more potential predictors were not included in the original model, or a newly discovered marker may need to be added
5	Re-estimation of all regression coefficients, using the data of the validation sample only	The strength of all predictors may be different in the validation sample, or the validation sample is much larger than the development sample
6	*Method 5* + stepwise selection of additional predictors	As in method 5, and one or more potential predictors were not included in the original model

A particular motivation to update a prognostic model is to replace existing predictors that suffer from substantial interobserver variability (such as physical examination, imaging, and histopathological techniques) with more reliably measured markers. Moreover, prognostic models that include factors or markers with a causal effect on the outcome under study may be expected to be more generalisable to other populations. Such models may also be better used, since they are linked to biological (or other) pathways rather than merely based on statistical association. While these suggestions are plausible, empirical evidence is lacking.

1.8 Impact studies

Impact studies aim to quantify whether the use of a prognostic model by practising physicians truly effectively improves their decision-making and ultimately patient outcomes. A prognostic model can influence patient outcome or the cost effectiveness of care only when changes in clinical management are made based on the prognostic information provided. Prognostic models are developed to provide estimates of outcome probabilities to complementary support clinical intuition and guidelines.[5,97] The effect of a previously developed and validated prognostic model on doctors decisions and behaviour and patient outcomes should be studied separately in so called impact studies. Validation and impact studies differ in their design, study outcome, statistical analysis, and reporting. A validation study ideally uses a prospective cohort design and does not require a control group.[109] For each patient, predictors and outcome are documented, and the rule's predictive performance is quantified. By contrast, impact studies quantify the effect of using a prognostic model on doctors' decisions, patient outcome, and/or cost effectiveness of care compared with and without using such model. They require a control group of healthcare professionals who provide usual care. Model impact studies thus follow a comparative intervention design, rather than the single cohort design used in model development or validation studies, and are ideally randomized trials.[46] If behaviour changes of professionals is the main outcome, a randomised study without follow-up of patients would suffice. Follow-up is required if patient outcome or cost effectiveness is assessed. However, since changes in outcome depend on changes in doctors' behaviour, it may be sensible to start with a randomised study assessing the model's impact on therapeutic decisions, especially when long follow-up times are needed to assess patient outcome.

Impact studies may use an assistive approach—simply providing the model's predicted probabilities of an outcome between 0% and 100%—or a decisive approach that explicitly suggests decisions for each probability category.[46] The assistive approach clearly leaves room for intuition and judgment, but a decisive approach may have greater effect.[132,133] Introduction of computerised patient records that automatically give predictions for individual subjects, enhances implementation and thus impact analysis of prognostic models in routine care.[132]

The comparison in impact studies is scientifically strongest when a cluster randomised trial is used.[93] One may randomize healthcare professionals (as clusters) or centres (practices). Randomising individual patients in an impact study may result in learning effects because the same doctor will alternately apply and not apply the model to subsequent patients, reducing the contrast between both randomised groups. Randomisation of doctors (clusters) is preferable, although this requires more patients.[134] Randomising centres is often the best method as it avoids exchange of experiences between doctors within a single centre. Although impact studies are scarce, are a few good examples exist.[135,136] An appealing variant, of a cluster randomised trial, particularly for complex or multifaceted interventions that need to be introduced into routine care, is the stepped-wedge (cluster randomised) trial.[137-139] Stepped wedge means that clusters for example, hospitals or general practitioner practices, are randomly allocated a time period when they are given the intervention, here the prediction model. All the clusters will be applying both care-as-usual (control) and the prediction model (intervention), but the time when they receive this prediction model is randomly ordered across the clusters. This is one-way crossover cluster trial, where the clusters cross over typically from control to intervention[137-140] at regular, randomly allocated time intervals.

Because randomised trials are expensive, time consuming and difficult to be properly conducted, other approaches are possible. One such approach is the prospective "*before-after*" impact study, in which comparison is made on the outcomes that are measured in a time period before the model was introduced versus a time period after which the model was made available to the same care providers. However, this design is sensitive to temporal changes in, for example, therapeutic approaches. A subtle variant to the beforeeafter approach, and therefore sharing the same limitations, is the "on-off" impact study where the outcome is measured in alternating time periods when the prediction model is or is not available in a particular centre.[46] Here, a problem is that the practising care providers in the centre may have changed over time, which may bias results.

An attractive alternative when outcomes are relatively rare, or when a long follow-up is required, is decision analytic modelling.[119] This approach starts with a well-developed and externally validated (and perhaps updated) model, and combines information on model predictions with information about the effectiveness of treatments from randomised therapeutic trials or meta-analyses. If such an approach fails to show improved outcome or favourable cost-effectiveness, a long-term randomised impact study may not even be indicated.

However, do all prognostic models require a three step assessment (development, validation/updating, evaluation in impact studies) before they are used in daily care? Does a model that has shown adequate prediction for its intended use in validation studies—adequately predicting the outcome—still require an impact analysis using a large, multicentre cluster randomised study? For models with less perfect performance, only an impact analysis can determine whether use of the model is better than usual care. Impact studies also provide the opportunity to study factors that may affect implementation of a prognostic model in daily care, including the acceptability of the prognostic model to clinicians and ease of use.

1.9 Evaluating prediction models

1.9.1 Discrimination

Discrimination is a measure of how well the prediction model can separate those who will and will not develop the outcome of interest. If the predicted values for cases are all higher than for non-cases, we say the model can discriminate perfectly, even if the predicted risk does not match the proportion with disease. The most popular measure of model fit in the cardiovascular literature has been the c statistic, a measure of discrimination also known as the area under the ROC curve, or the c index, its generalization for survival data.[92,141] The ROC curve and its associated c statistic are functions of the sensitivity and specificity for each value of the measure or model. The ROC curve is a plot of sensitivity versus 1-specificity (often called the false-positive rate) that offers a summary of sensitivity and specificity across a range of cut points for a continuous predictor. Discrimination is of most interest when classification into groups with or without prevalent disease is the goal, such as in diagnostic testing.[142]

ROC analysis is a useful tool for evaluating the performance of prognostic models and more generally for evaluating the accuracy of a statistical model (eg, logistic regression, linear discriminant analysis) that classifies subjects into 1 of 2 categories, diseased or nondiseased. The area under the curve (AUC), that uses the ROC curve, is a nonparametric test statistic (the Mann–Whitney U test equivalent to a Wilcoxon rank sum test) of equality of distribution of estimated risk in cases and controls. AUC equals 0.5 when the ROC curve corresponds to random chance and 1.0 for perfect accuracy. The value of 0.80 seems to be a common cut-off between acceptable and poor models. On rare occasions, the estimated AUC is <0.5, indicating that the test does worse than chance. Because the AUC is based solely on ranks, it is less sensitive than measures based on the likelihood or other global measures of model fit. This characteristic may make it a poor choice for the selection of variables to be used in a predictive model.

The difference between the clinical and statistical views is manifested in one particularly important way: the role of variability in the risk. Less variability, or more homogeneity, in risk within cases and controls increases the AUC and other measures of discrimination; in contrast, greater variation in risk in the study population for which decisions are to be made increases the potential for assigning an intervention for those at extreme risks different from the intervention appropriate for one at average risk. In other words, small variance in risk, conditional on disease, increases discrimination, but large unconditional variance increases the potential for clinical utility. Of course, the variation needs to be real, not a consequence of random variation of risk estimates or misclassification of markers in the model.

1.9.2 Calibration

The AUC (or c-statistic), achieved popularity in diagnostic testing, in which the test characteristics of sensitivity and specificity are relevant to discriminating diseased versus nondiseased patients. The AUC, however, may not be optimal in assessing models that predict future risk or stratify individuals into risk categories. In this setting, calibration is as important to the accurate assessment of risk. Calibration quantifies how closely the predicted probabilities of an event match the actual experience.[143] In other words, calibration is a measure of how well predicted probabilities agree with actual observed risk. When the average predicted risk within subgroups of a prospective cohort, for example, matches the proportion that actually develops disease, we say a model is well calibrated. The Hosmer-Lemeshow statistic compares these proportions directly and is a popular, though imperfect, means to assess model calibration.[144] When evaluating the performance of a model after addition of a new marker, it is essential to check for improvement (or at least no adverse effect if other measures improve) in calibration, which can be quantified by, for example, the Hosmer–Lemeshow's chisquare or its modifications.

Recalibration of prediction models to other settings can be done by adjustment at three levels: the baseline disease risk, the average predictor values, and the predictor-outcome associations. The most popular measure of calibration, the Hosmer-Lemeshow goodness-of-fit test, forms such subgroups, typically using deciles of estimated risk. Within each decile, the estimated observed proportion and average estimated predicted probability are estimated and compared. The statistic has a x^2 distribution with g – 2 degrees of freedom, where g is the number of subgroups formed. Although deciles are most commonly used to form subgroups, other categories, such as those formed on the basis of the predicted probabilities themselves (such as 0 to <5%, 5 to <10%, etc.), may be more clinically useful.

Because groups must be formed to evaluate calibration, this test is somewhat sensitive to the way such groups are formed.[144] Ideally the predicted probability would estimate the underlying or true risk for each individual (perfect calibration). Since we cannot know the underlying risk, but can only observe whether the individual gets the disease, a stochastic event, the Hosmer-Lemeshow statistic is a somewhat crude measure of model calibration.

1.9.3 Assessing improvement in model performance (net reclassification improvement (NRI), integrated discrimination improvement (IDI))

Identification of key factors associated with the risk of developing cardiovascular disease and quantification of this risk using multivariable prediction algorithms are among the major advances made in preventive cardiology and cardiovascular epidemiology in the 20th century. The ongoing discovery of new risk markers by scientists presents opportunities and challenges for statisticians and clinicians to evaluate these biomarkers and to develop new risk formulations that incorporate them. One of the key questions is how best to assess and quantify the improvement in risk prediction offered by these new models. Demonstration of a statistically significant association of a new biomarker with cardiovascular risk is not enough. Some researchers have advanced that the improvement in the area under the receiver-operating-characteristic curve (AUC) should be the main criterion, whereas others argue that better measures of performance of prediction models are needed. New risk factors or markers are being identified and proposed constantly, and as with each other for consideration for incorporation into risk prediction algorithms. The critical question arises as to how to evaluate the usefulness of a new marker. The most basic necessary condition required of any new marker is its statistical significance. It is hard to imagine that one would argue for an inclusion of a new marker into a risk prediction formulation if it is not related to the outcome of interest in a statistically significant manner. Statistical significance, however, does not imply either clinical significance or improvement in model performance. Indeed, many biomarkers with weak or moderate relations to the outcome of interest can be associated in a statistically significant fashion if examined using a large enough sample size.

Researchers, extending existing methodology, began evaluating new markers based on their ability to increase the AUC. It quickly became apparent that, for models containing standard risk factors and possessing reasonably good discrimination, very large 'independent' associations of the new marker with the outcome are required to result in a meaningfully larger AUC.[130,145,146] None of the numerous new markers proposed comes close in magnitude to these necessary levels of association. In response to this, some scientists have argued that we need to wait for new and better markers; other researchers have sought model performance measures beyond the AUC to evaluate the usefulness of markers. Reassignment of subjects into risk categories (reclassification tables) and predictiveness curves form opposite ends of the spectrum of new ideas.[147]

Net reclassification and integrated discrimination improvements have been proposed as alternatives to the increase in the AUC for evaluating improvement in the performance of risk assessment algorithms introduced by the addition of new phenotypic or genetic markers. These two metrics are used to assess improvement in model performance offered by a new marker. The NRI and IDI provide supplementary information over the difference in the areas under the receiver operating characteristic (ROC) curves (AUCs). The NRI focuses on reclassification tables constructed separately for participants with and without events, and quantifies the correct movement in categories—upwards for events and downwards for non-events. The improvement in AUC for a model containing a new marker is defined simply as the difference in AUCs calculated using a model with and without the marker of interest. This increase, however, is often very small in magnitude; for example, Wang et al. show that the addition of a biomarker score to a set of standard risk factors predicting CVD increases the model AUC only from 0.76 to 0.77.[148] Ware and Pepe show simple examples in which enormous odds ratios are required to meaningfully increase the AUC.[130,145] Reclassification tables have been gaining popularity in medical literature.[26,149] Unfortunately, reclassification tables constructed and interpreted in this manner offer limited means of evaluating improvement in performance. Relying solely on the number or percentage of subjects who are reclassified can be misleading. Additionally, calculating event rates among the reclassified individuals does not lead to an objective assessment of the true improvement in classification. For instance, even if someone reclassify 100 people from the 10–20% 10-year CVD risk category into the above 20% group and the 'actual' event rate among these individuals is 25%, improved the placement of 25 people, but not the remaining 75 who should have stayed in the lower risk category. Therefor a different way of constructing and interpreting the reclassification tables is suggested. The reclassification of people who develop and who do not develop events should be considered separately. Any "upward" movement in categories for event subjects (i.e. those with the event) implies improved classification, and any "downward" movement indicates worse reclassification. The interpretation is opposite for people who do not develop events. The improvement in reclassification can be quantified as a sum of differences in proportions of individuals moving up minus the proportion moving down for people who develop events, and the proportion of individuals moving down minus the proportion moving up for people who do not develop events. This sum is called NRI. Equivalently, the NRI can be calculated by computing the difference between the proportions of individuals moving up and the proportion of individuals moving down for those who develop events and the corresponding difference in proportions for those who not develop events, and taking a difference of these two differences.

The IDI assesses the improvement in average sensitivity without sacrificing average specificity. The IDI does not require categories, and focuses on differences between improve integrated sensitivities and one minus specificities' for models with and without the new marker. The IDI can be defined as the difference in discrimination slopes between two models-- one with, and the other without, the added variable. Discrimination slope was initially introduced as a "useful performance measure for it quantifies in a simple manner the separation of positive and negative outcomes”. Recently this argument was further supported by calling it “a highly recommendable” measure of explanatory power for binary outcome models. It is defined as a difference in the means of the model-based event probabilities, that is, a subtraction of the nonevents from the events.[150]

1.10 Clinical utility of prediction models

The clinical use of prognostic models should be dependent on evidence of successful validation and, preferably, on evidence of studies of clinical impact when using the model. Statistical and clinical perspectives on risk models can be very different, even with agreement on the objective: to develop accurate and precise risk estimates for rational, effective, and cost-effective prevention strategies. A good prediction model needs to be relatively easy to incorporate into routine clinical practice. A model will have no clinical impact unless measuring the variables needed for the model is feasible, using the model is acceptable to clinicians, and applying the model does not markedly increase workload or cost.[151,152] Application of prognostic models requires unambiguous definitions of predictors and reproducible measurements using methods available in clinical practice. Practitioners may be less experienced in properly coding this predictor for a patient, leading to misclassification that potentially compromises the rule's predictive performance. A complex prediction model that adds little to known risk factors will not (and probably should not) be used by most clinicians.[153]

A model that improves prognostic accuracy may be helpful to clinicians and to patients, but the model will have much more clinical impact if the outcome predicted (i.e. death, disease) can be prevented or delayed with effective treatment or if treatment can be individualized to improve outcomes. For example, the Framingham Risk Score[154] is widely used to determine which patients should receive lipid-lowering treatment for cardiovascular disease, and the Model for End-Stage Liver Disease (MELD)[155] score is used to prioritize patients for liver transplantation. The clinical usefulness of these scores depends on the availability of effective treatments for the predicted outcome.

Interestingly, a key factor for successful implementation of a prognostic model seems to be whether a model is supported by leading professionals in the field of application. Other factors that might be associated with use of prognostic models in practice include the complexity of the model (a few or many prognostic factors), the format of the model in which is available (as a score chart on paper, web based, or as standard part of an electronic patient record), the use of cut-off values for model predictions to guide decision making (rather than only providing the predicted probability), the ease of use in the consulting room, the clinical context, and the fear of "cookbook medicine" or medicolegal consequences of undue reliance on model based predictions and decisions.[46,93]

Like other tests, prediction models can increase harm to patients and costs if their use leads to testing or procedures that might not otherwise be performed. For example, a cardiac risk stratification algorithm that included cardiac computed tomographic scanning might increase radiation exposure compared with one that included only stress echocardiography.[35,156] Whether the potential harms and increased cost of a prediction model are balanced by improved diagnostic accuracy is an empirical question that should be addressed before a model is widely used. Of course, clinical utility also depends on the efficacy of available interventions; a model with prefect predictions has no clinical utility without an effective intervention.

1.11 Aims and outline

There is great interest in moving beyond established risk factors for the prediction of cardiovascular disease outcomes by incorporating a variety of information inclusive of demographic characteristics, biomarkers, and genetic factors among others into risk prediction models. Given the abundance of published prediction models across almost all clinical domains, critical appraisal and synthesis (whenever possible) of the available evidence is a requirement to allow researchers, care providers, and policymakers to identify possible pitfalls of newly introduced models, and in addition to determine which models may be useful in different situations. While the objective of all predictive models is to develop accurate and precise risk estimates for rational, effective, and cost-effective prevention and treatment strategies, the statistical and clinical perspectives on risk models can be very different. The aim of this work was to evaluate prognostic studies aimed at predicting outcomes using risk prediction models rather than studies investigating single variables. We aimed to evaluate the discriminating performance of predictive tools for death and the variability in this performance across different clinical settings and studies. Moreover, we evaluated the evidence on comparisons of established and widely used cardiovascular risk prediction models and collected comparative information on their relative prognostic performance. Finally, we focused on how often newly developed risk prediction models undergo independent external validation and how well they perform in such validations, an important step of model evaluation before wide application.

Section 2.

Predicting death: Seemingly well-validated predictive tools are not very accurate with a wide variation of predictive performance.

2. Predicting death: an empirical evaluation of predictive tools for mortality.[157]

Siontis GC, Tzoulaki I, Ioannidis JP.

Archives of Internal Medicine. 2011. (PMID: 21788535)

ORIGINAL INVESTIGATION

ONLINE FIRST

Predicting Death

An Empirical Evaluation of Predictive Tools for Mortality

George C. M. Siontis, MD; Ioanna Tzoulaki, PhD; John P. A. Ioannidis, MD, DSc

Background: The ability to predict death is crucial in medicine, and many relevant prognostic tools have been developed for application in diverse settings. We aimed to evaluate the discriminating performance of predictive tools for death and the variability in this performance across different clinical conditions and studies.

Methods: We used Medline to identify studies published in 2009 that assessed the accuracy (based on the area under the receiver operating characteristic curve [AUC]) of validated tools for predicting all-cause mortality. For tools where accuracy was reported in 4 or more assessments, we calculated summary accuracy measures. Characteristics of studies of the predictive tools were evaluated to determine if they were associated with the reported accuracy of the tool.

Results: A total of 94 eligible studies provided data on 240 assessments of 118 predictive tools. The AUC ranged from 0.43 to 0.98 (median [interquartile range], 0.77 [0.71-0.83]), with only 23 of the assessments reporting excellent discrimination (10%) (AUC, >0.90). For 10 tools, accuracy was reported in 4 or more assessments; only 1 tool had a summary AUC exceeding 0.80. Established tools showed large heterogeneity in their performance across different cohorts (I^2 range, 68%-95%). Reported AUC was higher for tools published in journals with lower impact factor ($P=.01$), with larger sample size ($P=.01$), and for those that aimed to predict mortality among the highest-risk patients ($P=.002$) and among children ($P<.001$).

Conclusions: Most tools designed to predict mortality have only modest accuracy, and there is large variability across various diseases and populations. Most proposed tools do not have documented clinical utility.

Arch Intern Med.
Published online July 25, 2011.
doi:10.1001/archinternmed.2011.334

Author Affiliations: Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece (Drs Siontis, Tzoulaki, and Ioannidis); Department of Epidemiology and Biostatistics, Imperial College of Medicine, London, England (Drs Tzoulaki and Ioannidis); the Institute for Clinical Research and Health Policy Studies, Department of Medicine, Tufts University School of Medicine, Boston, Massachusetts (Dr Ioannidis); the Department of Epidemiology, Harvard School of Public Health, Boston (Dr Ioannidis); and the Stanford Prevention Research Center, Stanford University School of Medicine, Stanford, California (Dr Ioannidis).

THE ABILITY TO PREDICT death accurately is crucial for conveying information to patients about their future; for making sound medical decisions for management, treatment, and prevention; and for having realistic expectations. Evidence suggests that physicians perform poorly in predicting when patients will die.[1,2] However, numerous models have been developed to predict mortality in diverse settings.[3-5]

Herein we aim to empirically evaluate the ability of available predictive tools (multivariate or single variables) to predict the risk of death accurately for diverse conditions and populations. We assess how accurately and consistently these tools perform to help understand their potential clinical utility.

METHODS

SEARCH STRATEGY

To evaluate recently published studies that assessed the accuracy (discrimination) of tools to predict mortality, we searched Medline for studies published in 2009 by using the Clinical Queries tool. For more details on our search strategy and data extraction, see the eAppendix (www.archinternmed.com).

STUDY SELECTION

We included studies of any design published in 2009 that assessed the accuracy of tools to predict mortality (either single predictors or multivariable models); included assessment of accuracy based on the area under the receiver operating characteristic curve (AUC) (aka, C statistic or C index); and focused on all-cause death as the primary outcome. The AUC[6-9] is the most commonly used metric for assessing the accuracy of predictive tools.[10] The AUCs can be compared across different tools, while relative risk metrics depend on the unit to which they are expressed and cannot directly compare predictive tools expressed for different units of measurement.[11]

We excluded studies that only had data on the development of a new predictive tool or validated the predictive tool in the same cohort where it was developed because new, nonvalidated predictive tools are likely to have inflated estimates of accuracy.[12-14] We also excluded articles that did not provide primary data (eg, reviews) and studies where death was part

2.1 BACKGROUND

The ability to predict death accurately is crucial for conveying information to patients about their future; for making sound medical decisions for management, treatment, and prevention; and for having realistic expectations. Prediction models (also called clinical prediction rules, clinical decision rules, or risk scores) are tools designed to assist clinical decision-making. Prediction models generally provide an estimate of the risk of disease, disease outcome, or the benefit of a diagnostic or therapeutic action.[52,97] For example, the well-known Framingham Risk Score[154] uses age, sex, total cholesterol level, high-density lipoprotein cholesterol level, smoking status, blood pressure, and use of hypertension medication to estimate the risk of myocardial infarction or coronary death during the next 10 years. Because prediction models are designed to guide clinical care, it is important that they be accurate and reliable.

Over the past few years, many studies describing the development and/or validation of prediction models have been published, undoubtedly facilitated by access to large electronic databases; whereas many others have never achieved to be published. Why is a good prediction model so hard to find? While prediction models are reasonably easy to develop, a prediction model that remains accurate across different populations and is clinically useful is rare. Prognostic tools should be evaluated in several sequential stages: initial model performance (model development), prospective validation in independent cohorts (external validation of a model), impact on patient management and outcome and cost-effectiveness. However, even for established and widely used prognostic tools, many of these steps suffer from methodological limitations and in many cases are missing.

With heightened interest in predictive medicine, many studies try to document information that can improve prediction of major clinical outcomes. Evidence suggests that physicians perform poorly in predicting when patients will die.[158-160] However, numerous models have been developed to predict mortality in diverse settings.[161-163] Herein we aim to empirically evaluate the ability of available predictive tools (multivariate or single variables) to predict the risk of death accurately for diverse conditions and populations. We assess how accurately and consistently these tools perform to help understand their potential clinical utility.

2.2 METHODS

Search Strategy

The PubMed literature search was filtered by the specific clinical study categories of "*prognosis*" and "*clinical prediction guides*", which was further limited by the filter of a narrow and specific search (Narrow/Specific [filter]). The terms "*AUC OR area under the curve OR c statistic OR c index*" and "*death OR mortality OR survival*" were applied. We set no limits for publication type or language. All items were initially evaluated for eligibility based on title and abstract. Potentially eligible studies were retrieved and scrutinized in full-text.

Study Selection

In this empirical evaluation, we included studies of any design published in 2009 that assessed the accuracy of tools to predict mortality (either single predictors or multivariable models); included assessment of accuracy based on the area under the receiver operating characteristic curve (AUC) (aka, C statistic or C index); and focused on all-cause death as the primary outcome. The AUC[92,148,164,165] is the most commonly used metric for assessing the accuracy/discriminatory ability of predictive tools.[166] The AUCs can be compared across different tools, while relative risk metrics depend on the unit to which they are expressed and cannot directly compare predictive tools expressed for different units of measurement.[130,166]

We excluded studies that only had data on the development of a new predictive tool or validated the predictive tool in the same cohort where it was developed because new, non-validated predictive tools are likely to have inflated estimates of accuracy.[50,111,167] We also excluded articles that did not provide primary data (eg, reviews) and studies where death was part of a composite outcome or was determined as cause-specific (rather than all-cause) mortality.

When there were several eligible predictive tools and/or they assessed the ability to predict death at different lengths of follow-up in the same cohort, each proposed predictive tool and each time of follow-up assessment was included separately. For example, studies that examined 2 or more different predictive tools at different follow-up periods.

Data Extraction

When studies examined previously developed predictive tools, we extracted the AUC values of all previously developed tools corresponding to all-cause mortality. When studies developed and proposed new predictive tools, we extracted the AUC values of all examined tools (newly and previously developed) in the external validation set only. We gave preference to keep information on the whole study population over subgroups. For each eligible AUC we extracted the specific value and the respective 95% confidence interval whenever available.

For each eligible prognostic study we recorded the first author, journal, impact factor of the journal (according to Thomson Journal Citation Reports), country of origin of the corresponding author or group investigators (USA, Europe, other), the study design (assessment of overall death prediction in a prospectively collected study population or in a retrospectively evaluated dataset), whether the assessment of the variables included in the predictive tool was blinded, and the percentage of losses to follow-up.

For each study and for each predictive tool, we recorded the total sample size and the number of deaths when a previously developed predictor/model was used and the sample size and number of deaths for the validation group, when a new tool was developed. Study populations were categorized according to their baseline status as healthy, acute disease (conditions that need acute evaluation and intervention e.g. trauma, sepsis, emergency surgery), chronic disease, or populations with both acute and chronic disease patients. Moreover, we recorded whether primarily adults or children were involved.

Studies were categorized into the following general fields based on the disease/clinical condition: cardiovascular diseases, critical illnesses, infectious diseases, gastroenterology-related diseases, malignancies, trauma, or other. We recorded the average follow-up corresponding to each of the extracted AUCs, giving preference to mean>median>in-hospital follow-up estimates. When no information was given in the text, duration of follow up was derived from Kaplan-Meier curves, where applicable. When only maximum follow-up duration under investigation was given, we approximated the mean follow-up by using the formula *mean follow-up=maximum follow-up * (1 - (0.5 * proportion of deaths))* which assumes that each dying patient contributes on average half of the maximum follow-up. The *death rate* (per month of follow up) was calculated by the *number of events / (sample size * mean follow-up).*

We noted whether a single predictor or a predictive model with multiple predictors was used. For each prognostic model we recorded the number and the included set of variables. Furthermore, we noted whether the accuracy of the prediction in each study was assessed by any measure of calibration,[168] and, if so, recorded the applied method, and whether the authors presented the calibration results. Calibration examines whether the risk prediction is equally good for patients at different levels of risk or there is a lack of fit. Finally, we identified the studies in which reclassification analysis[125,169,170] was performed and recorded the respective metrics. Reclassification examines whether the predictive tool helps classify patients in different, more appropriate risk categories compared with what could be done without its knowledge or compared with some other model.

The selection of the eligible studies and data extraction was performed independently by two investigators; whereas any discrepancies were resolved by consensus and arbitration by a third investigator.

Statistical Analysis

The AUC was defined as mean (SD) or median (interquartile range [IQR]). An AUC of 1 indicates perfect discrimination, while an AUC of 0.5 indicates discrimination no better than chance. While there are no absolute thresholds, usually an AUC of greater than 0.80 is considered to show very good discrimination, and AUC greater than 0.90 suggests excellent discrimination.[92] For predictive tools where there was more than 1 assessment available, we noted the range of AUC values. For predictive tools with at least 4 data sets where both the AUC and corresponding 95% confidence intervals were available, we summarized the AUC estimates using random effect models, weighting the AUC of each data set by the inverse of the sum of the between and within-study variances. We quantified the heterogeneity in AUC values by the I^2 metric and its 95% confidence intervals. The I^2 metric takes values between 0% and 100%, and it is independent of the number of data sets (50%-75% indicates moderate heterogeneity, while >75% indicates very large heterogeneity).[171] We compared the AUC values among pre-specified subgroups based on prevalence of disease and predictive tool characteristics using 1-way analysis of variance for categorical variables and the Spearman correlation coefficient for continuous variables. All analyses were performed with STATA software, version 10.0 (StataCorp LP, College Station, Texas).

2.3 RESULTS

Eligible Studies and Predictive Tools

Overall 544 items were retrieved from Medline, of which 235 were reviewed in full text. Of those, 94 articles (**2.5.1 Supplementary references**) were deemed eligible (**Figure 2.1**). The interrater agreement for the selection of the eligible studies had κ value of 0.86. These 94 manuscripts presented data on 240 assessments (224 multivariate models and 16 single predictors) of the accuracy of 118 predictive tools. Characteristics of studies and predictive tool assessments are provided below (**Table 2.1**).

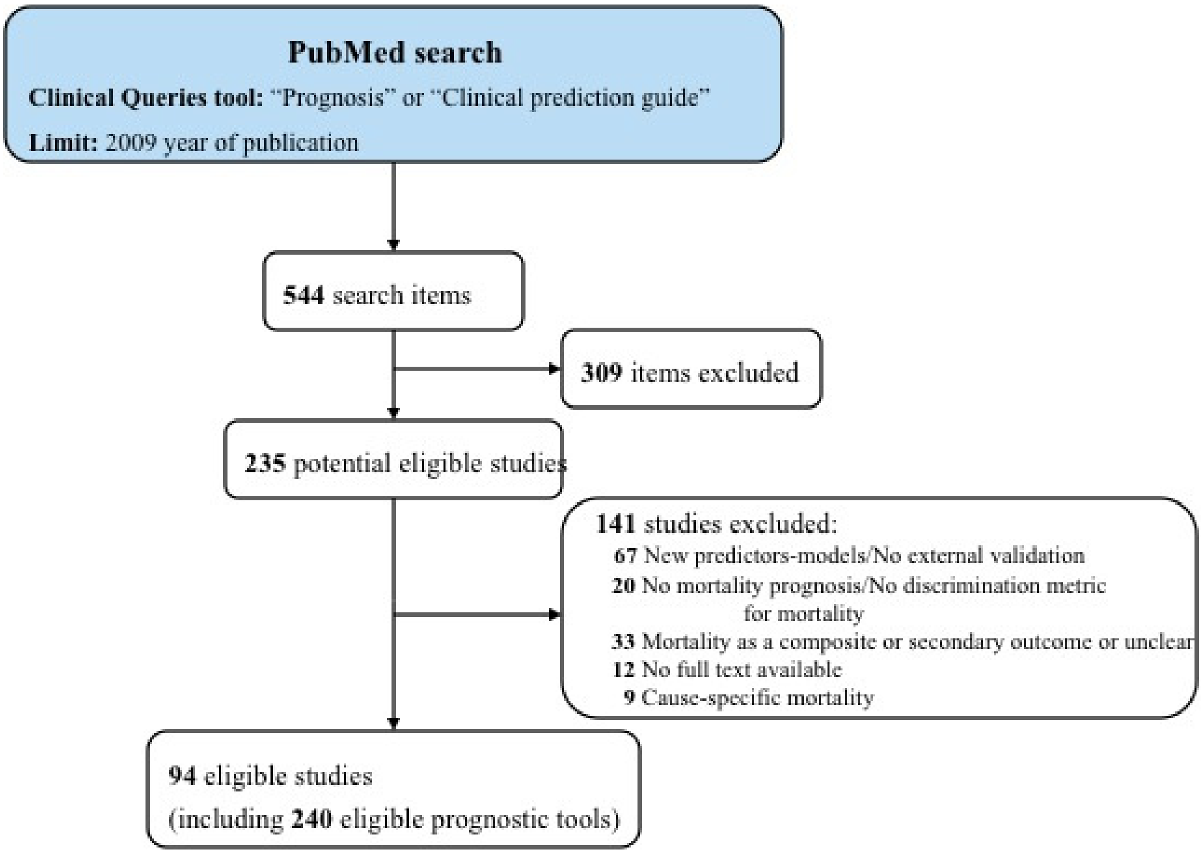
PubMed search
Clinical Queries tool: "Prognosis" or "Clinical prediction guide"
Limit: 2009 year of publication
544 search items
309 items excluded
235 potential eligible studies
141 studies excluded:
67 New predictors-models/No external validation
20 No mortality prognosis/No discrimination metric for mortality
33 Mortality as a composite or secondary outcome or unclear
12 No full text available
9 Cause-specific mortality
94 eligible studies
(including 240 eligible prognostic tools)

Figure 2.1: Selection of eligible studies of all-cause death prediction.

Most of the studies were performed in the United States or Europe, had a prospective cohort design, and pertained to acute disease conditions. Cardiovascular, critical-illness, infectious, gastroenterology-related,

and malignant diseases accounted for 83% of the cohorts, but many other diseases were also assessed (**Table 2.1**, **2.5.2** and **2.5.3 Supplementary tables**). The median (IQR) sample size for the assessments was 502 (185-2016); the median (IQR) number of deaths was 71 (32-157); the median (IQR) proportion of deaths was 14% (5%-29%); and the median (IQR) death rate was 13% (4%-44%) per month. Among the whole data set (94 studies), in only 1 study (S85 in **2.5.1 Supplementary references**) did the investigators review and abstract patient data blinded to patients' hospital course and clinical status (**Table 2.1**). For 78 studies, the percentage of losses to follow-up was available (70 studies reported no losses, while for the rest loss was generally low (median [IQR] loss to follow-up, 3.5% [1.25%-10.25%]).

Table 2.1: Characteristics of eligible studies and predictive tools.

Characteristic	**No. (%)** **Prognostic studies (n=94)**	 **Predictive tools (n=240)**
Type of study		
New externally validated tools	29 (31)	72 (30)
Previously developed predictive tools	65 (69)	168 (70)
Area of origin		
United States	21 (22)	49 (20)
Europe	43 (46)	113 (47)
Other	30 (32)	78 (33)
Study design		
Prospective	53 (56)	139 (58)
Retrospective	41 (44)	101 (42)
Disease status		
Acute disease	55 (59)	130 (54)
Chronic disease	29 (31)	91 (38)
Mixed (acute/chronic)	10 (11)	19 (8)
Study population		
Pediatric	5 (5)	7 (3)
Adult	88 (94)	225 (94)
Both	1 (1)	8 (3)
Disease/Clinical condition		
Cardiovascular	18 (19)	40 (17)
Critical illness	16 (17)	42 (18)
Gastroenterology	14 (15)	50 (21)
Infectious	15 (16)	37 (15)
Malignancies	9 (10)	30 (13)
Other	22 (23)	41 (17)
In-hospital mortality		
Yes	44 (47)	99 (41)
No	50 (53)	141 (59)
Predictive variables were assessed blinded to the outcome		
Yes	1 (1)	1 (0.4)
No	93 (99)	239 (99.6)
Information on loss to follow-up		
Not available	16 (17)	40 (17)
Available	78 (83)	200 (83)
Loss of follow-up 0%	70 (90)	158 (79)
Loss of follow-up >0%	8 (10)	42 (21)

Predictive Tools

Overall, 110 different predictive models and 8 different predictors were examined in the 240 assessments. The most commonly evaluated models included the Acute Physiology And Chronic Health Evaluation (APACHE) II model (n=19) and the MELD score (Model for End-Stage Liver Disease) (n=17) (**Table 2.2**). The predictive models included a wide range of variables (**2.5.2** and **2.5.3 Supplementary tables**). The number of variables in the models ranged from 2 to 30, and the median (IQR) number was 6 (4-12). All of the identified single predictors were biomarkers (**2.5.3 Supplementary tables**).

Table 2.2: AUC values of predictive tools examined >1 assessment.

Predictive tool	No. of assessments	AUC			
		Median	IQR	Min	Max
AMIS model	2	0.86	0.84-0.87	0.84	0.87
APACHE II	19	0.77	0.71-0.81	0.69	0.94
BCLC score	2	0.85	0.84-0.86	0.84	0.86
BISAP score	2	0.82	NA	0.82	0.82
BNP	3	0.66	0.63-0.69	0.63	0.69
CLIP score	5	0.88	0.64-0.88	0.62	0.96
CRIB II	2	0.91	0.90-0.92	0.90	0.92
CTP score	11	0.73	0.72-0.84	0.61	0.88
CURB-65 score	5	0.78	0.73-0.78	0.64	0.82
CCI	3	0.67	0.63-0.74	0.63	0.74
EuroSCORE	6	0.74	0.70-0.77	0.70	0.80
ISS	2	0.63	0.54-0.72	0.54	0.72
Intermountain Risk Score	3	0.87	0.84-0.87	0.84	0.87
JIS	5	0.85	0.64-0.87	0.59	0.87
MELD score	17	0.81	0.78-0.86	0.77	0.89
MELD - Sodium score	4	0.81	0.78-0.86	0.77	0.89
MESO index	3	0.87	0.69-0.88	0.69	0.88
MPI	3	0.80	0.79-0.83	0.79	0.83
MPM II	2	0.73	0.66-0.79	0.66	0.79
NT-proBNP	6	0.74	0.71-0.76	0.67	0.77
Pediatric death prediction model	2	0.92	0.91-0.94	0.91	0.94
PSI	7	0.75	0.69-0.81	0.63	0.83
Procalcitonin	2	0.73	0.65-0.81	0.65	0.81
RIFLE classification	3	0.75	0.70-0.91	0.70	0.91
Ranson's criteria	2	0.89	0.82-0.95	0.82	0.95
SAPS II	8	0.77	0.73-0.82	0.51	0.85
SAPS III	3	0.74	0.71-0.84	0.71	0.84
SOFA score	9	0.84	0.75-0.85	0.71	0.93
Simple risk index	2	0.80	0.78-0.82	0.78	0.82

TIMI risk score	5	0.73	0.72-0.75	0.68	0.84
TIMI-risk score & Labor index	2	0.77	0.76-0.78	0.76	0.78
TNM	2	0.80	NA	0.80	0.80
TRISS	2	0.75	0.64-0.85	0.64	0.85
Tokyo score	2	0.87	0.86-0.87	0.86	0.87

AUC, area under the curve; IQR, interquartile range; Min, minimum; Max, maximum; NA, not applicable; AMIS, Acute Myocardial Infarction in Switzerland; APACHE II, Acute Physiology And Chronic Health Evaluation II; BCLC, Barcelona Clinic Liver Cancer; BISAP, Bedside Index for Severity in Acute Pancreatitis; BNP, B-type natriuretic peptide; CLIP, Cancer of the Liver Italian Program; CRIB, Clinical Risk Index for Babies; CTP, Child–Turcotte–Pugh; CCI, Charlson Comorbidity Index; EuroSCORE, European system for cardiac operative risk evaluation; ISS, Injury Severity Score; JIS, Japan Integrated Staging; MELD, Model for End-Stage Liver Disease; MPI, Multidimensional Prognostic Index; MPM, Mortality Probability Models; NT-proBNP, N-terminal-pro-B-type natriuretic peptide; PSI, Pneumonia Severity Index; RIFLE, Risk of renal failure, Injury to the kidney, Failure of kidney function, Loss of kidney function, and End-stage renal disease; SAPS, Simplified Acute Physiology Score; SOFA, Sequential Organ Failure Assessment; TIMI, Thrombolysis In Myocardial Infarction; TRISS, Trauma Revised Injury Severity Score.

Accuracy

The AUC values ranged from 0.43 to 0.98 (**Figure 2.2**), and the median (IQR) AUC value was 0.77 (0.71-0.83). A total of 95 of the AUC values were higher than 0.80 (very good discrimination) (40%), but only 23 were higher than 0.90 (excellent discrimination) (10%).

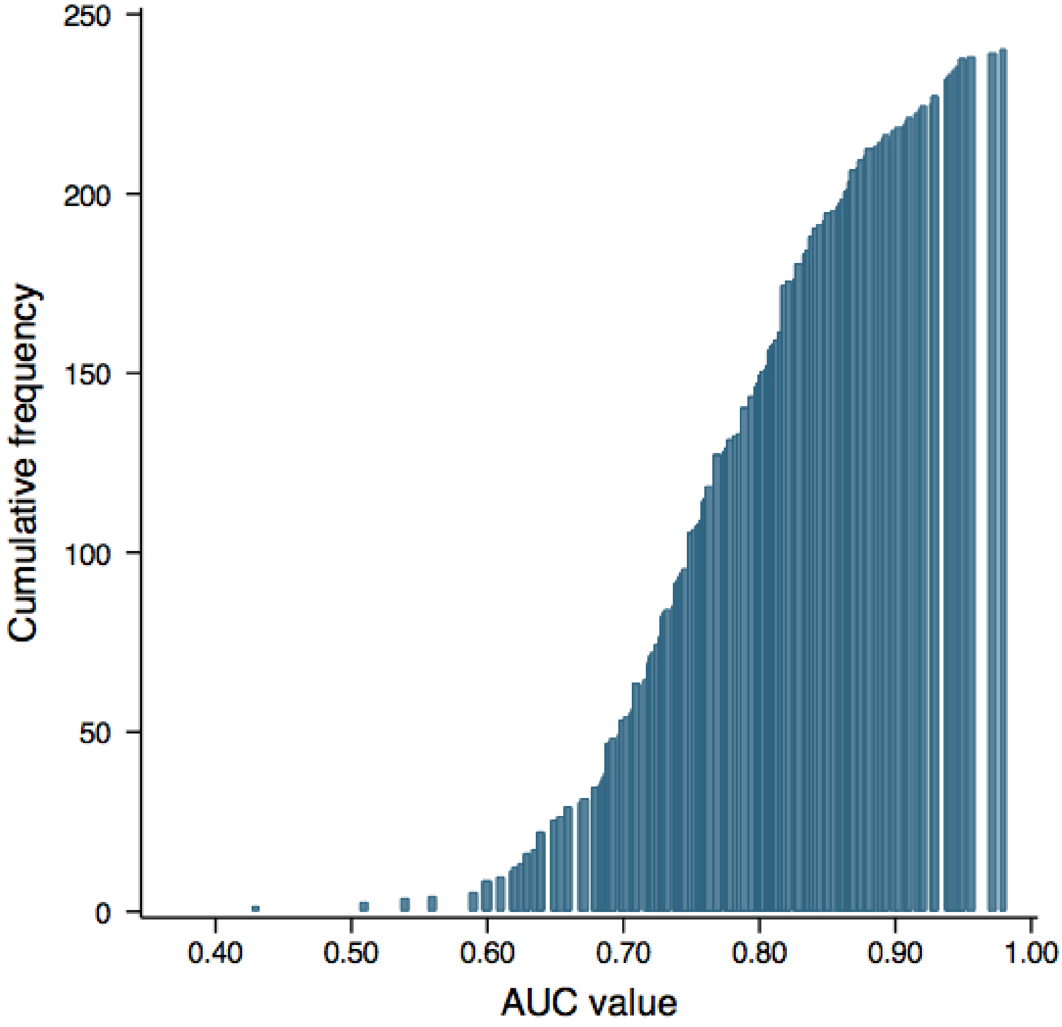

Figure 2.2: Cumulative frequency histogram of area under the receiver operating characteristic curve (AUC) values for mortality.

The AUC data for all predictive tools with 2 or more assessments are listed in (**Table 2.2**). For each of these 34 tools, the range of AUC estimates was large, sometimes spanning the spectrum from inaccurate to excellent accuracy. The median AUC values suggested modest accuracy. For only 2 predictive tools (Clinical Risk Index for Babies [CRIB] II (S25 and S27 in **2.5.1 Supplementary references**) and Pediatric death prediction model (S92 in **2.5.1 Supplementary references**)), the median AUC value suggested excellent accuracy (AUC, 0.91 and 0.92, respectively), but this was based on only 2 assessments of each tool. Four or more assessments of the accuracy of a predictive tool were available for only 9 tools (APACHE, MELD, SOFA [Sequential Organ Failure Assessment], CTP [Child-Turcotte-Pugh], SAPS [Simplified Acute Physiology Score] II, PSI, CLIP [Cancer of the Liver Italian Program], CURB-65 [confusion–blood urea nitrogen–respiratory rate–blood pressure–age >65 years], JIS [Japan Integrated Staging]) and 1 biomarker (NT-pro-BNP [N-terminal-pro-B-type natriuretic peptide]). Using random effects meta-analysis, we found that the summary AUC estimates for these 10 tools Ca (**Figure 2.3**). For each of the 9 multivariable tools, there was marked heterogeneity of AUC values across diverse settings and studies (heterogeneity I^2 estimates in AUC ranged from 68% to 95%). The 95% confidence intervals of the I^2 were also consistent with a large or very large heterogeneity. For NT-pro-BNP, the I^2 estimate was 25%. Meta-analyses retaining only the longest follow-up assessment when several follow-up assessments were available from the same study showed similar results (all changes in summary AUC estimates were <5% compared with the primary analysis including all data).

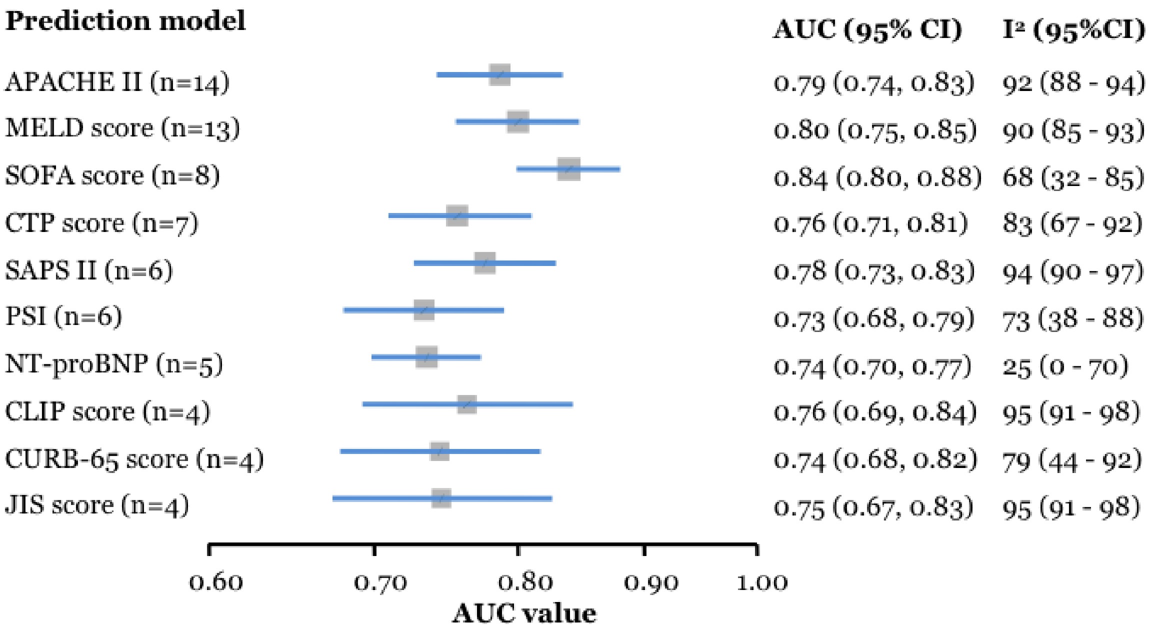

Figure 2.3: Area under the receiver operating characteristic curve (AUC) values for predictive tools that were examined in 4 or more assessments (n=number of assessments) with 95% confidence intervals (CIs). Summary results of AUC and 95% CIs are provided using random effects meta-analysis.

Calibration and Reclassification

Calibration of the examined predictive tools was examined in fewer than half of the included studies (n=45; 48%), mainly by using the Hosmer-Lemeshow statistic (n=35; 78%) and observed/predicted ratio (n=5; 11%). Results were available in 44 studies (105 predictive tool assessments), indicating lack of fit for 8 studies (17 predictive tools). Only 1 study (S83 in **2.5.1 Supplementary references**) examined reclassification analysis by means of the net reclassification improvement and the integrated discrimination index. This study investigated the added predictive value of radiographic ascites over and above the MELD-Na score in patients with cirrhosis.

Correlates of Accuracy

As listed in **Table 2.3**, predictive tools published in journals of lower impact factor had higher reported AUC estimates than those published in journals of higher impact factor. Predictive tools were more accurate in predicting mortality when a smaller proportion of study participants died. The AUC values were also higher in pediatric than in adult populations. Finally, studies with larger sample size tended to have higher AUC values than smaller studies. There was no evidence that study design (retrospective vs. prospective), area of origin, disease status, clinical condition examined, death rate per month, loss to follow- up, or number of variables included in the predictive tool were associated with the AUC values.

Table 2.3: Association between AUC values and study characteristics

Study characteristics	All predictive tools			
	No.[a]	Mean	SD	*p* value[b]
Journal impact factor	222			0.021
≤2.13	46	0.78	0.11	
2.13 – 2.32	45	0.79	0.07	
2.32 – 3.15	45	0.78	0.08	
3.15 – 5.39	43	0.77	0.07	
≥5.39	43	0.75	0.10	
Study population	240			<0.0001
Pediatric	7	0.92	0.02	
Adult	225	0.77	0.09	
Both	8	0.78	0.04	
Sample size	240			0.014
≤147	48	0.76	0.11	
147 – 287	49	0.76	0.11	
287 – 810	48	0.76	0.08	
810 – 2558	48	0.80	0.09	
≥2558	47	0.79	0.08	
Proportion of study participants who died	238			0.002
≤0.06	49	0.82	0.08	
0.06 – 0.13	47	0.76	0.10	
0.13 – 0.21	46	0.78	0.10	
0.21 – 0.33	50	0.78	0.06	
≥0.30	46	0.73	0.10	

Abbreviations: AUC, area under the curve; SD, standard deviation; USA, United States of America; NA, not applicable.

[a] number of the predictive tools related to the respective extracted variable.

[b] one-way ANOVA for categorical variables (study population) and Spearman correlation coefficient for continuous variables (impact factor, sample size, proportion of death).

2.4 DISCUSSION

This systematic evaluation of a large number of seemingly well-validated predictive tools reported in the recent literature shows that these tools are not very accurate and that there is wide variation in their predictive accuracy for death. Most of the tools included into this analysis are not sufficiently accurate for wide use in clinical practice. Moreover, calibration was assessed in fewer than half of the tools, and of those tested, several showed lack of fit, meaning that prediction was not equally good for patients at different levels of risk. Studies published in journals with lower impact factor tended to show better AUC values, while tools performed better when they tried to predict death only for the highest-risk patients.

For a proposed predictive tool to be useful in clinical practice, there are several prerequisites. The tool must be validated in populations other than the one in which was developed; it should be reproducible; and it should have good accuracy and calibration. Such a predictive tool can make accurate predictions in diverse settings across the range of both low- and high-risk patients. Few tools for predicting risk of death currently fit these criteria. Even tools that meet these criteria may not necessarily result in improvement in patient management and outcomes. This depends on whether effective, feasible interventions are available, the use of which is based on accurate knowledge of patient risk. However, reclassification, the ability to reclassify individuals into more appropriate risk categories where different actions/interventions might be indicated, is almost never assessed in the current literature of death prediction. Moreover, randomized trials on the use of predictive models, the ultimate proof of benefit, are few and difficult to conduct. Finally, clinicians are unlikely to use complex tools that require collection of extensive information, including data derived from expensive tests. It is possible that other predictive tools, based on far more limited clinical data, may perform equally well or better. In our empirical evaluation, models with more variables did not seem to perform clearly better than models with few variables.

Some characteristics of predictive tools were significantly associated with higher AUC estimates. For example, tools performed better when they tried to predict death only for the highest-risk patients. Excellent performance was seen in a small number of pediatric tools, while performance was substantially worse in predictive tools for adults. Larger studies tended to have slightly higher AUC estimates. These associations are exploratory and should be viewed with caution.

In this broad evaluation we focused on validated tools. However, even for some of the most widely applied predictive tools (such as APACHE II, MELD score, and SAPS II), we found great within-tool variability in accuracy across different studies and clinical settings. The observed variation of the accuracy for the same predictive tool may be partly ascribed to the selective analysis and reporting of studies of predictive tools that may lead to exaggerated results of predictive discrimination in some studies. Efforts at standardization of reporting are important in this regard.[172,173] The inverse correlation between journal impact factor and reported AUC that we observed may represent lower methodologic quality with spuriously high reported predictive performance in some articles published in journals with low impact factor.[174] Moreover, studies often test predictive tools in populations that are very different than the one the model was developed for and for a wide range of outcomes. This may further contribute to the variability seen in their discriminatory performance.

Some limitations should be mentioned. This empirical assessment was restricted to studies published during a single year. An effort to appraise the entire predictive literature would be a task requiring extensive international effort by hundreds of researchers, much as the Cochrane Collaboration has done for clinical trials. Moreover, we included only studies dealing with prediction of all-cause death, and we did not evaluate the accuracy of tools designed to predict other outcomes. However, death from any cause is a common outcome with great clinical impact, and it is possible to standardize unambiguously. Finally, we considered only predictive studies that assessed accuracy using the AUC. However, AUC is not the only metric to assess predictive ability,[124] and like any single metric, it can have limitations.[125,147,175,176] For example, the AUC does not provide information on the actual predicted probabilities, and it does not convey the exact risk distribution in the respective study population. Also, improvements in AUC are more difficult in the high-range values than when AUC is closer to 0.50.[164] Nevertheless, AUC is a very useful metric[125,147] and is the most widely used standardized metric in the predictive literature.

Conclusions

Given the very wide variability in the AUC, even for the same predictive tool, we believe that systematic efforts are needed to organize and synthesize the predictive literature, such as those proposed by the Cochrane Prognosis Methods Group. Such efforts are needed to enhance the evidence derived from predictive research and to establish standard methods for developing, evaluating, reporting, and eventually adopting new predictive tools in clinical practice. Clinicians should be cautious about adopting new, initially promising predictive tools, especially complex ones based on expensive measurements that have not been extensively validated and shown to be consistently useful in practice.

2.5 SUPPLEMENTARY MATERIAL

2.5.1 Supplementary references

S1. Trujillano J, Badia M, Serviá L, March J, Rodriguez-Pozo A. Stratification of the severity of critically ill patients with classification trees. BMC Med Res Methodol. 2009;9:83. [PMID: 20003229]

S2. España PP, Capelastegui A, Quintana JM, Bilbao A, Diez R, Pascual S, Esteban C, Zalacaín R, Menendez R, Torres A. Validation and comparison of SCAP as a predictive score for identifying low-risk patients in community-acquired pneumonia. J Infect. 2010;60(2):106-13. [PMID: 19961875]

S3. Durga P, Sahu BP, Mantha S, Ramachandran G. Development and validation of predictors of respiratory insufficiency and mortality scores: simple bedside additive scores for prediction of ventilation and in-hospital mortality in acute cervical spine injury. Anesth Analg. 2010;110(1):134-40. [PMID: 19933524]

S4. Kim JD, Choi JY, Kwon JH, Jang JW, Bae SH, Yoon SK, You YK, Kim DG. Performance of posttransplant Model for End-Stage Liver Disease (MELD) and delta-MELD scores on short-term outcome after living donor liver transplantation. Transplant Proc. 2009;41(9):3766-8. [PMID: 19917383]

S5. Hao K, Luk JM, Lee NP, Mao M, Zhang C, Ferguson MD, Lamb J, Dai H, Ng IO, Sham PC, Poon RT. Predicting prognosis in hepatocellular carcinoma after curative surgery with common clinicopathologic parameters. BMC Cancer. 2009;9:389. [PMID: 19886989]

S6. Papachristou GI, Muddana V, Yadav D, O'Connell M, Sanders MK, Slivka A, Whitcomb DC. Comparison of BISAP, Ranson's, APACHE-II, and CTSI scores in predicting organ failure, complications, and mortality in acute pancreatitis. Am J Gastroenterol. 2010;105(2):435-41. [PMID: 19861954]

S7. Raszeja-Wyszomirska J, Wasilewicz MP, Wunsch E, Szymanik B, Jarosz K, Wójcicki M, Milkiewicz P. Assessment of a modified Child-Pugh-Turcotte score to predict early mortality after liver transplantation. Transplant Proc. 2009;41(8):3114-6. [PMID: 19857689]

S8. Namendys-Silva SA, Baltazar-Torres JA, Rivero-Sigarroa E, Fonseca-Lazcano JA, Montiel-López L, Domínguez-Cherit G. Prognostic factors in patients with systemic lupus erythematosus admitted to the intensive care unit. Lupus. 2009;18(14):1252-8. [PMID: 19850662]

S9. Cholongitas E, Calvaruso V, Senzolo M, Patch D, Shaw S, O'Beirne J, Burroughs AK. RIFLE classification as predictive factor of mortality in patients with cirrhosis admitted to intensive care unit. J Gastroenterol Hepatol. 2009;24(10):1639-47. [PMID: 19788604]

S10. Tabak YP, Sun X, Johannes RS, Gupta V, Shorr AF. Mortality and need for mechanical ventilation in acute exacerbations of chronic obstructive pulmonary disease: development and validation of a simple risk score. Arch Intern Med. 2009;169(17):1595-602. [PMID: 19786679]

S11. Francia E, Torres O, Laiz A, Ruiz D, Gich I, Casademont J. Ability of physiological parameters versus clinical categories to predict mortality on admission to an internal medicine ward. Eur J Intern Med. 2009;20(6):636-9. [PMID: 19782928]

S12. Gotthardt D, Weiss KH, Baumgärtner M, Zahn A, Stremmel W, Schmidt J, Bruckner T, Sauer P. Limitations of the MELD score in predicting mortality or need for removal from waiting list in patients awaiting liver transplantation. BMC Gastroenterol. 2009;9:72. [PMID: 19778459]

S13. Khwannimit B, Bhurayanontachai R. The performance and customization of SAPS 3 admission score in a Thai medical intensive care unit. Intensive Care Med. 2010;36(2):342-6. [PMID: 19756506]

S14. Hariharan S, Chen D, Ramkissoon A, Taklalsingh N, Bodkyn C, Cupidore R, Ramdin A, Ramsaroop A, Sinanan V, Teelucksingh S, Verma S. Perioperative outcome of colorectal cancer and validation of CR-POSSUM in a Caribbean country. Int J Surg. 2009;7(6):534-8. [PMID: 19737634]

S15. Haase M, Bellomo R, Matalanis G, Calzavacca P, Dragun D, Haase-Fielitz A. A comparison of the RIFLE and Acute Kidney Injury Network classifications for cardiac surgery-associated acute kidney injury: a prospective cohort study. J Thorac Cardiovasc Surg. 2009;138(6):1370-6. [PMID: 19733864]

S16. Stauber RE, Wagner D, Stadlbauer V, Palma S, Gurakuqi G, Kniepeiss D, Iberer F, Smolle KH, Haas J, Trauner M. Evaluation of indocyanine green clearance and model for end-stage liver disease for estimation of short-term prognosis in decompensated cirrhosis. Liver Int. 2009;29(10):1516-20. [PMID: 19732329]

S17. Hampshire PA, Welch CA, McCrossan LA, Francis K, Harrison DA. Admission factors associated with hospital mortality in patients with haematological malignancy admitted to UK adult, general critical care units: a secondary analysis of the ICNARC Case Mix Programme Database. Crit Care. 2009;13(4):R137. [PMID: 19706163]

S18. Parsonage M, Nathwani D, Davey P, Barlow G. Evaluation of the performance of CURB-65 with increasing age. Clin Microbiol Infect. 2009;15(9):858-64. [PMID: 19702590]

S19. Lv XH, Liu HB, Wang Y, Wang BY, Song M, Sun MJ. Validation of model for end-stage liver disease score to serum sodium ratio index as a prognostic predictor in patients with cirrhosis. J Gastroenterol Hepatol. 2009;24(9):1547-53. [PMID: 19686416]
S20. Hamburger JN, Walsh SJ, Khurana R, Ding L, Gao M, Humphries KH, Carere R, Fung AY, Mildenberger RR, Simkus GJ, Webb JG, Buller CE. Percutaneous coronary intervention and 30-day mortality: the British Columbia PCI risk score. Catheter Cardiovasc Interv. 2009;74(3):377-85. [PMID: 19681116]

S21. Reid C, Billah B, Dinh D, Smith J, Skillington P, Yii M, Seevanayagam S, Mohajeri M, Shardey G. An Australian risk prediction model for 30-day mortality after isolated coronary artery bypass: the AusSCORE. J Thorac Cardiovasc Surg. 2009;138(4):904-10. [PMID: 19660369]

S22. Jalan R, Schnurr K, Mookerjee RP, Sen S, Cheshire L, Hodges S, Muravsky V, Williams R, Matthes G, Davies NA. Alterations in the functional capacity of albumin in patients with decompensated cirrhosis is associated with increased mortality. Hepatology. 2009;50(2):555-64. [PMID: 19642174]

S23. Keegan MT, Gali B, Findlay JY, Heimbach JK, Plevak DJ, Afessa B. APACHE III outcome prediction in patients admitted to the intensive care unit after liver transplantation: a retrospective cohort study. BMC Surg. 2009;9:11. [PMID: 19640303]

S24. Potocki M, Breidthardt T, Reichlin T, Morgenthaler NG, Bergmann A, Noveanu M, Schaub N, Uthoff H, Freidank H, Buser L, Bingisser R, Christ M, Mebazaa A, Mueller C.

Midregional pro-adrenomedullin in addition to b-type natriuretic peptides in the risk stratification of patients with acute dyspnea: an observational study. Crit Care. 2009;13(4):R122. [PMID: 19627611]

S25. Manktelow BN, Draper ES, Field DJ. Predicting neonatal mortality among very preterm infants: a comparison of three versions of the CRIB score. Arch Dis Child Fetal Neonatal Ed. 2010;95(1):F9-F13. [PMID: 19608556]

S26. Mbongo CL, Monedero P, Guillen-Grima F, Yepes MJ, Vives M, Echarri G. Performance of SAPS3, compared with APACHE II and SOFA, to predict hospital mortality in a general ICU in Southern Europe. Eur J Anaesthesiol. 2009;26(11):940-5. [PMID: 19606046]

S27. Rastogi PK, Sreenivas V, Kumar N. Validation of CRIB II for prediction of mortality in premature babies. Indian Pediatr. 2010;47(2):145-7. [PMID: 19578231]

S28. Jensen JK, Atar D, Kristensen SR, Mickley H, Januzzi JL Jr. Usefulness of natriuretic peptide testing for long-term risk assessment following acute ischemic stroke. Am J Cardiol. 2009;104(2):287-91. [PMID: 19576362]

S29. Park SK, Chun HJ, Kim DW, Im TH, Hong HJ, Yi HJ. Acute Physiology and Chronic Health Evaluation II and Simplified Acute Physiology Score II in predicting hospital mortality of neurosurgical intensive care unit patients. J Korean Med Sci. 2009;24(3):420-6. [PMID: 19543503]

S30. Dossett LA, Redhage LA, Sawyer RG, May AK. Revisiting the validity of APACHE II in the trauma ICU: improved risk stratification in critically injured adults. Injury. 2009;40(9):993-8. [PMID: 19535054]

S31. Brusselaers N, Juhász I, Erdei I, Monstrey S, Blot S. Evaluation of mortality following severe burns injury in Hungary: external validation of a prediction model developed on Belgian burn data. Burns. 2009;35(7):1009-14. [PMID: 19501970]

S32. Horne BD, May HT, Muhlestein JB, Ronnow BS, Lappé DL, Renlund DG, Kfoury AG, Carlquist JF, Fisher PW, Pearson RR, Bair TL, Anderson JL. Exceptional mortality prediction by risk scores from common laboratory tests. Am J Med. 2009;122(6):550-8. [PMID: 19486718]

S33. Mirsaeidi M, Peyrani P, Ramirez JA; Improving Medicine through Pathway Assessment of Critical Therapy of Hospital-Acquired Pneumonia (IMPACT-HAP) Investigators. Predicting mortality in patients with ventilator-associated pneumonia: The APACHE II score versus the new IBMP-10 score. Clin Infect Dis. 2009;49(1):72-7. [PMID: 19480582]

S34. Moore L, Lavoie A, Turgeon AF, Abdous B, Le Sage N, Emond M, Liberman M, Bergeron E. The trauma risk adjustment model: a new model for evaluating trauma care. Ann Surg. 2009;249(6):1040-6. [PMID: 19474674]

S35. Carosella VC, Navia JL, Al-Ruzzeh S, Grancelli H, Rodriguez W, Cardenas C, Bilbao J, Nojek C. The first Latin-American risk stratification system for cardiac surgery: can be used as a graphic pocket-card score. Interact Cardiovasc Thorac Surg. 2009;9(2):203-8. [PMID: 19454412]

S36. Adrie C, Francais A, Alvarez-Gonzalez A, Mounier R, Azoulay E, Zahar JR, Clec'h C, Goldgran-Toledano D, Hammer L, Descorps-Declere A, Jamali S, Timsit JF; Outcomerea Study Group. Model for predicting short-term mortality of severe sepsis. Crit Care. 2009;13(3):R72. [PMID: 19454002]

S37. Strand K, Søreide E, Aardal S, Flaatten H. A comparison of SAPS II and SAPS 3 in a Norwegian intensive care unit population. Acta Anaesthesiol Scand. 2009;53(5):595-600. [PMID: 19419352]

S38. Sun QF, Ding JG, Xu DZ, Chen YP, Hong L, Ye ZY, Zheng MH, Fu RQ, Wu JG, Du QW, Chen W, Wang XF, Sheng JF. Prediction of the prognosis of patients with acute-on-chronic hepatitis B liver failure using the model for end-stage liver disease scoring system and a novel logistic regression model. J Viral Hepat. 2009;16(7):464-70. [PMID: 19413694]

S39. Visser JJ, Williams M, Kievit J, Bosch JL; 4-A Study Group. Prediction of 30-day mortality after endovascular repair or open surgery in patients with ruptured abdominal aortic aneurysms. J Vasc Surg. 2009;49(5):1093-9. [PMID: 19394540]

S40. Raum MR, Nijsten MW, Vogelzang M, Schuring F, Lefering R, Bouillon B, Rixen D, Neugebauer EA, Ten Duis HJ; Polytrauma Study Group of the German Trauma Society. Emergency trauma score: an instrument for early estimation of trauma severity. Crit Care Med. 2009;37(6):1972-7. [PMID: 19384227]

S41. Noda T, Sasaki Y, Yamada T, Eguchi H, Yano M, Ohigashi H, Ishikawa O, Imaoka S. Usefulness of the CLIP scoring system for prediction of postoperative prognosis of patients with large hepatocellular carcinoma. J Hepatobiliary Pancreat Surg. 2009;16(4):538-45. [PMID: 19367360]

S42. Huang DT, Angus DC, Kellum JA, Pugh NA, Weissfeld LA, Struck J, Delude RL, Rosengart MR, Yealy DM. Midregional proadrenomedullin as a prognostic tool in community-acquired pneumonia. Chest. 2009;136(3):823-31. [PMID: 19363212]

S43. Garcia PC, Eulmesekian P, Branco RG, Perez A, Sffogia A, Olivero L, Piva JP, Tasker RC. External validation of the paediatric logistic organ dysfunction score. Intensive Care Med. 2010;36(1):116-22. [PMID: 19360395]

S44. Callaway DW, Shapiro NI, Donnino MW, Baker C, Rosen CL. Serum lactate and base deficit as predictors of mortality in normotensive elderly blunt trauma patients.J Trauma. 2009;66(4):1040-4. [PMID: 19359912]

S45. Ranucci M, Castelvecchio S, Menicanti LA, Scolletta S, Biagioli B, Giomarelli P. An adjusted EuroSCORE model for high-risk cardiac patients. Eur J Cardiothorac Surg. 2009;36(5):791-7. [PMID: 19359191]

S46. Namendys-Silva SA, Texcocano-Becerra J, Herrera-Gómez A. Application of the Sequential Organ Failure Assessment (SOFA) score to patients with cancer admitted to the intensive care unit. Am J Hosp Palliat Care. 2009;26(5):341-6. [PMID: 19357377]

S47. Pilotto A, Addante F, Ferrucci L, Leandro G, D'Onofrio G, Corritore M, Niro V, Scarcelli C, Dallapiccola B, Franceschi M. The multidimensional prognostic index predicts short- and long-term mortality in hospitalized geriatric patients with pneumonia. J Gerontol A Biol Sci Med Sci. 2009;64(8):880-7. [PMID: 19349589]

S48. Voors AA, von Haehling S, Anker SD, Hillege HL, Struck J, Hartmann O, Bergmann A, Squire I, van Veldhuisen DJ, Dickstein K; OPTIMAAL Investigators. C-terminal provasopressin (copeptin) is a strong prognostic marker in patients with heart failure after an acute myocardial infarction: results from the OPTIMAAL study. Eur Heart J. 2009;30(10):1187-94. [PMID: 19346228]

S49. Barbieri A, Pinna C, Basso GP, Molinari R, Giuliani E, Fruggeri L, Nolli M. Specificity and reliability of prognostic indexes in intensive care evaluation: the spontaneous cerebral haemorrhage case. J Eval Clin Pract. 2009;15(2):242-5. [PMID: 19335479]

S50. Inal MT, Memiş D, Kargi M, Sut N. Prognostic value of indocyanine green elimination assessed with LiMON in septic patients. J Crit Care. 2009;24(3):329-34. [PMID: 19327336]

S51. Juneja D, Gopal PB, Kapoor D, Raya R, Sathyanarayanan M, Malhotra P. Outcome of patients with liver cirrhosis admitted to a specialty liver intensive care unit in India. J Crit Care. 2009;24(3):387-93. [PMID: 19327335]

S52. Jones AE, Trzeciak S, Kline JA. The Sequential Organ Failure Assessment score for predicting outcome in patients with severe sepsis and evidence of hypoperfusion at the time of emergency department presentation. Crit Care Med. 2009;37(5):1649-54. [PMID: 19325482]

S53. Ueda T, Takeyama Y, Yasuda T, Kamei K, Satoi S, Sawa H, Shinzeki M, Ku Y, Kuroda Y, Ohyanagi H. Utility of the new Japanese severity score and indications for special therapies in acute pancreatitis. J Gastroenterol. 2009;44(5):453-9. [PMID: 19308309]

S54. Singh VK, Wu BU, Bollen TL, Repas K, Maurer R, Johannes RS, Mortele KJ, Conwell DL, Banks PA. A prospective evaluation of the bedside index for severity in acute pancreatitis score in assessing mortality and intermediate markers of severity in acute pancreatitis. Am J Gastroenterol. 2009;104(4):966-71. [PMID: 19293787]

S55. Dominguez-Rodriguez A, Abreu-Gonzalez P, Jimenez-Sosa A, Samimi-Fard S, Idaira HB. Does ischemia-modified albumin add prognostic value to the Thrombolysis In Myocardial Infarction risk score in patients with ST-segment elevation myocardial infarction treated with primary angioplasty? Biomarkers. 2009;14(1):43-8. [PMID: 19283523]

S56. Gummert JF, Funkat A, Osswald B, Beckmann A, Schiller W, Krian A, Beyersdorf F, Haverich A, Cremer J. EuroSCORE overestimates the risk of cardiac surgery: results from the national registry of the German Society of Thoracic and Cardiovascular Surgery. Clin Res Cardiol. 2009;98(6):363-9. [PMID: 19262978]

S57. Huo TI, Hsia CY, Huang YH, Lin HC, Lee PC, Lui WY, Chiang JH, Chiou YY, Loong CC, Lee SD. Selecting a short-term prognostic model for hepatocellular carcinoma: comparison between the model for end-stage liver disease (MELD), MELD-sodium, and five cancer staging systems. J Clin Gastroenterol. 2009;43(8):773-81. [PMID: 19262404]

S58. Giles KA, Schermerhorn ML, O'Malley AJ, Cotterill P, Jhaveri A, Pomposelli FB, Landon BE. Risk prediction for perioperative mortality of endovascular vs open repair of abdominal aortic aneurysms using the Medicare population. J Vasc Surg. 2009;50(2):256-62. [PMID: 19249184]

S59. Capelastegui A, España PP, Quintana JM, Bilbao A, Menendez R, Zalacain R, Torres A. Development of a prognostic index for 90-day mortality in patients discharged after admission to hospital for community-acquired pneumonia. Thorax. 2009;64(6):496-501. [PMID: 19237392]

S60. Boursier J, Cesbron E, Tropet AL, Pilette C. Comparison and improvement of MELD and Child-Pugh score accuracies for the prediction of 6-month mortality in cirrhotic patients. J Clin Gastroenterol. 2009;43(6):580-5. [PMID: 19197195]

S61. Leaphart CL, Graham D, Pieper P, Celso BG, Tepas JJ 3rd. Surgical quality improvement: a simplified method to apply national standards to pediatric trauma care. J Pediatr Surg. 2009;44(1):156-9. [PMID: 19159735]

S62. Frankenstein L, Clark AL, Goode K, Ingle L, Remppis A, Schellberg D, Grabs F, Nelles M, Cleland JG, Katus HA, Zugck C. The prognostic value of individual NT-proBNP values in chronic heart failure does not change with advancing age. Heart. 2009;95(10):825-9. [PMID: 19147626]

S63. Menéndez R, Martínez R, Reyes S, Mensa J, Filella X, Marcos MA, Martínez A, Esquinas C, Ramirez P, Torres A. Biomarkers improve mortality prediction by prognostic scales in community-acquired pneumonia. Thorax. 2009;64(7):587-91. [PMID: 19131448]

S64. Rello J, Rodriguez A, Lisboa T, Gallego M, Lujan M, Wunderink R. PIRO score for community-acquired pneumonia: a new prediction rule for assessment of severity in intensive care unit patients with community-acquired pneumonia. Crit Care Med. 2009;37(2):456-62. [PMID: 19114916]

S65. Belgian Outcome in Burn Injury Study Group. Development and validation of a model for prediction of mortality in patients with acute burn injury. Br J Surg. 2009;96(1):111-7. [PMID: 19109825]

S66. Dehing-Oberije C, Yu S, De Ruysscher D, Meersschout S, Van Beek K, Lievens Y, Van Meerbeeck J, De Neve W, Rao B, van der Weide H, Lambin P. Development and external validation of prognostic model for 2-year survival of non-small-cell lung cancer patients treated with chemoradiotherapy. Int J Radiat Oncol Biol Phys. 2009;74(2):355-62. [PMID: 19095367]

S67. Lin CY, Kao KC, Tian YC, Jenq CC, Chang MY, Chen YC, Fang JT, Huang CC, Tsai YH, Yang CW. The RIFLE score increases the accuracy of outcome prediction in patients with acute respiratory distress syndrome undergoing open lung biopsy. Respiration. 2009;77(4):398-406. [PMID: 19077383]

S68. Ruiz-Alvarez MJ, García-Valdecasas S, De Pablo R, Sanchez García M, Coca C, Groeneveld TW, Roos A, Daha MR, Arribas I. Diagnostic efficacy and prognostic value of serum procalcitonin concentration in patients with suspected sepsis. J Intensive Care Med. 2009;24(1):63-71. [PMID: 19054806]

S69. Limquiaco JL, Wong GL, Wong VW, Lai PB, Chan HL. Evaluation of model for end stage liver disease (MELD)-based systems as prognostic index for hepatocellular carcinoma. J Gastroenterol Hepatol. 2009;24(1):63-9. [PMID: 19054256]

S70. Naveau S, Gaudé G, Asnacios A, Agostini H, Abella A, Barri-Ova N, Dauvois B, Prévot S, Ngo Y, Munteanu M, Balian A, Njiké-Nakseu M, Perlemuter G, Poynard T. Diagnostic and prognostic values of noninvasive biomarkers of fibrosis in patients with alcoholic liver disease. Hepatology. 2009;49(1):97-105. [PMID: 19053048]

S71. Schuetz P, Müller B, Nusbaumer C, Wieland M, Christ-Crain M. Circulating levels of GH predict mortality and complement prognostic scores in critically ill medical patients. Eur J Endocrinol. 2009;160(2):157-63. [PMID: 19022915]

S72. Chiu PW, Ng EK, Cheung FK, Chan FK, Leung WK, Wu JC, Wong VW, Yung MY, Tsoi K, Lau JY, Sung JJ, Chung SS. Predicting mortality in patients with bleeding peptic ulcers after therapeutic endoscopy. Clin Gastroenterol Hepatol. 2009;7(3):311-6; quiz 253. [PMID: 18955161]

S73. Kotanidou A, Karsaliakos P, Tzanela M, Mavrou I, Kopterides P, Papadomichelakis E, Theodorakopoulou M, Botoula E, Tsangaris I, Lignos M, Ikonomidis I, Ilias I, Armaganidis A, Orfanos SE, Dimopoulou I. Prognostic importance of increased plasma amino-terminal pro-brain natriuretic peptide levels in a large noncardiac, general intensive care unit population. Shock. 2009;31(4):342-7. [PMID: 18791494]

S74. Zheng Z, Li Y, Zhang S, Hu S; Chinese CABG Registry Study. The Chinese coronary artery bypass grafting registry study: how well does the EuroSCORE predict operative risk for Chinese population? Eur J Cardiothorac Surg. 2009;35(1):54-8. [PMID: 18778949]

S75. Peterson PN, Rumsfeld JS, Liang L, Albert NM, Hernandez AF, Peterson ED, Fonarow GC, Masoudi FA; American Heart Association Get With the Guidelines-Heart Failure Program. A validated risk score for in-hospital mortality in patients with heart failure from the American Heart Association get with the guidelines program. Circ Cardiovasc Qual Outcomes. 2010;3(1):25-32. [PMID: 20123668]

S76. Barba P, Piñana JL, Martino R, Valcárcel D, Amorós A, Sureda A, Briones J, Delgado J, Brunet S, Sierra J. Comparison of two pretransplant predictive models and a flexible HCT-CI using different cut off points to determine low-, intermediate-, and high-risk groups: the flexible HCT-CI Is the best predictor of NRM and OS in a population of patients undergoing allo-RIC. Biol Blood Marrow Transplant. 2010;16(3):413-20. [PMID: 19922807]

S77. Hernández D, Sánchez-Fructuoso A, González-Posada JM, Arias M, Campistol JM, Rufino M, Morales JM, Moreso F, Pérez G, Torres A, Serón D; Spanish Late Allograft Dysfunction Study Group. A novel risk score for mortality in renal transplant recipients beyond the first posttransplant year. Transplantation. 2009;88(6):803-9. [PMID: 19920780]

S78. O'Brien SM, Clarke DR, Jacobs JP, Jacobs ML, Lacour-Gayet FG, Pizarro C, Welke KF, Maruszewski B, Tobota Z, Miller WJ, Hamilton L, Peterson ED, Mavroudis C, Edwards FH. An empirically based tool for analyzing mortality associated with congenital heart surgery. J Thorac Cardiovasc Surg. 2009;138(5):1139-53. [PMID: 19837218]

S79. Justice AC, McGinnis KA, Skanderson M, Chang CC, Gibert CL, Goetz MB, Rimland D, Rodriguez-Barradas MC, Oursler KK, Brown ST, Braithwaite RS, May M, Covinsky KE, Roberts MS, Fultz SL, Bryant KJ; VACS Project Team. Towards a combined prognostic index for survival in HIV infection: the role of 'non-HIV' biomarkers. HIV Med. 2010;11(2):143-51. [PMID: 19751364]

S80. Thiem U, Niklaus D, Sehlhoff B, Stückle C, Heppner HJ, Endres HG, Pientka L. C-reactive protein, severity of pneumonia and mortality in elderly, hospitalised patients with community-acquired pneumonia. Age Ageing. 2009;38(6):693-7. [PMID: 19729453]

S81. Elbarouni B, Goodman SG, Yan RT, Welsh RC, Kornder JM, Deyoung JP, Wong GC, Rose B, Grondin FR, Gallo R, Tan M, Casanova A, Eagle KA, Yan AT; Canadian Global Registry of Acute Coronary Events (GRACE/GRACE(2)) Investigators. Validation of the Global Registry of Acute Coronary Event (GRACE) risk score for in-hospital mortality in patients with acute coronary syndrome in Canada. Am Heart J. 2009;158(3):392-9. [PMID: 19699862]

S82. Quach S, Hennessy DA, Faris P, Fong A, Quan H, Doig C. A comparison between the APACHE II and Charlson Index Score for predicting hospital mortality in critically ill patients. BMC Health Serv Res. 2009;9:129. [PMID: 19643010]

S83. Somsouk M, Guy J, Biggins SW, Vittinghoff E, Kohn MA, Inadomi JM. Ascites improves upon [corrected] serum sodium plus [corrected] model for end-stage liver disease (MELD) for predicting mortality in patients with advanced liver disease. Aliment Pharmacol Ther. 2009;30(7):741-8. [PMID: 19604177]

S84. Simons JP, Ng SC, Hill JS, Shah SA, Bodnari A, Zhou Z, Tseng JF. In-hospital mortality for liver resection for metastases: a simple risk score. J Surg Res. 2009;156(1):21-5. [PMID: 19577250]

S85. Caterino JM, Kulchycki LK, Fischer CM, Wolfe RE, Shapiro NI. Risk factors for death in elderly emergency department patients with suspected infection. J Am Geriatr Soc. 2009;57(7):1184-90. [PMID: 19558478]

S86. Chamogeorgakis T, Toumpoulis I, Tomos P, Ieromonachos C, Angouras D, Georgiannakis E, Michail P, Rokkas C. External validation of the modified Thoracoscore in a new thoracic surgery program: prediction of in-hospital mortality. Interact Cardiovasc Thorac Surg. 2009;9(3):463-6. [PMID: 19549645]

S87. Gomez EV, Bertot LC, Oramas BG, Soler EA, Navarro RL, Elias JD, Jiménez OV, Abreu Vazquez Mdel R. Application of a biochemical and clinical model to predict individual survival in patients with end-stage liver disease. World J Gastroenterol. 2009;15(22):2768-77. [PMID: 19522028]

S88. van Walraven C, Austin PC, Jennings A, Quan H, Forster AJ. A modification of the Elixhauser comorbidity measures into a point system for hospital death using administrative data. Med Care. 2009;47(6):626-33. [PMID: 19433995]

S89. Capuzzo M, Scaramuzza A, Vaccarini B, Gilli G, Zannoli S, Farabegoli L, Felisatti G, Davanzo E, Alvisi R. Validation of SAPS 3 Admission Score and comparison with SAPS II. Acta Anaesthesiol Scand. 2009;53(5):589-94. [PMID: 19419351]

S90. Giraldez RR, Sabatine MS, Morrow DA, Mohanavelu S, McCabe CH, Antman EM, Braunwald E. Baseline hemoglobin concentration and creatinine clearance composite laboratory index improves risk stratification in ST-elevation myocardial infarction. Am Heart J. 2009;157(3):517-24. [PMID: 19249423]

S91. Brandão A, Fuchs SC, Gleisner AL, Marroni C, Zanotelli ML, Cantisani G; Liver Transplantation Group. MELD and other predictors of survival after liver transplantation. Clin Transplant. 2009;23(2):220-7. [PMID: 19210688]

S92. Feudtner C, Hexem KR, Shabbout M, Feinstein JA, Sochalski J, Silber JH. Prediction of pediatric death in the year after hospitalization: a population-level retrospective cohort study. J Palliat Med. 2009;12(2):160-9. [PMID: 19207060]

S93. Soler-Cataluña JJ, Martínez-García MA, Sánchez LS, Tordera MP, Sánchez PR. Severe exacerbations and BODE index: two independent risk factors for death in male COPD patients. Respir Med. 2009;103(5):692-9. [PMID: 19131231]

S94. Kurz DJ, Bernstein A, Hunt K, Radovanovic D, Erne P, Siudak Z, Bertel O. Simple point-of-care risk stratification in acute coronary syndromes: the AMIS model. Heart. 2009;95(8):662-8. [PMID: 19066189]

2.5.2 Supplementary table: Assessed prediction models and their variables.

sRef	Disease/Clinical condition	Predictive model	Set of variables in each predictive model
S15	Cardiovascular disease	Acute Kidney Injury Network (AKIN)	serum creatinine criteria or urine output criteria
S67	Critical illness	Acute Lung Injury (ALI) score	chest X-ray, hypoxemia, PEEP, compliance
S94	Cardiovascular disease	Acute Myocardial Infarction in Switzerland (AMIS) model	age, Killip class, systolic blood pressure, heart rate, prehospital cardiopulmonary resuscitation, history of heart failure, history of cardiovascular disease
S1 S6 S8 S9 S17 S26 S29 S30 S33 S36 S50 S51 S53 S54 S64 S67 S71 S73 S82	Critical illness Gastroenterology-related Malignancies Infectious disease Other	Acute Physiology And Chronic Health Evaluation (APACHE) II	temperature, mean arterial pressure, heart rate, respiratory rate, oxygenation or PaO_2, arterial pH, serum sodium, serum potassium, serum creatinine, hematocrit, white blood cell (WBC) count, Clasgow Coma Score

S23	Critical illness	Acute Physiology And Chronic Health Evaluation (APACHE) III	pulse, mean blood pressure (BP), temperature, respiratory rate, PaO2, A-aDO2, Hct, WBC, Cr-No ARF, Cr-ARF, urine output, BUN (mmol/l), soium (mmol/l), albumin (g/l), bilirubin (mmol/l), Glu (mmol/l)
S31	Other	Prognostic model (unnamed)	age, burned surface area (BSA), inhalation injury
S41	Malignancies	American Joint Committee on Cancer/International Union Against Cancer (AJCC/UICC) TNM classification	Child-Pugh score, tumor morphology, serum a-fetoprotein (ng/dL), portal vein thrombosis
S64	Infectious disease	American Thoracic Society/Infectious Diseases Society of America (ATS/IDSA) major criteria	respiratory rate>30 breaths/min, PaO2/FiO2 ratio<250, multilobar infiltrates, confusion/disorientation, uremia (BUN level > 20mg/dL, leukopenia (WBC count, <4000 cells/mm, thrombocytopenia (PLT<100000 cells/mm), hypothermia (core temperature <36 C), hypotension requiring aggressive fluid resusciatation, invasive mechanical ventilation, septic shock with the need for vasopressors
S78	Cardiovascular disease	Aristotle Basic Complexity (ABC) score	mortality component, morbidity component, technical difficulty component
S70	Gastroenterology-related	AST to platelet ratio index (APRI)	platelet count, aspartate aminotransferase (AST)
S21	Cardiovascular disease	AusSCORE	age, New York Heart Association (NYHA) class, urgency of procedure, ejection fraction estimate, previous cardiac surgery, hypercholesterolemia (lipid-lowering treatment), peripheral vascular disease, cardiogenic shock
S57	Malignancies	Barcelona Clinic Liver Cancer (BCLC)	early stage (A) includes patients with asymptomatic early tumors suitable for radical therapies, intermediate stage (B) comprises patients with asymptomatic multinodular HCC, advanced stage (C) includes patients with symptomatic tumors and/or an invasive tumoral pattern (vascular invasion/extrahepatic spread), end-stage disease (D) contain patients with extremely grim prognosis

S6 S54	Gastroenterology-related	Bedside Index for Severity in Acute Pancreatitis (BISAP) score	BUN>25 mg/dl, impaired metal status (Glasgow coma scale score<15), systematic inflammatory response syndrome (SIRS), age>60 years, pleural effusion detected on imaging
S93	Other	BODE index	BMI, airflow obstruction, dyspnea, exercise capacity
S20	Cardiovascular disease	British Columbia Percutaneous Coronary Intervention (PCI) risk score	age, gender, emergency, left main disease, triple vessel disease, left ventricular ejection fraction (LVEF), New York Heart Association (NYHA), critical preprocedural sate, ST-elevation myocardial infarction (STEMI), other acute coronary syndrome (ACS), creatinine
S1	Critical illness	C4.5 classification tree	inotropic therapy (INOT), Glasgow value, (A-a)O2, gradient ((A-a)O2), age, chronic organ insufficiency (COI), mean arterial pressure (MAP)
S41 S57 S69	Malignancies	Cancer of the Liver Italian Program (CLIP)	Child-Pugh score, tumor morphology, serum alfa-fetoprotein (ng/dL) levels, portal vein thrombosis
S1	Critical illness	CHAID algorithm	inotropic therapy (INOT), Glasgow value, (A-a)O_2 gradient ((A-a)O_2), age, chronic organ insufficiency (COI), mechanical ventilation, trauma
S76 S82 S88	Critical illness Other	Charlson Comorbidity Index (CCI)	age, AIDS, cerebrovascular disease, chronic pulmonary disease, congestive heart failure, connective tissue disease, dementia, hemiplegia, leukemia, malignant lymphoma, myocardial infarction, peripheral vascular disease, ulcer disease, diabetes mellitus, liver disease, renal disease, malignant solid tumor
S7 S9 S12 S19 S22 S60 S87 S91	Critical illness Gastroenterology-related	Child–Turcotte–Pugh (CTP) score	bilirubin, albumin, prothrombin time prolong, ascites, encephalopathy

S60	Other	CHS index	shrinking, weakness, poor energy, slowness, low physical activity
S1	Critical illness	Classification And Regression Trees (CART)	inotropic therapy (INOT), Glasgow value, (A-a)O2 gradient ((A-a)O2), age, chronic organ insufficiency (COI)
S25	Other	Clinical Risk Index for Babies (CRIB)	birthweight, gestational age, maximum and minimum fraction of inspired oxygen and maximum base excess during the first 12 h, presence of congenital malformations
S25 S27	Other	Clinical Risk Index for Babies (CRIB) II	sex, birthweight, gestational age, temperature at admission, base excess
S59	Infectious disease	Community-acquired pneumonia-90 (CAP-90) index	pre-illness functional status, Charlson index (composite measure of co-morbid illnesses) and severity on admission
S6	Gastroenterology-related	Computed Tomography Severity Index (CTSI)	Balthazar grade and necrosis percentage
S63	Infectious disease	CRB65	new confusion, respiratory rate >30/min, systolic blood pressure <90 mmHg or diastolic blood pressure =<60 mmHg, age >65 years
S14	Malignancies	CR-POSSUM	age, pre-existing cardiac failure, systolic blood pressure, pulse, haemoglobin, serum urea nitrogen, operative severity, peritoneal soiling, cancer stage, mode of surgery, observed 30-day mortality
S2 S18 S42 S63 S80	Infectious disease	CURB-65 score	new confusion, urea>7 mM, respiratory rate >30/min, systolic blood pressure <90 mmHg or diastolic blood pressure =<60 mmHg, age >65 years
S13	Critical illness	customized Simplified Acute Physiology Score (SAPS) III	age, co-morbidities, length of stay before ICU admission, intra-hospital location before ICU admission, use of major therapeutic options before ICU admission, ICU admission: planned or unplanned, reasons for ICU admission, surgical status at ICU admission, anatomical site of surgery, acute

			infection at ICU admission, estimated GCS (lowest), total bilirubine (highest) in mg/dL, total bilirubin (highest) µmol/L, body temperature (highest), creatinine (highest) in mg/dL, creatinine (highest) µmol/L, heart rate (highest), leukocytes, hydrogen ion concettration (lowest), platelets, systolic blood pressure, oxygenation
S88	Other	Elixhauser comorbidity score	congestive heart failure, cardiac arrhythmias, valvular disease, pulmonary circulation disorders, peripheral vascular disorders, hypertension (uncomplicated and complicated), paralysis, other neurological disorders, chronic pulmonary disease, diabetes - uncomplicated, diabetes complicated, hypothyroidism, lymphoma, metastatic cancer, solid tumor without metastasis, rheumatoid arthritis/collagen vascular disease, coagulopathy, obesity, weight loss, fluid and electrolyte disorders, blood loss anemia, deficiency anemias, alcohol abuse, drug abuse, psychoses, depression
S40	Other	Emergency trauma score (EMTRAS)	age, Glagow Coma Scale, base excess, prothrombin time
S21 S35 S45 S56 S74	Cardiovascular disease	European system for cardiac operative risk evaluation (EuroSCORE)	age (years), gender, chronic pulmonary disease, extracardiac arteriopathy, neurological dysfunction, previous cardiac surgery, creatinine > 200 Mmol/L, active endocarditis, critical preoperative state, unstable angina, LV function, recent MI, pulmonary hypertension, emergency, operation other than isolated CABG, surgery on thoracic aorta, post infarct septal rupture
S70	Gastroenterology-related	FIB4	platelet count, alanine aminotransferase (ALT), aspartate aminotransferase (AST)
S70	Gastroenterology-related	FibrometerA	prothrombin index (PI), a-2 macroglobulin, hyaluronic acid, age
S70	Gastroenterology-related	Fibrosis staging at biopsy	scale runs from 0 to 4

S70	Gastroenterology-related	FibroTest	α-2 macroglobulin, haptoglobin, gamma glutamyl transpeptidase (GGT), apolipoprotein A1, total bilirubin corrected for age, gender
S70	Gastroenterology-related	Forns	age, gamma-glutamyl transferase (GGT), cholesterol, platelet count, prothrombin time
S39	Cardiovascular disease	Glasgow Aneurysm Score (GAS)	age, shock, myocardial disease, cerebrovascular disease, renal insufficiency
S49	Critical illness	Glasgow Coma Scale (GCS)	eye opening, verbal response, motor response
S53	Gastroenterology-related	Glasgow criteria	on admission (age >55 yrs, WBC Count >15 x109/L, Blood Glucose >200 mg/dL (No Diabetic History), Serum Urea >16 mmol/L (No response to IV fluids), Arterial Oxygen Saturation <76 mmHg), within 48 hours (Serum Calcium <2 mmol/L, Serum Albumin <34 g/L, LDH >219 units/L, AST/ALT >96 units/L)
S81	Cardiovascular disease	Global Registry of Acute Coronary Events (GRACE) risk score	age (years), heart rate (bpm), systolic blood pressure (mmHg), creatinine (mg/dL), Killip class, cardiac arrest at admission, elevated cardiac markers, ST segment deviation
S26	Critical illness	Global Simplified Acute Physiology Score (SAPS) III	age (years), co-morbidities, length of stay before intensive care unit (ICU) admission, intra-hospital location before ICU admission, use of major therapeutic options before ICU admission, ICU admission (planned or unplanned), reasons for ICU admission, surgical status at ICU admission, anatomical site of surgery, acute infection at ICU admission, estimated Glagow Coma Scale, total bilirubine (mg/dL), total bilirubine (µmol/L), body temperature, degrees celcius, creatinine (mg/dL), creatinine (µmol/L), heart rate (beats/min), leukocytes, hydrogen ion concentration, platelets, systolic blood pressure, oxygenation

S76	Other	Hematopoietic Cell Transplantation Comorbidity Index (HCT-CI)	arrhythmia, cardiac, inflammatory bowel disease, diabetes, cerebrovascular disease, psychiatric disturbance, hepatic (mild), obesity, infection, rheumatologic, peptic ulcer, moderate/severe renal, moderate pulmonary, prior solid tumor, heart valve disease, severe pulmonary, moderate/severe hepatic
S70	Gastroenterology-related	Hepascore	bilirubin, gamma glutamyl transpeptidase (GGT), hyaluronic acid, α-2 macroglobulin, age, gender
S79	Infectious disease	HIV biomarkers	CD4 cell count, HIV RNA, AIDS-defining conditions
S79	Infectious disease	HIV biomarkers + Non-HIV biomarkers	CD4 cell count, HIV RNA, AIDS-defining conditions, haemoglobin, transaminases, platelets, creatinine, hepatitis B and C serology
S30 S40	Other	Injury Severity Score (ISS)	regions of injury (head and neck, face, chest, abdomen, extremity, external)
S17	Malignancies	Intensive Care National Audit & Research Centre (ICNARC) physiology score	highest heart rate, lowest systolic blood pressure, highest temperature, lowest respiratory rate, PaO_1/FiO_2 ratio, lowest arterial pH, highest serum urea, highest serum creatinine, highest serum sodium, urine output, lowest white blood count, sedated-paralyzed- Glagow Coma Scale
S32	Other	Intermountain Risk Score	age, sex, hematocrit, white blood cell count, platelet count, mean corpuscular volume, mean corpuscular hemoglobin concentration, red cell distribution width, mean platelet volume, sodium, potassium, bicarbonate, calcium, glucose, creatinine
S61	Other	International Classification Injury Severity Score (ICISS)	simply determining the product of the survival risk ratios (SRRs) for each individual injury ICD-9 codes. Included variables not given.
S49	Critical illness	Intra Cerebral Haemorrhage (ICH) score	Glagow Coma Scale, age over 80, ICH volume over 30 ml, intraventricular hemorrhage, infratentorial origin of hemorrhage

S41 S57 S69	Malignancies	Japan Integrated Staging (JIS)	Child-Pugh score, tumor morphology, serum a-fetoprotein (ng/dL), portal vein thrombosis
S90	Cardiovascular disease	Laboratory index (LI)	hemoglobin (Hb) levels and renal function (creatinine clearance)
S41	Malignancies	Liver Cancer Study Group of Japan (LCSGJ)	Child-Pugh score, tumor morphology, serum a-fetoprotein (ng/dL), portal vein thrombosis
S1	Critical illness	Logistic Regression model	Age, heart rate, Glasgow Coma Scale, (A-a)O_2 gradient, inotropic therapy, mechanical ventilation, acute renal failure, COI, trauma
S7 S19	Gastroenterology-related	MESO index	MELD to SNa ratio x 10
S57 S60 S83	Gastroenterology-related Malignancies	Model for End-Stage Liver Disease (MELD) - Sodium score	serum creatinine, the international normalized ratio (INR) for prothrombin time, serum bilirubin, cirrhosis etiology (alcohol or cholestasis, other), sodium
S4 S7 S9 S12 S16 S19 S22 S38 S57 S60 S83 S87 S91	Critical illness Gastroenterology-related Infectious disease Malignancies Other	Model for End-Stage Liver Disease (MELD) score	serum creatinine, the international normalized ratio (INR) for prothrombin time, serum bilirubin, cirrhosis etiology (alcohol or cholestasis, other)
S5	Malignancies	Prognostic model for hepatocellular carcinoma	alpha-fetoprotein, total albumin concentration, venous infiltration, tumor size, new AJCC stage, number of tumor nodule

S11	Other	Mortality Probabilistic Model at 24-Hours (MPMHOS-24)	age, type of admission, chronic heart failure, chronic respiratory failure, chronic liver disease, cancer, dementia, haemoglobin <110 mg/dL, creatinine ≥2 mg/dL
S1 S36	Critical illness Infectious disease	Mortality Probability Models (MPM) II	age, prothrombin time, PaO_2<60 mmHg, vasoactive drugs>1 hour, mechanical ventilation, intracranial mass effect, confirmed infection, coma, urine output<150 mL/8 hours, creatinine >2.0 mg/dl, cirrhosis, metastatic neoplasm, medical or unscheduled surgery admission
S72	Gastroenterology-related	Mortality risk model among patients with bleeding peptic ulcers	age>70 y, presence of listed comorbidities, more than 1 listed comorbidity, hematemesis, initial systolic blood pressure<100 mmHg, in-hospital bleeders, presence of H. pylori, development of rebleeding, need of operation.
S47	Infectious disease	Multidimensional Prognostic Index (MPI)	age, sex, the presence of comorbid illnesses, vital sign abnormalities, and some laboratory and radiographic abnormalities
S40	Other	New Injury Severity Score (NISS)	sums the severity score for the three most severe injuries, regardless of body region (according to ISS)
S53	Gastroenterology-related	new Japanese severity score (JSS)	age>70 years, SIRS score>3, CRP>15 mg/dl, Ca<7.5 mg/dl, PLT<1 x 10000/mm3, LDH>2 folds of upper normal limit, BUN>40 mg/dl or creatinine>2 mg/dl, PaO_2<60 mmHg or respiratory failure, BE<-3 mEq/L or shock
S79	Infectious disease	Non-HIV biomarkers	haemoglobin, transaminases, platelets, creatinine, hepatitis B and C serology
S43	Other	Paediatric Logistic Organ Dysfunction (PELOD) score	heart rate, systolic blood pressure, PaO_2/FiO_2, $PaCO_2$, mechanical ventilation,creatinine, Glascow Coma Scale, pupillary reactions, white blood cell count, platelet count, serum glutamic oxaloacetic transaminase, prothrombin time or international normalized ratio, pulmonary score, CVS score, hepatic score, neurologic score, renal score, hematologic score

S35	Cardiovascular disease	Parsonnet score 2000-version	age, gender, body weight, aortic stenosis, congenital heart defect, arterial hypertension, pulmonary hypertension, LV aneurysm, LV ejection fraction, asthma, dialysis, acute renal failure, diabetes, paraplegia, pacemaker, intra-aortic balloon pump, cardiogenic shock, combined surgery, urgent/emergency operation, reoperation
S92	Other	Pediatric death prediction model	Not given
S2 S42 S47 S63 S80	Infectious disease	Pneumonia Severity Index (PSI)	age of more than 50 years, five coexisting illnesses (neoplastic disease, congestive heart failure, cerebrovascular disease, renal disease, and liver disease), and five physical-examination findings (altered mental status; pulse, 125 per minute; respiratory rate, 30 per minute; systolic blood pressure, 90 mm Hg; and temperature, 35°C or 40°C), male sex, nursing home residence, blood urea nitrogen concentration (30 mg per deciliter [11 mmol per liter]), glucose concentration (250 mg per deciliter [14 mmol per liter]), hematocrit (30 percent), sodium concentration (130 mmol per liter), partial pressure of oxygen (60 mm Hg), arterial pH, pleural effusion
S3	Other	Predictors of Respiratory Insufficiency and Mortality (PRIM) score	severe injury (Asia impairement Scales A and B), hemodynamic instability, neurological deterioration, mechanical ventilation
S72	Gastroenterology-related	Pre-endocopic prediction score	age>70 y, presence of listed comorbidities, more than 1 listed comorbidity, hematemesis, initial systolic blood pressure<100 mmHg, in-hospital bleeders
S76	Other	Pretransplantation Assessment of Mortality (PAM)	age, donor type (related matched, unrelated, related mismatched), disease risk category, conditioning regimens, pretransplant serum creatinine (mg/dL), serum glutamic pyruvic transaminase (mg/dL), percent of predicted forced expiratory volume in one secong (FEV1), percent of predicted carbon monoxide diffusion capacity adjusted for hematocrit

S53	Gastroenterology-related	old Japanese severity score (JSS)	The included variables are not listed.
S70	Gastroenterology-related	Pugh prognostic score	Not given
S6 S53	Gastroenterology-related	Ranson's criteria	on admission (age in years > 55 years, white blood cell count > 16000 /mcL, blood glucose > 11 mmol/L (>200 mg/dL), serum AST > 250 IU/L, serum LDH > 350 IU/L); after 48 hours (Haematocrit fall > 11.3444%, increase in BUN by 1.8 or more mmol/L (5 or more mg/dL) after IV fluid hydration, hypocalcemia (serum calcium < 2.0 mmol/L (<8.0 mg/dL)), hypoxemia (PO2 < 60 mmHg), base deficit > 4 Meq/L, estimated fluid sequestration > 6 L)
S11	Other	Rapid Emergency Medicine Score (REMS)	Age, blood pressure, respiratory rate, heart rate, Glasgow coma scale, peripheral oxygen saturation
S40	Other	Revised Trauma Score (RTS)	Glasgow Coma Scale, systolic blood pressure, respiratory rate
S78	Cardiovascular disease	Risk Adjustment for Congenital Heart Surgery (RACHS-1) categories	scale runs from 1 to 6
S85	Infectious disease	Risk model for elderly emergency department (ED) patients	respiratory failure (respiratory rate>20, pulse oximetry<90%, pulse oximetry<94% on supplemental oxygen, or need for intubation), tachycardia, cardiac failure (systolic blood pressure<90 mmHg after a fluid challenge, need for vasopressors, or venous lactic acid level > 4 mmol/L), pre-existing terminal illness, platelets<150.000/µL
S58	Cardiovascular disease	risk model for perioperative mortality of endovascular vs open repair of Abdominal Aortic Aneurysm	age, sex, renal failure, congestive heart failure (CHF), peripheral vascular disease (PVD) or cerebrovascular disease (CBVD)

S36	Infectious disease	Risk model for short-term mortality of severe sepsis	logistic organ dysfunction, septic shock, multiple sites of infection, SAPS, fatal illness by McCabe Score, no chroni illness (one, two, or more)
S66	Malignancies	Risk model for survival of NSCLC patients	gender, World Health Organization performance status (WHO-PS), forced expiratory volume in 1 s (FEV1), number of positive lymph node sations (PLNSs), gross tumor volume (GTV)
S9 S15 S67	Critical illness Cardiovascular disease	Risk of renal failure, Injury to the kidney, Failure of kidney function, Loss of kidney function, and End-stage renal disease (RIFLE) classification	serum creatinine, glomerular filtration rate (GFR)
S84	Malignancies	Risk score for in-hospital mortality for Liver Resection for Metastases	age group, Charlson score, procedure type (RFA/enucleation, wedge resection, lobectomy), sex, hospital type (teaching, nonteaching)
S75	Cardiovascular disease	Risk score for in-hospital mortality in patients hospitalized with heart failure	age, systolic blood pressure, blood urea nitrogen, heart rate, sodium, chronic obstructive pulmonary disease, nonblack race
S77	Other	Risk score for mortality in Renal Transplant Recipients	age, pretransplant diabetes, positive Hepatitis C virus antibodies, new onset of diabetes after transplantation at the first year, serum creatinine at the first year (mg%), proteinuria>1 g at the first year, use of tacrolimus at the first year, use of mycophenolate mofetil at the first year
S72	Gastroenterology-related	Rockall score	age, shock, comorbidity, diagnosis, evidence of bleeding
S65	Other	Scoring system predicting mortality following acute burn injury	age (years), burned surface area (%), inhalation injury

S9 S22 S26 S46 S51 S52 S67 S68	Critical illness Gastroenterology-related Infectious disease Malignancies	Sequential Organ Failure Assessment (SOFA) score	respiratory system (PaO_2/FiO_2), nervous system (Glasgow coma scale), cardio-vascular system (mean arterial pressure or administration of vasopressors required), liver (bilirubin (mg/dl)), coagulation (platelet count x1000/mcl), renal system (creatinine (mg/dl) (or urine output)
S2	Infectious disease	Severe Community Acquired Pneumonia (SCAP) score	systolic blood pressure <90 mm Hg, arterial pH<7.30, respiratory rate >30 breaths/min, blood urea nitrogen (BUN) >30 mg/dl, oxygen arterial pressure <54 mm Hg or PaO2/FiO2 <250 mm Hg, altered mental status, age>80 yr, multilobar/bilateral lung affectation in X-rays
S35	Cardiovascular disease	Simple graphic pocket-card score for cardiac surgery	age, gender, indulin dependent, renal failure, peripheral vascular disease, reoperation, urgent-emergent-salvage status, preoperative intra-aortic balloon pump, aortic-mitral valve replacement, aortic-mitral valve repair, thoracic aorta replacement, aortic acute dissection, heart transplant, surgery combined, one-two-three vessel disease, moderate-severe left ventricular dysfunction
S94	Cardiovascular disease	Simple risk index	age, heart rate, systolic blood pressure
S10	Other	Simple risk score	serum urea nitrogen level>25mg/dL, acute mental status change, pulse>109/min, age<65 years
S1 S17 S29 S36 S37 S49 S71 S89	Critical illness Infectious disease Malignancies	Simplified Acute Physiology Score (SAPS) II	type of admission, chronic diseases, Glasgow coma scale, age, systolic blood pressure, heart rate, temperature, mechanical ventilation or CPAP PaO_2/FiO_2, urine output, serum urea or BUN, white blood cell count, potassium, sodium, HCO_3, bilirubin

S37 S89	Critical illness	Simplified Acute Physiology Score (SAPS) III	age, co-morbidities, use of vasoactive drugs before intensive care unit (ICU) admission, intrahospital location before ICU admission, length of stay in the hospital before ICU admission, reason(s) for ICU admission, planned/unplanned ICU admission, surgical status at ICU admission, anatomical site of surgery, presence of infection at ICU admission and place acquired, lowest estimated GCS, highest heart rate, kowest systolic blood pressure, highest bilirubine, highest body temperature, highest creatinine, highest leukocytes, lowest platelets, lowest hydrogen ion concentration (pH), ventilatory support and oxygenation
S78	Cardiovascular disease	Society of Thoracic Surgeons (STS)-European Association for Cardiothoracic Surgery (EACTS) categories	scale runs from 1 to 5
S78	Cardiovascular disease	Society of Thoracic Surgeons (STS)-European Association for Cardiothoracic Surgery (EACTS) score	scale runs from 0.1 to 5
S86	Other	Thoracoscore	patient's age, gender, priority of the procedure, ASA class, Zubrod score, number of co-morbidities, presence of malignancy, dyspnea score, and type of procedure
S55 S90 S94	Cardiovascular disease	Thrombolysis In Myocardial Infarction (TIMI) risk score	age >75 and 65-74 years, systolic blood pressure <100mmHg, heart rate > 100 beats/min, Killip classes II-IV, anterior ST-elevation myocardial infarction or left-bundle branch block, history of diabetes, hypertension, or angina, body weight < 67 kg, time to start of intravenous thrombolysis of >4 h
S90	Cardiovascular disease	TIMI-risk score + Laboratory index	(age >75 and 65-74 years, systolic blood pressure <100mmHg, heart rate > 100 beats/min, Killip classes II-IV, anterior ST-elevation myocardial infarction or left-bundle

			branch block, history of diabetes, hypertension, or angina, body weight < 67 kg, time to start of intravenous thrombolysis of >4 h) and hemoglobin levels, baseline creatinine clearance
S57	Malignancies	TNM	size and nuber of tumors, node, metastasis
S57	Malignancies	Tokyo score	serum albumin, bilirubin, size and number of tumours
S34	Other	Trauma Injury Severity Score (TRISS)	Injury Severity Score (ISS), Revised Trauma Score (RTS) (including ISS, systolic blood pressure, respiratory rate, coma score), and age
S30 S40	Other	Trauma Revised Injury Severity Score	age, Glagow Coma Scale, base excess, prothrombin time
S34	Other	Trauma Risk Adjustment Model (TRAM)	anatomic injury severity is represented by the AIS score of the 2 most severe injuries and the body region of the most severe injury, physiological response to injury by the Glasgow coma scale, systolic blood pressure and heart rate; and physiological reserve by age and number of comorbidities

2.5.3 Supplementary table: Assessed single predictors.

sRef	Disease/Clinical condition	Single predictor	Type of predictor
S44	Other	Base deficit	Biomarker
S24 S48 S68	Cardiovascular disease Infectious disease Other	B-type natriuretic peptide (BNP)	Biomarker
S68	Infectious disease	C2	Biomarker
S68	Critical illness Infectious disease	C-reactive protein (CRP)	Biomarker
S44	Other	lactate	Biomarker
S42	Infectious disease	Midregional proadrenomedullin (MR-proADM)	Biomarker
S24 S28 S48 S62 S73	Cardiovascular disease Critical illness Other	N-terminal-pro-B-type natriuretic peptide (NT-proBNP)	Biomarker
S42 S68	Infectious disease	procalcitonin	Biomarker

Section 3.

Comparisons of established risk prediction models for cardiovascular disease.

3. Comparisons of established risk prediction models for cardiovascular disease: systematic review.[116]

Siontis GC, Tzoulaki I, Siontis KC, Ioannidis JP.

British Medical Journal. 2012. (PMID: 22628003)

BMJ

BMJ 2012;344:e3318 doi: 10.1136/bmj.e3318 (Published 24 May 2012)

Comparisons of established risk prediction models for cardiovascular disease: systematic review

George C M Siontis *research associate*[1], Ioanna Tzoulaki *lecturer*[1], Konstantinos C Siontis *research associate*[1], John P A Ioannidis *professor*[2]

[1]Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece; [2]Stanford Prevention Research Center, Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305-5411, USA

Abstract

Objective To evaluate the evidence on comparisons of established cardiovascular risk prediction models and to collect comparative information on their relative prognostic performance.

Design Systematic review of comparative predictive model studies.

Data sources Medline and screening of citations and references.

Study selection Studies examining the relative prognostic performance of at least two major risk models for cardiovascular disease in general populations.

Data extraction Information on study design, assessed risk models, and outcomes. We examined the relative performance of the models (discrimination, calibration, and reclassification) and the potential for outcome selection and optimism biases favouring newly introduced models and models developed by the authors.

Results 20 articles including 56 pairwise comparisons of eight models (two variants of the Framingham risk score, the assessing cardiovascular risk to Scottish Intercollegiate Guidelines Network to assign preventative treatment (ASSIGN) score, systematic coronary risk evaluation (SCORE) score, Prospective Cardiovascular Münster (PROCAM) score, QRESEARCH cardiovascular risk (QRISK1 and QRISK2) algorithms, Reynolds risk score) were eligible. Only 10 of 56 comparisons exceeded a 5% relative difference based on the area under the receiver operating characteristic curve. Use of other discrimination, calibration, and reclassification statistics was less consistent. In 32 comparisons, an outcome was used that had been used in the original development of only one of the compared models, and in 25 of these comparisons (78%) the outcome-congruent model had a better area under the receiver operating characteristic curve. Moreover, authors always reported better area under the receiver operating characteristic curves for models that they themselves developed (in five articles on newly introduced models and in three articles on subsequent evaluations).

Conclusions Several risk prediction models for cardiovascular disease are available and their head to head comparisons would benefit from standardised reporting and formal, consistent statistical comparisons. Outcome selection and optimism biases apparently affect this literature.

Introduction

Cardiovascular disease carries major morbidity and mortality.[1] To effectively implement prevention strategies clinicians need reliable tools to identify individuals without known cardiovascular disease who are at high risk of a cardiovascular event.[2 3] For this purpose, multivariable risk assessment tools, such as the Framingham risk score, are recommended for clinical use.[4] Besides the Framingham risk score, several other risk prediction tools combining different sets of variables have been developed and validated.[5 6] Some investigators have evaluated the performance of two or more risk prediction models in the same populations.

We evaluated the evidence on comparisons of established cardiovascular risk prediction models. We systematically collected comparative information on discrimination, calibration, and reclassification performance and evaluated whether specific biases may have affected the inferences of studies comparing such models.

Correspondence to: J P A Ioannidis jioannid@stanford.edu

Extra material supplied by the author (see http://www.bmj.com/content/344/bmj.e3318?tab=related#webextra)

Appendix: search strategy
Figure showing selection of eligible studies of risk models comparisons
Table 1: details of examined risk models for cardiovascular disease prediction
Table 2: reporting and management of missing data
Table 3: potential biases and authors' comments on performance of risk models
Table 4: discrimination performance according to metrics other than area under the receiver operating characteristic curve
Table 5: calibration metrics

3.1 BACKGROUND

Cardiovascular disease carries major morbidity and mortality.[1] To effectively implement prevention strategies clinicians need reliable tools to identify individuals without known cardiovascular disease who are at high risk of a cardiovascular event.[177,178] Risk prediction models have great potential to support clinical decision making and are increasingly incorporated into clinical guidelines. For this purpose, multivariable risk assessment tools, such as the Framingham risk score, are recommended for clinical use;[179] whereas besides the Framingham risk score, several other risk prediction tools combining different sets of variables have been developed and validated for cardiovascular disease - SCORE, QRISK, and the Reynolds risk score— to mention just a few.[180,181] With so many prediction models for similar outcomes or target populations, clinicians have to decide which model should be used on their patients. To make this decision they need to know, as a minimum, how well the score predicts disease in people outside the populations used to develop the model ("what is the external validation?") and which model performs best.[109] A plea for more direct comparisons is increasingly heard in the field of therapeutic intervention and diagnostic research and may be echoed in that of prediction model validation studies. Many more prediction models have been developed than have been validated in independent datasets. Moreover, few models developed for similar outcomes and target populations have been directly validated and compared.[109]

Some investigators have evaluated the performance of two or more risk prediction models in the same populations. The purpose of this evaluation was to systematically summarize the available evidence on comparisons of established cardiovascular risk prediction models. For this, we systematically collected comparative information on discrimination, calibration, and reclassification performance and evaluated whether specific biases may have affected the inferences of studies comparing such models.

3.2 METHODS

Eligible Models and Literature Search

Prediction models for the risk of cardiovascular disease in general populations that were considered in two recent expert reviews[180,181] were considered: the Framingham risk score[24,64,154] (and the national cholesterol education program–adult treatment panel III version[182]), the assessing cardiovascular risk to Scottish Intercollegiate Guidelines Network to assign preventative treatment (ASSIGN) score[183], systematic coronary risk evaluation (SCORE) score[184], Prospective Cardiovascular Münster (PROCAM) score[185], QRESEARCH cardiovascular risk (QRISK1 and QRISK2) algorithms[29,31], Reynolds risk score[25,26], and the World Health Organization/International Society of Hypertension score.[186] Different versions of the Framingham risk score were categorised as Framingham risk score (including the Framingham risk score described by Anderson et al for risk of coronary heart disease and stroke[24] and the Framingham risk score proposed by Wilson et al[154]) (also proposed by National Institute for Health and Clinical Excellence guidelines) and as FRS (CVD) (which included the global Framingham risk score equations to predict cardiovascular disease[64]). Additional details are provided in **Table 3.1**.

Medline (last update July 2011) was searched for articles with data on the performance of at least two of these models. We also scrutinised the received citations (through SCOPUS) of the primary publications of these models (whenever applicable) and the references of all eligible papers for any additional relevant studies. The primary screening algorithm for the identification of eligible articles consisted of the full name of the prognostic models, the respective abbreviation, and any other known variation of these terms (*(Framingham OR FRS OR Framingham risk score OR NCEP ATP III OR National Cholesterol Education Program Adult Treatment Panel III), (Systematic Coronary Risk Evaluation), (Assessing Cardiovascular Risk to Scottish Intercollegiate Guidelines Network/SIGN to Assign Preventative Treatment score OR ASSIGN-SCORE OR ASSIGN SCORE OR ASSIGN OR SHHEC OR Scottish Heart Extended Cohort), (QRISK* OR QRESEARCH), (PROCAM OR Prospective Cardiovascular*

Münster Study OR Prospective Cardiovascular Münster Scoring System), (WHO/ISH OR WHO ISH OR World Health Organization/International Society of Hypertension), (RRS OR Reynolds Risk Score)). Titles and abstracts were screened first and potentially eligible articles scrutinised in full text. No year or language restrictions were applied.

Table 3.1: Details of the examined risk models for cardiovascular disease prediction.

Risk Model	Set of variables	Outcomes	Geographical origin	Web site/Risk calculator
FRS[24,64,154]	age, diabetes, HTN-related medications, HDL, sex, SBP, smoking, TC	CHD (angina, MI, sudden death)	United States	www.nhlbi.nih.gov/guidelines/cholesterol/index/htm and www.framinghamheartstudy.com and http://cvdrisk.nhlbi.nih.gov/calculator.asp
SCORE[184]	age, sex, smoking, SBP, TC or TC/HDL ratio (higher & lower CVD risk)	CVD mortality	Europe	www.HeartScore.org
ASSIGN-SCORE[183]	age, area-based index, diabetes, family history, HDL, sex, SBP, smoking, TC	CVD mortality, CHD admission, coronary revascularization (CABG, PTCA)	Scotland	www.assign-score.com
QRISK1[31], **QRISK2**[29]	**QRISK1:** age, area-based index of deprivation, BMI, diabetes, family history, HTN-related medications, sex, SBP, smoking, TC/HDL ratio **QRISK2:** QRISK1 variables & chronic disease, ethnicity	CVD (MI, CHD, stroke, TIA)	United Kingdom	www.qrisk.co.uk

PROCAM[185]	age, diabetes, HDL, LDL, sex, SBP, smoking	major CV events (coronary and cerebrovascular)	Germany	www.chd-taskforce.com/calculator
WHO/ISH[186]	age, diabetes, sex, SBP, smoking, +/- TC (different charts for worldwide regions)	CVD	None	www.who.int/cardiovascular_diseases/guidelines/Pocket_GL_information/en/index.html
RRS[25,26]	age, family history, HbA1c (if diabetes), HDL, hsCRP, sex, SBP, smoking, TC	CVD mortality, coronary ravsvularization, MI, stroke	United States	www.reynoldsriskscore.com and http://www.reynoldsriskscore.org/default.aspx

FRS, Framingham risk score; NCEP ATP III, National Cholesterol Education Program Adult Treatment Panel III; SCORE, Systematic Coronary Risk Evaluation; ASSIGN-SCORE, Assessing Cardiovascular Risk to Scottish Intercollegiate Guidelines Network/SIGN to Assign Preventative Treatment score; SHHEC, Scottish Heart Extended Cohort; PROCAM, Prospective Cardiovascular Münster Scoring System; WHO/ISH, World Health Organization/International Society of Hypertension; RRS, Reynolds Risk Score; BP, blood pressure; HTN, hypertension; TC, total cholesterol; HDL, high-density lipoprotein; CHD, coronary heart disease; SBP, systolic blood pressure; MI, myocardial infarction; CABG, coronary artery bypass graft; PTCA, percutaneous transluminal coronary angioplasty; TIA, transient ischemic attack; BMI, body-mass index.

Study Eligibility

Articles were eligible if they examined at least two pertinent risk models for the prediction of cardiovascular disease in populations without cardiovascular disease or general populations. We included original articles irrespective of sample size and duration of follow-up. Eligible outcomes were cardiovascular disease (and any composite cardiovascular disease end point), cardiovascular disease mortality, and coronary heart disease, including stable disease and acute coronary syndromes. When different published data on identical comparisons were identified comparing the same models, in the same cohort, and for the same outcome, we kept only the data that included the largest number of events. We excluded cross sectional studies, studies where all-cause mortality was the only outcome, studies that used models to calculate the baseline risk without providing outcome data, and studies including exclusively patients with specific morbidities—that is, patients with known cardiovascular disease, diabetes, or other diseases. Two investigators independently carried out the literature searches and assessed the studies for eligibility. Discrepancies were resolved by consensus and arbitration by two other investigators.

Data Extraction

Two investigators independently extracted data from the main paper and any accompanying supplemental material. The following items of interest were recorded in standardised forms: study design (prospective or retrospective), year of publication, sample size, type of population, percentage of baseline population with pre-existing cardiovascular disease, and reported risk models. We recorded the clinical end points assessed in each study (cardiovascular disease, cardiovascular disease mortality, coronary heart disease) and the respective number of events. When multiple different eligible outcomes or populations were identified in the same model comparison, we considered each outcome or cohort separately. Similarly, when more than two prognostic models were presented in an article, we considered all possible pairwise comparisons as eligible. Whenever a study also examined subgroups, such as males and females, we focused on the whole population unless only data per subgroup were provided; in those cases, we extracted data for each eligible subgroup separately.

Moreover, for each study we also captured whether the authors reported the presence of missing data on examined outcomes and on variables included in risk prediction models; and, if so, we recorded how missing data were managed (with imputation and by which methods, exclusion of missing observations, or other). We further extracted information on the geographical origin of each study and noted whether it was the same country to the one in which one (or both) of the compared models was initially developed.

For each model in each article we extracted metrics on discrimination (area under the receiver operating characteristic curve (or the equivalent C statistic), D statistic, R^2 statistic, and Brier score), their 95% confidence intervals, and the *P* value for comparison between models when available.[147,165] We also captured calibration[168] and reclassification[125,169] metrics. We extracted information on whether the observed versus predicted ratio and lack of fit statistics were reported, and whether the calibration plot was shown. Finally, we extracted information on reclassification statistics, such as the net reclassification index, and on the classification percentages of each model along with the thresholds used by each study.

Data Analysis and Evaluation of Biases

We analysed each risk model pairwise comparison separately. For each comparison we noted the model with a numerically higher area under the receiver operating characteristic curve estimate, and whether there was formal statistical testing of the difference in areas under the receiver operating characteristic curve. When confidence intervals were not available, we estimated them as previously proposed.[92] We also recorded separately which pairwise comparisons had a relative difference in area under the receiver operating characteristic curve exceeding 5% (for example, if the worse score had an area under the receiver operating characteristic curve of 0.70, the better score had one >0.70×1.05=0.735). The choice of a 5% threshold was chosen for descriptive purposes only. Furthermore, we noted whether models differed in other performance metrics. Calibration was considered better when the observed to predicted ratio was closer to 1.

We also evaluated the potential for outcome selection and optimism biases. Some of the examined risk scores have been originally developed for different cardiovascular outcomes (**Table 3.1**). We evaluated whether the examined outcome in each comparison was used in the original development of only one of the two compared models and, if so, whether the outcome-congruent model showed better performance. Owing to optimism bias, a new model may have better performance than the competing standard model when it is first presented, but not in subsequent comparisons. Therefore we noted whether each article described the application of previously established models or was the first to describe or validate a specific model or models. Moreover, authors who developed one model may favour publishing results that show its superiority against competing models. We thus noted whether any of the study authors had been involved in the development of any of the assessed models. Finally, we recorded the authors' comments on the relative performance of the model and examined whether these were affected by such potential biases. Analyses were done in Stata 10.1 (College Station, TX). P values are two tailed.

3.3 RESULTS

Inclusion of Studies

Of 672 published articles screened at title and abstract level, 74 were identified as potentially eligible for inclusion in the review. Of these, 58 articles were excluded because they only compared models using a baseline risk calculation without association with outcomes (n=20); assessed only patients with specific conditions (diabetes (n=11), HIV infection (n=4), known cardiovascular disease (n=3), liver transplantation (n=1), schizoaffective disorder (n=1), systemic lupus erythematosus or rheumatoid arthritis (n=1)); or had ineligible model comparisons (n=10), ineligible outcomes (non-cardiovascular disease outcomes) (n=6), or duplicate comparisons (n=1) (**Figure 3.1**). Searches of references and citations yielded another four eligible articles. Overall, 20 articles[26-29,31,183,185,187-199] were analysed (**Table 3.2**).

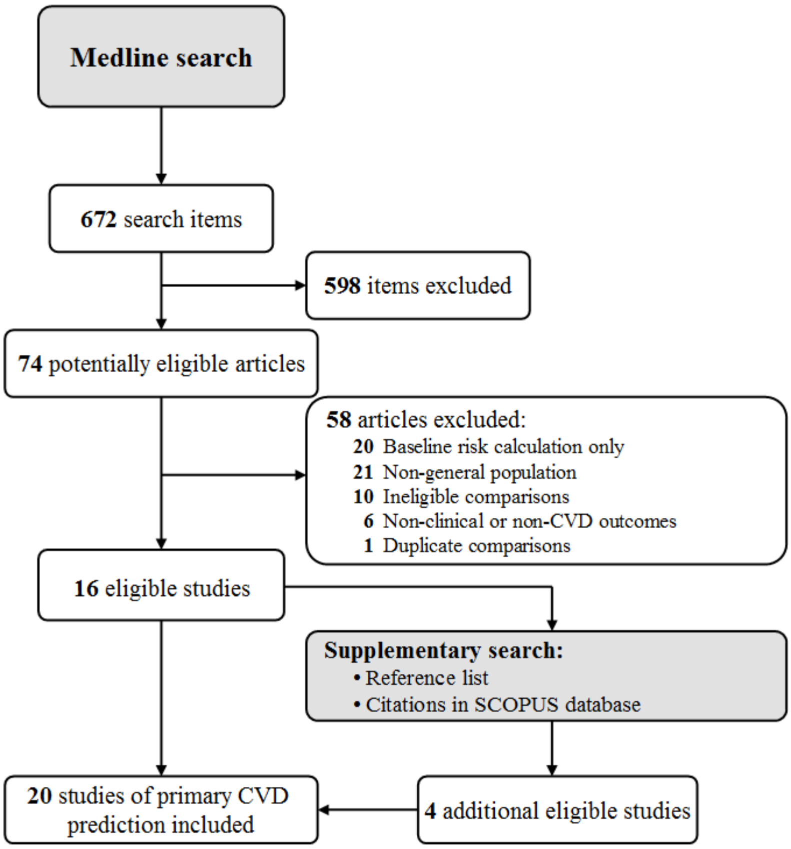
Medline search
672 search items
598 items excluded
74 potentially eligible articles
58 articles excluded:
20 Baseline risk calculation only
21 Non-general population
10 Ineligible comparisons
6 Non-clinical or non-CVD outcomes
1 Duplicate comparisons
16 eligible studies
Supplementary search:
• Reference list
• Citations in SCOPUS database
20 studies of primary CVD prediction included
4 additional eligible studies

Figure 3.1: Selection of eligible studies of risk models comparisons.

Characteristics of Eligible Studies and Risk Models

All articles were published after 2002 (**Table 3.2**). All but two[187,189] studies had prospective designs. Most (n=17) articles assessed populations of European descent. The median sample size was 8958 (interquartile range 2365-327 136).

Table 3.2: Characteristics of the included studies.

First author	Year	Data collection period	Study design	Study population	Sample size (men/women)	Models	Outcome(s)	Events (men/women)
Pandya et al.[189]	2011	1988-1994	retrospective	NHANES III cohort	5999 (3501/2498)	FRS, FRS (CVD)[d], SCORE (low and high risk)	CVD mortality	176 (118/58)
de la Iglesia et al.[188]	2011	1995-2006	prospective	THIN cohort	1072289 (529506/ 542783)	FRS, FRS (CVD)[d], ASSIGN	CVD (MI, CHD, stroke, TIA)	44375 (26202/18173)
Barroso et al.[187]	2010	ND	retrospective	Cohort in Spain	608 (263/345)	FRS, SCORE	CHD (angina, fatal and non-fatal MI), CVD mortality	57 (41/16)
Collins et al.[27]	2010	1993-2008	prospective	THIN cohort	1583106 (785733/ 797373)	FRS, QRISK1, QRISK2	CVD (angina, MI, CHD, stroke, TIA)	71465 (42408/29057)
van der Heijden et al.[190]	2009	1989-1992	prospective	Cohort in Netherlands	1125 (509/616)[a]	FRS, SCORE	CHD, CHD mortality	108 (CHD), 27 (fatal CHD)
Chen et al.[191]	2009	2003-2005	prospective	Cohort in Australia	1998 (808/1190)	FRS, SCORE (low and high risk)	CVD mortality	62 (36/26)
Collins et al.[28]	2009	1995-2006	prospective	THIN cohort	1072800 (529813/ 542987)	FRS (CVD)[d], QRISK1	CVD (MI, CHD, stroke, TIA)	43990 (25963/18027)
Woodward et al.[192]	2009	1984-1987, 1989, 1992, 1995	prospective	SHHEC cohort	13060 (6509/6551)	FRS, ASSIGN	CVD mortality, CHD or cerebrovascular disease, CABG or PTCA	2626 (1634/992)

Scheltens et al.[193]	2008	1987-1992	prospective	Cohort in Netherlands	40316 (18814/21502)	FRS, SCORE	CVD mortality	256 (189/67)
Hippisley-Cox et al.[29]	2008	1993-2008	prospective	QRESEARCH cohort	750232 (374469/ 375763)[b]	FRS, QRISK1, QRISK2	CVD (CHD, stroke, TIA)	ND[e]
Hippisley-Cox et al.[31]	2007	1995-2007	prospective	QRESEARCH cohort	614553 (305140/ 309413)	FRS, QRISK1, ASSIGN	CVD (MI, CHD, stroke, TIA)	30812 (17705/13107)
Mainous et al.[194]	2007	1987-1989	prospective	ARIC study	14343 (6239/8104)	FRS, SCORE	CHD (MI, fatal CHD, cardiac procedure)	1108
Ridker et al.[26]	2007	1992-2004	prospective	WHS cohort	8158	FRS, Reynolds Risk Score	CVD (MI, ischemic stroke, coronary revascularization, CVD mortality)	262
Woodward et al.[183]	2007	1984-1987, 1989, 1992, 1995	prospective	Cohort in Scotland	13297 (6540/6757)	FRS, ASSIGN	CVD mortality, CHD or cerebrovascular disease, CABG or PTCA	1165 (743/422)
Störk et al.[195]	2006	ND	prospective	Cohort in Netherlands	403[c]	FRS, PROCAM	CVD and all-cause mortality	31[c]
Cooper et al.[196]	2005	ND	prospective	Cohort in United Kingdom	2732[c]	FRS, PROCAM	CHD	219[c]
Ferrario et al.[197]	2005	1982-1996	prospective	CUORE study	6865[c]	FRS, PROCAM	fatal and non-fatal major CHD	312[c]
Dunder et al.[198]	2004	1970-1973	prospective	Uppsala Longitudinal Study of Adult Men cohort	534[c]	FRS, PROCAM	fatal and non-fatal MI	116[c]

Empana et al.[199]	2003	1991-1993	prospective	PRIME cohort (Belfast cohort)	2399[c]	FRS, PROCAM	CHD (angina, fatal CHD, MI)	120[c]
	2003	1991-1993	prospective	PRIME cohort (France cohort)	7359[c]	FRS, PROCAM	CHD (angina, fatal CHD, MI)	197[c]
Assmann et al.[185]	2002	1979-1985	prospective	PROCAM cohort	5389[c]	FRS, PROCAM	MI or CHD mortality	325[c]

ND, no data; NHANES III, National Health and Nutrition Examination Survey III; THIN, The Health Improvement Network; SHHEC, Scottish Heart Health Extended Cohort Study; ARIC, Atherosclerosis Risk In Communities; WHS, Women Health Study; PRIME, Prospective Epidemiological Study of Myocardial Infarction; PROCAM, Prospective Cardiovascular Münster; CVD, cardiovascular disease; FRS, Framingham risk score; ASSIGN-SCORE, Assessing Cardiovascular Risk to Scottish Intercollegiate Guidelines Network/SIGN to Assign Preventative Treatment score; SCORE, Systematic Coronary Risk Evaluation; PROCAM, Prospective Cardiovascular Münster score; MI, myocardial infarction; CHD, coronary heart disease; TIA, transient ischemic attack; CABG, coronary artery bypass graft; PTCA, percutaneous transluminal coronary angioplasty.

[a] Cohort subpopulation with normal glucose tolerance.

[b] Derived from validation cohort.

[c] Only males.

[d] Global FRS for total CVD prediction.[64]

[e] Data not shown, available from the corresponding author.

Eight different risk models were evaluated (all of those considered upfront eligible, except the World Health Organization/International Society of Hypertension score). Of the 28 possible types of pairwise comparisons of these eight risk scores, 14 existed in the literature. After excluding overlapping data (same models compared, same outcome, same cohort), independent data were available on 56 individual comparisons of risk models. Eight articles reported data for men and women separately (44 comparisons), four reported overall data (four comparisons), seven assessed only males (seven comparisons), and one assessed only women (one comparison, **Table 3.3**). The Framingham risk score or FRS (CVD) were involved in 50 of 56 comparisons (**Tables 3.2** and **3.3**). In four articles (eight comparisons) the authors reported information on missing data on the examined outcomes, and in all cases the investigators excluded the respective participants (**Table 3.4**). Information on missing data for variables included in risk models was reported in 11 articles (44 comparisons). Different strategies were implemented to deal with missing data and sometimes different strategies were applied to different predictors: exclusion of participants with missing data[27-29,31,190-192,198] (27 comparisons), multiple imputation technique[27,29,31] (16 comparisons), value generation by multivariate regression methods[189] (10 comparisons), replacement by the mean value of the variable[28,188,196] (nine comparisons), and assumption that participants without information on smoking were non-smokers[28,188] (eight comparisons, also see **Table 3.4**). In 25 comparisons, the geographical origin of the study population was the same as the origin of the population in which at least one of the examined models was initially developed (see **3.5.1 Supplementary table**).

Table 3.3: Discrimination performance according to the AUC metric.

First author	Year	Model	Outcome	AUC (95% CI)		
				Men	Women	Overall
Pandya et al.[189]	2011	FRS	CVD mortality	0.781 (0.738-0.823)	0.821 (0.766-0.876)	ND
		FRS (CVD)[d]		0.776 (0.733-0.819)	0.834 (0.782-0.885)	ND
		SCORE		Low risk: 0.785 (0.743-0.826) High risk: 0.785 (0.743-0.826)	Low risk: 0.792 (0.730-0.854) High risk: 0.792 (0.731-0.854)	ND
de la Iglesia et al.[188]	2011	FRS	CVD (MI, CHD, stroke, TIA)	0.740 (0.736-0.744)[c]	0.765 (0.761-0.769)[c]	ND
		FRS (CVD)[d]		0.752 (0.749-0.755)[c]	0.771 (0.767-0.775)[c]	ND
		ASSIGN		0.756 (0.753-0.759)[c]	0.792 (0.788-0.796)[c]	ND
Barroso et al.[187]	2010	FRS	CHD (angina, fatal and non-fatal MI), CVD mortality	-	-	0.70 (0.63-0.78)
		SCORE		-	-	0.86 (0.77-0.96)
Collins et al.[27]	2010	FRS	CVD (angina, MI, CHD, stroke, TIA)	0.75 (0.747-0.753)[c]	0.774 (0.771-0.777)[c]	ND
		QRISK1		0.771 (0.768-0.774)[c]	0.799 (0.796-0.802)[c]	ND
		QRISK2		0.773 (0.770-0.776)[c]	0.801 (0.798-0.804)[c]	ND
van der Heijden et al.[190]	2009	FRS	CHD, CHD mortality	ND	ND	0.68 (0.63-0.74)
		SCORE		ND	ND	0.71 (0.66-0.76)
		FRS		ND	ND	0.71 (0.61-0.82)
		SCORE		ND	ND	0.79 (0.70-0.87)
Chen et al.[191]	2009	FRS	CVD mortality	0.72 (0.65-0.80)	0.72 (0.64-0.80)	ND
		SCORE		Low risk: 0.75 (0.68-0.83), High risk: 0.75 (0.68-0.82)	Low risk: 0.70 (0.62-0.79), High risk: 0.70 (0.62-0.79)	ND

Collins et al.[28]	2009	FRS (CVD)[d]	CVD (MI, CHD, stroke, TIA)	0.752 (0.749-0.755)[c]	0.770 (0.766-0.774)[c]	ND
		QRISK1		0.762 (0.759-0.765)[c]	0.789 (0.785-0.793)[c]	ND
Woodward et al.[192]	2009	FRS	CVD mortality, CHD or cerebrovascular disease, CABG or PTCA	0.7183 (0.715-0.7213)	0.737 (0.733-0.741)	ND
		ASSIGN		0.7248 (0.722-0.728)	0.7618 (0.757-0.766)	ND
Scheltens et al.[193]	2008	FRS	CVD mortality	ND	ND	0.86 (0.84-0.88)
		SCORE		ND	ND	0.85 (0.83-0.87)
Hippisley-Cox et al.[29]	2008	FRS	CVD (CHD, stroke, TIA)	0.779 (0.776-0.782)	0.800 (0.797-0.803)	ND
		QRISK1		0.788 (0.786-0.791)	0.814 (0.811-0.817)	ND
		QRISK2		0.792 (0.789-0.794)	0.817 (0.814-0.820)	ND
Hippisley-Cox et al.[31]	2007	FRS	CVD (MI, CHD, stroke, TIA)	0.7598 (0.756-0.764)[c]	0.7744 (0.77-0.778)[c]	ND
		QRISK1		0.7674 (0.763-0.772)[c]	0.7879 (0.785-0.79)[c]	ND
		ASSIGN		0.7644 (0.760-0.769)[c]	0.7841 (0.78-0.787)[c]	ND
Mainous et al.[194]	2007	FRS	CHD (MI, fatal CHD, cardiac procedure)	0.691 (0.67-0.712)	0.808 (0.792-0.823)	ND
		SCORE		0.619 (0.597-0.641)	0.687 (0.668-0.705)	ND
Ridker et al.[26]	2007	FRS	CVD (MI, ischemic stroke, coronary revascularization, CVD mortality)	NA	0.787 (0.754-0.82)[c]	NA
		RRS		NA	0.808 (0.776-0.84)[c]	NA
Woodward et al.[183]	2007	FRS	CVD mortality, CHD or cerebrovascular disease, CABG or PTCA	0.716 (0.694-0.738)[c]	0.741 (0.72-0.762)[c]	ND
		ASSIGN		0.727 (0.706-0.748)[c]	0.765 (0.744-0.786)[c]	ND

Störk et al.[195]	2006	FRS	CVD and all-cause mortality	0.60 (0.49-0.69)	NA	NA
		PROCAM		0.55 (0.45-0.65)	NA	NA
Cooper et al.[196]	2005	FRS	CHD	0.62 (0.58-0.66)	NA	NA
		PROCAM		0.63 (0.59-0.67)	NA	NA
Ferrario et al.[197]	2005	FRS	fatal and non-fatal major CHD	0.723 (0.670-0.779)	NA	NA
		PROCAM		0.735 (0.678-0.790)	NA	NA
Dunder et al.[198]	2004	FRS	fatal and non-fatal MI	0.61 (0.55-0.67)[c]	NA	NA
		PROCAM		0.63 (0.57-0.69)[c]	NA	NA
Empana et al.[199]	2003	FRS[a]	CHD (angina, fatal CHD, MI)	0.66 (0.606-0.714)[c]	NA	NA
		PROCAM[a]		0.61 (0.555-0.665)[c]	NA	NA
		FRS[b]		0.68 (0.638-0.722)[c]	NA	NA
		PROCAM[b]		0.64 (0.598-0.682)[c]	NA	NA
Assmann et al.[185]	2002	FRS	MI or CVD mortality	0.778 (0.748-0.808)[c]	NA	NA
		PROCAM		0.824 (0.796-0.852)[c]	NA	NA

AUC, area under the curve; CI, confidence interval; FRS, Framingham risk score; ATP III, Adult Treatment Panel III; SCORE, Systematic Coronary Risk Evaluation; ASSIGN-SCORE, Assessing Cardiovascular Risk to Scottish Intercollegiate Guidelines Network/SIGN to Assign Preventative Treatment score; RRS, Reynolds Risk Score; PROCAM, Prospective Cardiovascular Münster Scoring System; CVD, cardiovascular disease; NICE, National Institute for Health and Clinical Excellence; MI, myocardial infraction; CHD, coronary heart disease; TIA, transient ischemic attack; CV, cardiovascular; ND, no data; NA, not applicable.

[a] North Ireland cohort

[b] France cohort

[c] Confidence intervals calculated as desribed in reference 24.

[d] Global FRS for total CVD prediction.[64]

Table 3.4: Reporting and management of missing data.

First author	Outcome(s)		Variables included in risk models	
	Missing data	**Management of missing data**	**Missing data**	**Management of missing data**
Pandya et al.[189]	No/Not reporting	NA	currently smoker, history of diabetes, systolic blood pressure, total cholesterol, high density lipoprotein, body-mass index	Independent draws from predictive distributions generated by using multivariate regression methods
de la Iglesia et al.[188]	No/Not reporting	NA	smoking, systolic blood pressure, total cholesterol, high density lipoprotein, body-mass index	missing data on smoking status: the patient was assumed to be a non-smoker other missing values: replaced by the mean for the sex and age-band (5 years bands)
Barroso et al.[187]	No/Not reporting	NA	Not reporting	NA
Collins et al.[27]	No/Not reporting	NA	Townsend scores (social deprivation), smoking status, systolic blood pressure, total serum cholesterol:high density lipoprotein ratio, body mass index.	missing Townsend scores: these patients were excluded other missing values: multiple imputation technique

van der Heijden et al.[190]	Yes	Patients excluded from the study	Not specified which	Patients were excluded from the study
Chen et al.[191]	Yes	Patients excluded from the study	smoking, blood pressure, total cholesterol	Patients were excluded from the study
Collins et al.[28]	No/Not reporting	NA	Townsend scores (social deprivation), smoking status, systolic blood pressure, total serum cholesterol:high density lipoprotein ratio, body mass index.	missing Townsend scores: patients were excluded missing data on smoking status: patients assumed to be non-smokers other missing values: replaced with unpublished age-sex reference values from the QRESEARCH cohort
Woodward et al.[192]	No/Not reporting	NA	Not specified which	Patients were excluded from the study
Scheltens et al.[193]	Yes	Patients excluded from the study	Not reporting	NA
Hippisley-Cox et al.[29]	No/Not reporting	NA	Townsend scores (social deprivation), smoking status, systolic blood pressure, total serum cholesterol:high density lipoprotein ratio, body mass index.	missing Townsend scores: these patients were excluded other missing values: multiple imputation technique

Hippisley-Cox et al.[31]	No/Not reporting	NA	Townsend scores (social deprivation), others (not specified)	missing Townsend scores: these patients were excluded other missing values: multiple imputation technique
Mainous et al.[194]	No/Not reporting	NA	Not reporting	NA
Ridker et al.[26]	No/Not reporting	NA	No	NA
Woodward et al.[183]	No/Not reporting	NA	No	NA
Störk et al.[195]	No/Not reporting	NA	Not reporting	NA
Cooper et al.[196]	No/Not reporting	NA	Not specified which	Average values were used
Ferrario et al.[197]	Yes	Patients excluded from the study	No	NA
Dunder et al.[198]	No/Not reporting	NA	Not specified which	Patients were excluded from the study
Empana et al.[199]	No/Not reporting	NA	Not reporting	NA
Assmann et al.[185]	No/Not reporting	NA	Not reporting	NA

Discrimination Performance

Area under the receiver operating characteristic curve estimates were available for all 56 pairwise comparisons (**Table 3.3**). Confidence intervals were given for only 20 pairs and P values for the comparison of area under the receiver operating characteristic curve were available for only two comparisons (in a single study[183]). The relative difference between the area under the receiver operating characteristic curve estimates exceeded 5% in only 10 (18%) comparisons, but even these differences were inconsistent: compared with SCORE, the Framingham risk score was worse in two cases but better in another two; compared with PROCAM, the Framingham risk score was worse in one case but better in another three; finally, FRS (CVD) was worse than SCORE in two cases. Among the 50 comparisons that included variants of the Framingham risk score, in 37 (74%) the area under the receiver operating characteristic curve estimate was higher for the comparator model. Use of other discrimination metrics (D statistic, R^2 statistic, Brier score) was inconsistent. At least one of these metrics was available for 26 comparisons (see **3.5.2 Supplementary table**).

Calibration

Calibration performance was reported in 38 comparisons (see **3.5.3 Supplementary table**). Observed versus predicted ratio estimates were available for 23 comparisons and results were quite inconsistent. The Framingham risk score was better than FRS (CVD) in one comparison but worse in another. The Framingham risk score was worse than ASSIGN in two comparisons, SCORE in two, QRISK1 in five, and PROCAM in one comparison, but it was better than ASSIGN in two comparisons, PROCAM in two, and QRISK1 in one comparison. FRS (CVD) was worse than ASSIGN in two comparisons and QRISK1 in one comparison, but it was better than QRISK1 in another comparison. Finally, QRISK1 was better than ASSIGN in two comparisons. The 95% confidence intervals of the observed to predicted ratio were available in only two comparisons, so we could not tell whether differences were beyond chance.

Risk Reclassification

Reporting of risk classification and reclassification was uncommon; information was available for 10 comparisons. In nine comparisons a dichotomous cut-off point of 20% 10 year risk was used; one study used 0-5, 5-10, 10-20, >20% as risk thresholds. All comparisons reported the number of participants reclassified with use of alternative models along with the predicted and observed risk in each risk category. The net reclassification index was calculated for six comparisons between non-nested models, all using the 20% threshold: ASSIGN versus Framingham risk score (n=2, net reclassification index 4%, 16%), ASSIGN versus FRS (CVD) (n=2, 0%, 12%), and FRS (CVD) versus Framingham risk score (n=2, 4% for both).

Outcome Selection Bias

In 13 comparisons the examined outcome was the one for which both compared models had been developed and validated, whereas in 32 comparisons only one of the compared models had been originally developed for that outcome, and in the other 11 comparisons none of the compared models had been developed originally for that outcome. When an outcome was used that had been used in the original development of only one of the compared models, it was more common for the outcome-congruent model to have a better area under the receiver operating characteristic curve than the comparator (25 versus 7, P<0.001, based on point estimates).

Optimism Bias

Five articles[26,29,31,183,185] (12 comparisons) described a model for the first time **Table 3.5**. In all 12 comparisons, the new model had a higher area under the receiver operating characteristic curve estimate than Framingham risk score versions, although the relative improvement exceeded 5% only for one model[185] (PROCAM better than Framingham risk score). Ten subsequently published articles addressed one or more of these same comparisons (**Table 3.5**). In three[29,31,192] articles at least one of the authors had been previously involved in the development of one of the compared models, and that model continued to have a better area under the receiver operating characteristic curve. Conversely, two[195,199] of the seven[27,188,195-199] articles published by entirely independent authors showed the older model to have a better area under the receiver operating characteristic curve.

Table 3.5: Potential optimism bias.

	First description of a model			Subsequent comparisons[a]	
First author	Model	Comparator	Performed better than the comparator(s)[a]	Involving some of the same authors	Involving independent authors
Hippisley-Cox et al.[29]	QRISK2	FRS, QRISK1	Yes	None	QRISK2>FRS and QRISK1 [27]
Hippisley-Cox et al.[31]	QRISK1	FRS, ASSIGN	Yes	QRISK1>FRS[29]	QRISK1>FRS[27]
Ridker et al.[26]	RRS	FRS	Yes	None	None
Woodward et al.[183]	ASSIGN	FRS	Yes	ASSIGN>FRS[31,192]	ASSIGN>FRS[188]
Assmann et al.[185]	PROCAM	FRS	Yes	None	PROCAM<FRS[195,199]; PROCAM>FRS[196-198]

RRS, Reynolds Risk Score; ASSIGN, Assessing Cardiovascular Risk to Scottish Intercollegiate Guidelines Network/SIGN to Assign Preventative Treatment score; PROCAM, Prospective Cardiovascular Münster Scoring System; FRS, Framingham risk score.

[a] Better performance of models in comparisons is based on AUC point estimates.

Author Interpretation

Overall, the authors claimed superiority of one model in 31 of 56 comparisons (**3.5.1 Supplementary table**). In 25 of these 31 comparisons a Framingham risk score version was one of the models compared and in all 25 cases the comparator model was claimed to be superior: SCORE>Framingham risk score (n=3), ASSIGN>Framingham risk score (n=6), PROCAM>Framingham risk score (n=1), QRISK1>Framingham risk score (n=4), QRISK2>Framingham risk score (n=4), FRS (CVD)>Framingham risk score (n=2), ASSIGN>FRS (CVD) (n=2), QRISK1>FRS (CVD) (n=2), and Reynolds risk score>Framingham risk score (n=1). The other six pairs where superiority was claimed were QRISK2>QRISK1 (n=4) and QRISK1>ASSIGN (n=2). For 22 comparisons the authors either claimed that both models had good or equal discriminatory ability or did not comment on their relative performance. In eight articles the authors favoured models they had themselves developed (five first publications, three subsequent publications). Authors involved in the development of a model never favoured a comparator.

3.4 DISCUSSION

Comparative studies on the relative performance of established risk models for prediction of cardiovascular disease often suggest that one model may be better than another. In particular, the Framingham risk score usually had inferior performance compared with other models newly developed, but the results were sometimes inconsistent across studies, and inferences may be susceptible to potential biases and methodological shortcomings. Most studies did not compare statistically the models that they examined. Models were usually reported to be superior against comparators when the examined outcome was the one that the model was developed for but not the one for which the comparator was developed. Articles presenting new models or including authors involved in the original development of a model favoured the model that the authors had developed.

Comparison with other studies

Head to head comparisons of emerging risk models are important to perform so as to document improvements in risk prediction. We showed that such data are limited and, when available, difficult to interpret. Discrimination, the ability of a statistical model to distinguish those who experience cardiovascular disease events from those who do not, was presented for all comparisons but the differences were usually small. Only in 18% of the comparisons did the relative difference between the two areas under the receiver operating characteristic curve exceed 5%. Most studies did not report the confidence intervals of the area under the receiver operating characteristic curve or the P values for the comparison between models. Calibration, which assesses how closely predicted estimates of absolute risk agree with actual outcomes, was reported in two thirds of the comparisons, but again formal statistical testing was lacking. Although the area under the receiver operating characteristic curve is the most commonly used discrimination metric, it has limitations.[164] Similarly, assessment of model calibration by the Hosmer-Lemeshow goodness of fit test is sensitive to sample size and gives no information on the extent or direction of miscalibration.[200,201] Evaluating calibration graphically either by 10ths of predicted risk or by key prognostic variables, such as age, is more informative than a single P value.

Assessment of risk reclassification was sparse and, when assessed, it was sub-optimally described, in agreement with previous empirical evaluations.[166,202] Reclassification is a clinically useful concept. It makes most sense when the categories of risk are clearly linked to different indications for interventions. It may be informative to report the percentage of patients changing risk categories and their direction of change. However, summary metrics such as the net reclassification index are problematic, especially when the compared models are non-nested (that is, they include different predictors and are derived from different datasets), and the problems are even worse when at least one model is poorly calibrated.[150]

Choices of comparators and outcomes are particularly important in such studies. Models were often claimed to be superior when the outcome examined was different from what the comparator model had been developed for. In those cases, the comparator is disadvantaged and becomes a strawman comparator towards which superiority can easily be claimed; a phenomenon analogous to that observed in clinical trial studies where an intervention is compared against a placebo or ineffective intervention.[203] In addition, we observed some evidence of potential optimism bias, with potentially unwarranted belief in the predictive performance of newer models[204] by the scientists developing them. Authors consistently claimed superiority of the models that they have developed versus comparators. While genuine progress in predictive ability is a possible explanation for this pattern, it is worthwhile to ensure that such favourable results are also validated by completely independent investigators.

Limitations

Some limitations need to be mentioned. Firstly, most of the analysed studies and models pertained to populations of European descent. Risk models may, however, perform differently in populations of different racial or ethnic backgrounds.[143,205] Systematic efforts for model validation in other populations are essential.[206] Secondly, most confidence intervals of area under the receiver operating characteristic curve estimates were unavailable and were derived as previously described.[92] We examined whether 95% confidence intervals did or did not overlap. A more formal statistical testing would have required access to individual level data to account for the fact that models were evaluated in the same population in each comparison using the pairwise individual level correlation in the calculations.[207]

Conclusions

- Direct comparisons of the most established risk prediction models for CVD are few.
- Studies that suggest one model is better than another often have potential biases and methodological shortcomings. Current studies comparing predictive models often have limitations or are missing information, which makes it difficult to reach robust conclusions about the best model or the ranking of performance of models. It should also be acknowledged that the

answers to these questions may be different in different populations and settings. The box shows several items and pieces of information that would be useful to consider in the design and reporting of results in studies comparing different predictive models to make these evaluations more useful, unbiased, and transparent, and to allow a balanced interpretation of the relative performance of these models.

- The Framingham risk score may often require recalibrating when used as a comparator. In many of the studies examined in our evaluation a new model was compared against the Framingham risk score. Although the Framingham risk score—developed in the United States during the 1970s—has stood the test of time, it has been shown to be miscalibrated in several other settings.[208] It is not surprising that without recalibration comparisons against it will often favour the new model, especially if the validation dataset covers specific subpopulations that were not covered in the original Framingham study.
- There is a lack of consistency between studies that compare prediction models because different statistical measures are used to describe the performance of the models. Statistical properties such as discrimination and calibration are widely recommended characteristics to evaluate; yet calibration is rarely examined. As important as the statistical characteristics of the model are, they do not ensure its clinical usefulness. There should therefore be more emphasis on demonstrating net benefit, for example[209], or, preferably, on conducting a randomised trial to evaluate the model's ability to change clinicians' decision making and patient outcomes.[93,105]
- The clinical usefulness of these models should be ultimately established on the basis of their potential for affecting decisions on treatment and prevention and improving health outcomes.[210] Ideally, this would require randomised trials where patients are allocated to being managed using information from different predictive models. Given that such trials are difficult to perform and costly, evidence from well conducted studies of comparative predictive performance will remain important. Our empirical

evaluation suggests that such studies may benefit from using standardised reporting of discrimination, calibration, and reclassification metrics with formal statistical comparisons; and standardised outcomes that are clinically appropriate and, whenever possible, relevant to both compared models. Finally, improved performance of new models versus established ones should ideally be documented in several studies carried out by independent investigators.

Box: Proposed items and pieces of information that would be useful to be considered in the design and reporting of results in studies comparing different predictive models.

Suggestions for studies comparing risk prediction models

- Comparative studies should be carried out in independent samples from those where each model was originally developed, and ideally by investigators other than those who originally proposed these models.
- The study setting, country, and type of population should be described; it should also be recognised whether these characteristics are expected to offer any clear advantage to one of the compared models.
- The main outcome of the study should be clearly defined and clinically relevant; it should be recognised that models originally developed to predict other outcomes may exhibit inferior predictive performance.
- Models should be calculated using the same exact predictors and coefficients as when they were originally developed and validated.
- The follow-up time should correspond to the same follow-up as when the models were developed (for example, 10 year risk); deviations should be clarified and an explanation about choice given.
- The discrimination of each model should be given with point estimates and confidence intervals; differences between the discrimination of compared models should be formally tested, reporting the magnitude of the difference and the accompanying uncertainty.
- The calibration of each model may be assessed with statistical tests, but there is no good formal test for comparing calibration performance; it is useful to also show graphically the expected versus predicted risk for different levels of risk or levels of predictors.
- Examination of reclassification performance of examined risk scores is meaningful when there are well established clinically relevant risk thresholds; it is useful to provide information on the number of correct and incorrect classifications; avoid using the net reclassification improvement for non-nested models.
- The extent of missing information for outcomes and predictors should be described, also explaining how missing information was handled.

3.5 SUPPLEMENTARY MATERIAL

3.5.1 Supplementary table: Potential biases and authors’ comments on the performance of risk models.

Models	Number of available comparisons	Geographic origin of cohort(s) [a]	Comparisons potentially affected by		Authors claim		Authors’ comment
			Optimism bias [b]	Outcome selection [b]	First model is superior	Second model is superior	
FRS vs. SCORE	14	**Australia**; **Netherlands**; **Spain**; **USA**	0	0	0	3	“…conclusion that the SCORE model provides a more accurate prediction than the Framingham one… we conclude that the former should be chosen over the latter to categorize the risk of cardiovascular …”[187], “…The use of the Framingham function for prediction of the first CHD event is likely to overestimate an individual's absolute CHD risk. In CHD prevention, application of the SCORE and UKPDS functions might be useful in the absence of a more valid tool…”[190], “…The findings of this study show that both the SCORE and the Framingham model function have a good discriminative ability but are insufficient in predicting absolute risks… ”[193], “… Every score discriminated risk of CVD death well… We observed strong agreement in risk characterization between the non-laboratory based and laboratory-based scores, and that all scores performed well in discriminating 10-year risk of CVD death in an external validation cohort (the NHANES III population)… NO FURTHER COMMENT…”[189]. No comment on FRS vs. SCORE”[191,194].
FRS vs. ASSIGN	8	**UK**	2 (out of 6)	4 (out of 6)	0	6	“…ASSIGN showed better discrimination for both men and women with recorded family history who appear to be at a much higher risk of the disease according to the K-M incidence.”[188], “…The slightly higher AUCs found for ASSIGN than for Framingham in this study were, as expected, due to socioeconomic status being accounted for in ASSIGN only…”[192], “…Our analysis shows that neither the Framingham nor ASSIGN equations is well calibrated for this UK population, with both scores tending to over-predict risk.”[31], “…The ASSIGN score receiver operating characteristic area under the curve was significantly (but marginally) higher than the Framingham equivalent in both sexes…”[183].

FRS vs. PROCAM	7	**France**; **Germany**; **Italy**; **Netherlands**; **Sweden**; **UK**	1 (out of 1)	0	0	1	“…Both the Framingham Score and the PROCAM Risk Function had no discriminatory power when applied to our cohort…”[195], “…the present study shows that while the use of Framingham and possibly PROCAM risk functions may be suitable for ordering individuals according to their estimated CHD absolute risk, their use seems inappropriate to estimate CHD absolute risk of healthy middle-aged men from low risk (France) and high-risk (Belfast) populations since it leads to a clear overestimation.”[199], “…The area under the ROC curve derived by use of the Framingham score (77.8%) was significantly less than that achieved with either the PROCAM Cox model (82.9%) or the PROCAM score (82.4%, P<0.001 for both comparisons)…”[185] “…No comment on FRS vs. PROCAM”[197,198].
FRS vs. QRISK1	4	**UK**	0	4 (out of 4)	0	4	“In this large cohort of 1.6 million patients, the NICE Framingham equation had inferior performance compared with either QRISK2 or its predecessor, QRISK1.”[27], “…The QRISK2 algorithm, like its predecessor, has better calibration and is a better discriminator of risk of cardiovascular disease than the modified Framingham score…”[29].
FRS vs. QRISK2	4	**UK**	2 (out of 4)	4 (out of 4)	0	4	“…We have assessed the performance of QRISK2 against the NICE version of the Framingham equation and have provided evidence to support the use of QRISK2 in favour of the NICE Framingham equation.”[27], “…The QRISK2 algorithm, like its predecessor, has better calibration and is a better discriminator of risk of cardiovascular disease than the modified Framingham score…”[29].
FRS vs. FRS(CVD)	4	**USA**; **UK**	0	0	0	2	“… Every score discriminated risk of CVD death well … We observed strong agreement in risk characterization between the non-laboratory based and laboratory-based scores, and that all scores performed well in discriminating 10-year risk of CVD death in an external validation cohort (the NHANES III population)… NO FURTHER COMMENT …”[189], “… Generally, Anderson Framingham made worse predictions than ASSIGN and Cox Framingham… ”[188].
QRISK1 vs. QRISK2	4	**UK**	2 (out of 4)	0	0	4	“…The difference in performance between QRISK2 and QRISK1 was slight, with QRISK2 marginally outperforming QRISK1.…”[27], “…The QRISK2 model was marginally superior to the original QRISK1 equation”[29]**.**

FRS(CVD) vs. SCORE	4	**USA**	0	0	0	0	“… Every score discriminated risk of CVD death well… We observed strong agreement in risk characterization between the non-laboratory based and laboratory-based scores, and that all scores performed well in discriminating 10-year risk of CVD death in an external validation cohort (the NHANES III population)… NO FURTHER COMMENT …”[189].
FRS(CVD) vs. ASSIGN	2	**UK**	0	0	0	2	“…ASSIGN showed better discrimination for both men and women with recorded family history who appear to be at a much higher risk of the disease according to the K-M incidence…”[188].
FRS(CVD) vs. QRISK1	2	**UK**	0	2 (out of 2)	0	2	“…the QRISK model gives a more accurate estimate of predicted risk compared with either Framingham equation…”[28].
QRISK1 vs. ASSIGN	2	**UK**	2 (out of 2)	0	2	0	“…QRISK performed at least as well as the Framingham model for discrimination and was better calibrated to the UK population than either the Framingham model or ASSIGN… ”[31].
FRS vs. RRS	1	**USA**	1 (out of 1)	1 (out of 1)	0	1	“…We developed, validated, and demonstrated highly improved accuracy of 2 clinical algorithms for global cardiovascular risk prediction that reclassified 40% to 50% of women at intermediate risk into higher- or lower-risk categories…”[26].

FRS, Framingham risk score; SCORE, Systematic Coronary Risk Evaluation; ASSIGN, Assessing Cardiovascular Risk to Scottish Intercollegiate Guidelines Network/SIGN to Assign Preventative Treatment score; PROCAM, Prospective Cardiovascular Münster Scoring System; RRS, Reynolds Risk Score; CVD, cardiovascular disease.

[a] Geographic origin of cohorts used in each comparison pair.
[b] Among comparisons where one model is claimed to be superior to the other.

3.5.2 Supplementary table: Discrimination performance according to metrics other than the AUC.

First author	Model	*D* statistic (95% CI)		R^2 statistic (95% CI)		Brier score (95% CI)	
		Men	Women	Men	Women	Men	Women
de la Iglesia et al.[188]	FRS	1.26 (1.24-1.28)	1.39 (1.36-1.41)	27.57 (27.1-28.1)	31.5 (30.9-32.2)	0.0536	0.0335
	FRS (CVD)[a]	1.32 (1.30-1.34)	1.41 (1.39-1.44)	29.52 (29-30.2)	32.37 (31.6- 33)	0.0535	0.0334
	ASSIGN	1.35 (1.33-1.37)	1.58 (1.56-1.60)	30.5 (29.8-31.2)	37.4 (36.7-37.9)	0.0517	0.0351
Barroso et al.[187]	FRS	ND	ND	ND	ND	0.1200	0.0396
	SCORE	ND	ND	ND	ND	0.0221	0.0079
Collins et al.[27]	FRS	1.30 (1.12-1.48)	1.47 (1.29-1.64)	28.7 (23.1-34.3)	33.8 (28.5-39.2)	0.08 (0.08-0.09)	0.05 (0.05-0.06)
	QRISK1	1.42 (1.28-1.55)	1.61 (1.50-1.71)	32.3 (28.3-36.4)	38.2 (35.1-41.3)	0.08 (0.07-0.08)	0.05 (0.05-0.05)
	QRISK2	1.45 (1.31-1.59)	1.66 (1.56-1.76)	33.3 (28.9-37.8)	39.5 (36.6-42.4)	0.08 (0.07-0.08)	0.05 (0.05-0.05)
Collins et al.[28]	FRS (CVD)[a]	1.33 (1.31-1.34)	1.41 (1.39-1.44)	29.5 (28.9-30.1)	32.3 (31.6-33.1)	0.0530	0.0330
	QRISK1	1.39 (1.38-1.41)	1.56 (1.53-1.58)	31.7 (31.1-32.3)	36.6 (35.9-37.3)	0.0470	0.0321
Hippisley-Cox et al.[29]	FRS	1.495 (1.47-1.52)	1.632 (1.61-1.66)	34.8 (34.1-35.5)	38.- (38.1-39.6)	0.18 (0.17-0.18)	0.09 (0.09-0.09)
	QRISK1	1.59 (1.568-1.61)	1.776 (1.75-1.80)	37.6 (36.9-38.3)	42.9 (42.2-43.7)	0.13 (0.13-0.13)	0.08 (0.08-0.08)
	QRISK2	1.62 (1.594-1.64)	1.795 (1.77-1.82)	38.4 (37.8-39.0)	43.5 (42.8-44.2)	0.14 (0.13-0.14)	0.09 (0.08-0.09)
Hippisley-Cox et al.[31]	FRS	1.31 (SE: 0.012)	1.39 (SE: 0.014)	29.1 (SE: 0.38)	31.7 (SE: 0.44)	ND	ND
	QRISK1	1.45 (SE: 0.013)	1.55 (SE: 0.014)	33.3 (SE: 0.39)	36.4 (SE: 0.43)	ND	ND
	ASSIGN	1.36 (SE: 0.012)	1.47 (SE: 0.014)	30.5 (SE: 0.38)	34.1 (SE: 0.43)	ND	ND

FRS, Framingham risk score; ASSIGN, Assessing Cardiovascular Risk to Scottish Intercollegiate Guidelines Network/SIGN to Assign Preventative Treatment score; SCORE, Systematic Coronary Risk Evaluation; NICE, National Institute for Health and Clinical Excellence; PROCAM, Prospective Cardiovascular Münster score; ATP III, Adult Treatment Panel III; ND, no data; SE, standard error.

[a] Global FRS for total CVD prediction.[64]

3.5.3 Supplementary table: Calibration metrics.

First author	Model	Predicted/Observed ratio		Other information
		Men	Women	
de la Iglesia et al.[188]	FRS	1.25	1.02	calibration plot
	FRS (CVD)[a]	1.25	1.04	calibration plot
	ASSIGN	1.20	1.20	calibration plot
Barroso et al.[187]	FRS	1.33	2.50	ND
	SCORE	1.30	1.55	ND
Collins et al.[27]	FRS	ND	ND	calibration plot (men/women)
	QRISK1	ND	ND	calibration plot (men/women)
	QRISK2	ND	ND	calibration plot (men/women)
van der Heijden et al.[190]	FRS	ND	ND	calibration plot (overall)
	SCORE	ND	ND	calibration plot (overall)
	FRS	ND	ND	calibration plot (overall)
	SCORE	ND	ND	calibration plot (overall)
Chen et al.[191]	FRS	ND	ND	H-L=11.60 (men) and 12.92 (women)
	SCORE	ND	ND	H-L=4.40 (men low risk) and 12.92 (women low risk); 32.78 (men high risk) and 27.25 (women high risk)
Collins et al.[28]	FRS (CVD)[a]	1.25	1.04	calibration plot
	QRISK1	0.87	0.90	calibration plot
Scheltens et al.[193]	FRS	ND	ND	H-L=64 (overall)
	SCORE	ND	ND	H-L=35 (overall)

Hippisley-Cox et al.[29]	FRS	ND	ND	calibration plot (men/women)
	QRISK1	ND	ND	ND
	QRISK2	ND	ND	calibration plot (men/women)
Hippisley-Cox et al.[31]	FRS	1.47	1.18	ND
	QRISK1	1.00	1.02	ND
	ASSIGN	1.35	1.38	ND
Ridker et al.[26]	FRS	NA	ND	H-L p value <0.001
	Reynolds Risk Score	NA	ND	H-L p value =0.62
Cooper et al.[196]	FRS	0.47	NA	ND
	PROCAM	0.46	NA	ND
Ferrario et al.[197]	FRS	ND	NA	H-L=27.1 & calibration plot
	PROCAM	ND	NA	H-L=220.3 & calibration plot
Empana et al.[199]	FRS[b]	1.34 (1.12-1.60)	NA	calibration plot
	PROCAM[b]	1.78 (1.38-2.28)	NA	calibration plot
	FRS[c]	2.35 (2.05-2.71)	NA	calibration plot
	PROCAM[c]	2.76 (2.28-3.34)	NA	calibration plot

FRS, Framingham risk score; ATP III, Adult Treatment Panel III; SCORE, Systematic Coronary Risk Evaluation; ASSIGN, Assessing Cardiovascular Risk to Scottish Intercollegiate Guidelines Network/SIGN to Assign Preventative Treatment score; PROCAM, Prospective Cardiovascular Münster Scoring System; CVD, cardiovascular disease; NICE, National Institute for Health and Clinical Excellence; MI, myocardial infraction; CHD, coronary heart disease; TIA, transient ischemic attack; CV, cardiovascular; ND, no data; NA, not applicable; H-L, Hosmer-Lemeshow statistic.

[a] Global FRS for total CVD prediction[9]
[b] North Ireland cohort
[c] France cohort

Section 4.

External validation of new risk prediction models: infrequent with worse prognostic discrimination.

4. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination.[22]

Siontis GC, Tzoulaki I, Castaldi PJ, Ioannidis JP.

Journal of Clinical Epidemiology. 2015. (PMID: 25441703)

Journal of Clinical Epidemiology 68 (2015) 25−34

Journal of Clinical Epidemiology

ORIGINAL ARTICLES

External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination

George C.M. Siontis[a,1], Ioanna Tzoulaki[a,b], Peter J. Castaldi[c], John P.A. Ioannidis[d,e,f,*]

[a]*Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, University Campus, P.O. Box 1186, 45110 Ioannina, Greece*

[b]*Department of Epidemiology and Biostatistics, Imperial College London, Norfolk Place W2 1PG, London, United Kingdom*

[c]*Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, 181 Longwood Avenue, Boston, MA 02115, USA*

[d]*Department of Medicine, Stanford Prevention Research Center, Stanford University School of Medicine, 1265 Welch Rd, MSOB X306, Stanford, CA 94305, USA*

[e]*Department of Health Research and Policy, Stanford University School of Medicine, Stanford, CA 94305, USA*

[f]*Department of Statistics, Stanford University School of Humanities and Sciences, Stanford, CA 94305, USA*

Accepted 4 September 2014; Published online 23 October 2014

Abstract

Objectives: To evaluate how often newly developed risk prediction models undergo external validation and how well they perform in such validations.

Study Design and Setting: We reviewed derivation studies of newly proposed risk models and their subsequent external validations. Study characteristics, outcome(s), and models' discriminatory performance [area under the curve, (AUC)] in derivation and validation studies were extracted. We estimated the probability of having a validation, change in discriminatory performance with more stringent external validation by overlapping or different authors compared to the derivation estimates.

Results: We evaluated 127 new prediction models. Of those, for 32 models (25%), at least an external validation study was identified; in 22 models (17%), the validation had been done by entirely different authors. The probability of having an external validation by different authors within 5 years was 16%. AUC estimates significantly decreased during external validation vs. the derivation study [median AUC change: −0.05 ($P < 0.001$) overall; −0.04 ($P = 0.009$) for validation by overlapping authors; −0.05 ($P < 0.001$) for validation by different authors]. On external validation, AUC decreased by at least 0.03 in 19 models and never increased by at least 0.03 ($P < 0.001$).

Conclusion: External independent validation of predictive models in different studies is uncommon. Predictive performance may worsen substantially on external validation. © 2015 Elsevier Inc. All rights reserved.

Keywords: Risk prediction model; Prognostic models; External validation; Discrimination; Area under the receiver operating characteristics curve; Derivation study

1. Introduction

Risk prediction models can be useful tools to guide clinical decision making, including treatment selection and patient counseling. Numerous such models are constantly being developed in the medical literature; however, very few of them are actually used in clinical practice [1]. Some models are only described once and never used afterward in subsequent publications. For example, there are 94 models to assess risk of incident diabetes in the medical literature, and only 14 of those have been calculated again in subsequent publications [2].

Successful application of a risk prediction model requires validation in different populations (external validation) [3]. External validation may be done in different geographical areas, periods, and settings (eg, secondary vs. primary care), and this may involve the same authors or different authors. Moreover, external validation may be performed as part of the same article that describes the original development of the model, a different article by the same or overlapping authors, or by completely different teams in different investigations. These steps of increasing independence document that the model can

Conflict of interest: None.

Funding: This study was supported in part by an unrestricted gift by Sue and Bob O'Donnell to Stanford Prevention Research Center.

[1] Current address: Department of Cardiology, Bern University Hospital, Bern, Switzerland.

* Corresponding author. Tel.: 650-725-5465.

E-mail address: jioannid@stanford.edu (J.P.A. Ioannidis).

http://dx.doi.org/10.1016/j.jclinepi.2014.09.007

4.1 BACKGROUND

Risk prediction models, as have been highlighted above, can be useful tools to guide clinical decision making, including treatment selection and patient counseling. Numerous such models are constantly being developed in the medical literature; however, very few of them are actually used in clinical practice.[211] Some models are only described once and never used afterward in subsequent publications. For example, there are 94 models to assess risk of incident diabetes in the medical literature, and only 14 of those have been calculated again in subsequent publications.[212]

Successful application of a risk prediction model requires validation in different populations (external validation).[109] External validation may be done in different geographical areas, periods, and settings (eg, secondary vs. primary care), and this may involve the same authors or different authors. Moreover, external validation may be performed as part of the same article that describes the original development of the model, a different article by the same or overlapping authors, or by completely different teams in different investigations. These steps of increasing independence document that the model can perform well in diverse circumstances and in the hands of different investigators. Clinicians who trust the original claim of predictive ability of a new model that has not been externally validated may have an unrealistically optimal impression about how good the predictive tool is. For example, the Mortality Probability Models (MPM II) for the prediction of mortality in critically ill patients had very good discriminating ability when it was first developed [area under the curve (AUC), 0.837][213], but in a study published 16 years later by different investigators, its performance was very modest (AUC, 0.66) and could not compete with other models.[214]

Authors who develop a new risk prediction model using their data and then compare it with an existing model often report better performance for the new model. Prediction models tend to perform better on the dataset from which they were developed and usually, if not always, perform better than existing models when validated on that dataset. This is simply because the model is tuned to the dataset at hand, which is why a model's performance should be evaluated in other datasets, preferably by independent investigators. However, some form of reporting bias must play a role here,[215] because a newly developed prediction model that performed worse than an existing one would probably not be submitted or published. Greater emphasis should therefore be placed on methodologically sound and appropriately detailed external validation studies, ideally of multiple models at once, to show which model is most useful.[93]

Methodologists have long established the importance and implications of external validation of multivariate models.[50,102] However, it is unclear whether these practices are adhered in the literature, and lack of proper external independent validation may result in unrealistic expectations for the performance of these models. For example, for highly cited and popular single biomarkers[216,217], validation efforts in large studies have shown much smaller (or even null) effects compared with early studies. Comparisons of risk prediction models for cardiovascular disease have also shown allegiance or optimism bias: when the authors of the comparative study have developed one of the models, they report favorable results for their own model.[116] To our knowledge, there is no large-scale systematic evaluation on the performance of diverse proposed risk prediction models when these are tested in external independent validation efforts by the same or different teams than those who originally developed them. An evaluation of a large number of such studies is needed to get a sense of the external independent performance of such models because single models and validations may have results that are difficult to generalize.

Here, we aimed to perform an empirical evaluation on the external independent validation practices of risk prediction models. We aimed to evaluate how often external validations were performed, in particular by different authors than those who had developed the model. We also evaluated whether the estimates of model performance deteriorated substantially during external validation efforts by overlapping or different authors.

4.2 METHODS

Literature Search

Two different searches in ISI Thomson Web of Science database with the following keywords were adopted: Search A: Title = ("risk score" OR "risk model" OR "prognostic model" OR "prognostic score" OR "predictive model" OR "predictive score") AND Title = ("new" OR "novel") and search B. Topic = ("novel risk score" OR "novel risk model" OR "novel prognostic model" OR "novel prognostic score" OR "novel predictive model" OR "novel predictive score" OR "new risk score" OR "new risk model" OR "new prognostic model" OR "new prognostic score" OR "new predictive model" OR "new predictive score"). The search strategies did not aim to identify all newly developed risk prediction models but to generate a pool of articles that would be enriched in eligible studies where a new model was presented for the first time. Search was limited to derivation studies published until the end of 2010, so as to allow at least 2.5 years for the publication of subsequent external validation studies.

Eligibility Criteria for Derivation Studies of Risk Prediction Models

For this evaluation, we deemed eligible those original derivation studies that describe risk prediction models that are developed for the first time, built from a set of candidate predictors, pertain to biomedical application (eg, we excluded prediction models on economics), and include more than one variable (eg, excluding single biomarkers). We did not consider animal studies, reviews, editorials, letters, and studies not published in English.

Eligibility Criteria and Search Strategy for Subsequent Validation Studies

For each eligible derivation study, we searched the citations made to this study by subsequently published articles. These citations were retrieved from ISI Web of Science (search last updated on July 1, 2013), and among them, we identified citing articles that have claimed to validate the same model in different populations (validation studies) by either at least one author in common with the initial author group (overlapping author(s) validation) or completely different authors (different authors validation). As a validation study, we considered any study that claimed to apply the same model for the same disease and same outcome as the derivation study. The searched time span was "all years," and retrieved citations were limited to those classified as "articles" by ISI Web of Science. We searched the citations of each eligible derivation study starting with the first (oldest) citation that had received and moving to the newer ones in the chronological order of entry in the Web of Science. We identified and included the first (oldest) appeared validation study with at least one overlapping author(s) and one with completely different author group compared with the derivation study.

Data Extraction

From each eligible derivation study where the new model was built, we extracted the listed authors and year of publication, type of population in relation to nature of underlying disease (acute, chronic, or mixed), type of patients (cardiovascular diseases, gastrointestinal-related diseases, malignancies, or other), study sample size, the newly developed risk prediction model and the included variables of the proposed model, the examined outcome(s) for which the model was developed [mortality related (death or composite including also death) or other], and model's discriminatory ability (expressed as area under the receiver operating characteristic curve[92] or equivalently C index or C statistic) for every examined outcome. Any given calibration metric in derivation and validation studies was also recorded.

The derivation study publication could have assessed the discriminatory ability of a model (1) only directly in the training set (thus expected to suffer from optimism due to overfitting), (2) through some unbiased procedure (crossvalidation, split sample, and bootstrap) on the training set (unbiased in the sense that, if properly conducted, it does not suffer from training set overestimation of performance due to overfitting), and/or (3) in a totally different testing set. We preferred (3) over (2) over (1) to get the most unbiased and generalizable estimate of discriminatory ability. We also gave preference to keep information on the whole study population over subgroups. When only a subgroup of the derivation study was eligible for our analysis, we kept the respective information only for that sample. Finally, when different risk prediction models were developed for different outcomes and/or follow-up periods in the same study, we extracted the respective information for each model in separate.

From each eligible subsequent validation article (validation by overlapping or different authors) that cited an eligible derivation study and validated the newly proposed model, we recorded the listed authors and year of publication, type of population and disease/clinical condition, the included outcome(s), study sample size, and the discriminatory performance (based on AUC) of the validated model for any given outcome matched to the respective model in the derivation study.

Statistical Analysis

Descriptive statistics are presented as median and interquartile range (IQR) or absolute counts and percentages. When models for different outcomes and/or different follow-up periods were proposed/included in the same article, these were considered independently in the analyses. We noted how many of the eligible risk prediction models have been validated externally in subsequent publications and of those how many were validated by overlapping and/or by different authors. When the model's AUC performance was not given directly by the authors, we estimated it by using the available number of patients with an event or the probability of being event free at specified follow-up in each group of risk according to the evaluated model.

Kaplan-Meier plots assessed the probability of having a published validation study as a function of time from the publication of each newly proposed model. Time was censored at the time of citation search (July 2013). We further investigated the relationship between the time of any external independent validation and the year of publication of the derivation study. For each validated model, we examined whether the estimated AUC deteriorated with subsequent more stringent validation steps: derivation, overlapping, and different authors validation. In addition, we performed a sensitivity analysis for the probability of external independent validation limited to derivation studies published in the last decade (2003-2013). Derivation vs. validation AUCs were compared with paired t-test, and the frequency of AUC changes of 0.03 or more and 0.05 or more in either direction was also assessed with the sign test. The correlation of AUC between derivation and validation studies was evaluated with Spearman's correlation coefficient.

We also performed a sensitivity analysis in which we also counted as first external independent validations by overlapping authors those validations that were part of the derivation studies but had used entirely different data sets than the training data set (not just splitting the same sample but using data sets from different time and/or location of recruitment). All analyses were performed in Stata software, version 12.0 (StataCorp, College Station, TX, USA). P-values are two tailed.

4.3 RESULTS

Eligible Studies

The study selection process of eligible derivation studies and identification of newly proposed models is illustrated in **Figure 4.1**. Search A: Seventy-nine potentially eligible studies were evaluated in full-text for eligibility. Of those, 11 studies were excluded for the following reasons: non-multivariable prediction tool (n=3), not in-vivo model (n=3), based on previously developed models (n=2), study not included in biomedicine field (n=1), study not in English (n=1) and study without any prognostic impact (n=1). Finally, 68 derivation studies each proposing newly introduced models were identified and included in our analysis. Search B: Ninety studies were initially evaluated for eligibility, while only 20 studies were finally deemed eligible. Seventy studies were excluded: studies in duplicate (Search A) (n=30), nonmultivariable prediction tool (n=3), previously developed models (n=18), study written not in English (n=1), study without any prognostic impact (n=16), and studies not available in full text (n=2). In total, 88 derivation studies proposing 127 new prediction models were deemed eligible and included in our analysis. Of the 127 risk prediction models, 95 (75%)

had no subsequently published external validation study (S1-S66 listed in **4.5.1 Supplementary references**) (**4.5.2 Supplementary table**). Of the remaining 32 models[12,31,218-241] (**4.5.3 Supplementary table**), 10 models[31,218,220,221,227,241] had been validated in subsequent publications only by overlapping author(s)[30,242-245], 16 models[219,222-224,228-237,240] only by different authors[246-260], and only 6 models[12,225-227,238,239] had subsequent validation publications by both overlapping[261-266] and different authors[267-272] (**4.5.4 Supplementary table**).

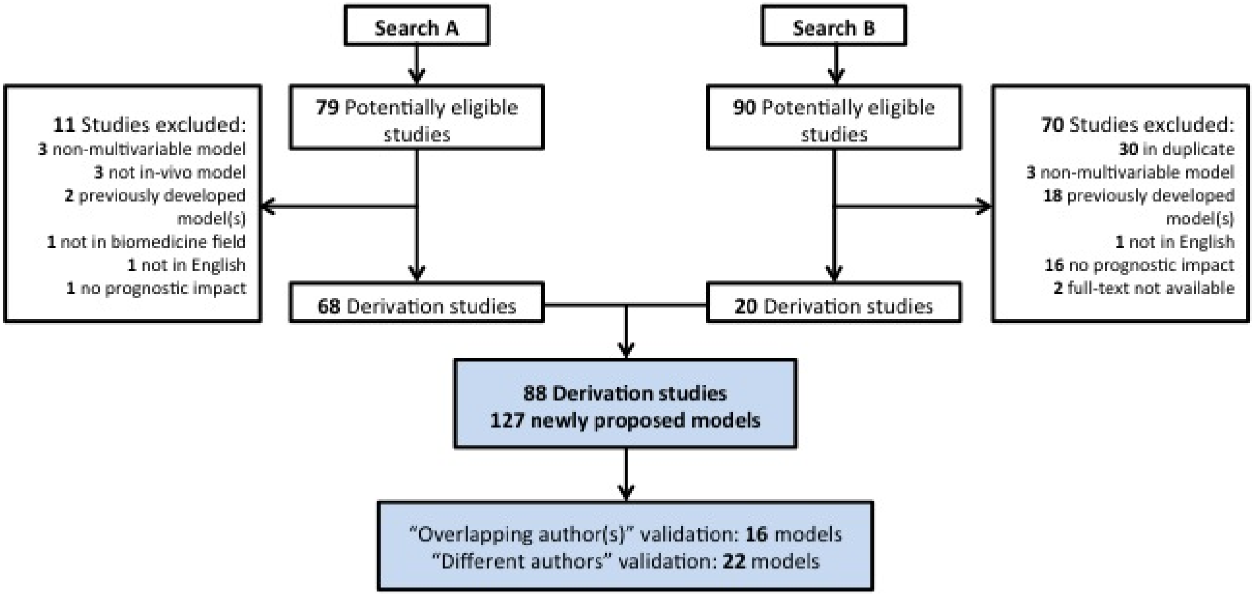
Search A
Search B
79 Potentially eligible studies
90 Potentially eligible studies
11 Studies excluded:
3 non-multivariable model
3 not in-vivo model
2 previously developed model(s)
1 not in biomedicine field
1 not in English
1 no prognostic impact
70 Studies excluded:
30 in duplicate
3 non-multivariable model
18 previously developed model(s)
1 not in English
16 no prognostic impact
2 full-text not available
68 Derivation studies
20 Derivation studies
88 Derivation studies
127 newly proposed models
"Overlapping author(s)" validation: 16 models
"Different authors" validation: 22 models

Figure 4.1: Selection of eligible derivation studies proposing new risk prediction models and their subsequent validation studies through two different searches.

Characteristics of Derivation and Validation Studies

Derivation studies with or without subsequent validation(s) were published between 1973 and 2010 (median 2007; **Table 4.1**, **4.5.2** and **4.5.3 Supplementary table**). Study populations included both acute (36%) and chronic (53%) settings. Most models pertained to malignancies or cardiovascular diseases. Most studies examined only one outcome (76%). Death or composite outcomes including death were chosen in 36% of the studies. The median sample size (IQR) was 445 (153 - 1,127). Calibration metrics were suboptimally reported (**Table 4.1**, **4.5.2** and **4.5.3 Supplementary table**).

Table 4.1: Characteristics of derivation studies proposing new predictive models and of subsequent validation studies.

Characteristic	Derivation studies (n=88)	Validation studies Overlapping author(s) (n=11)	 Different authors (n=20)
Year of publication [range, median (IQR)]	1973-2010 2007 (2003-2009)	1982-2012 2008 (2006-2011)	1987-2013 2009 (2005-2011)
Type of population (n (%))			
Acute	32 (36)	5 (45)	10 (50)
Chronic	47 (53)	5 (45)	9 (45)
Other	9 (11)	1 (10)	1 (5)
Disease/Clinical condition (n (%))			
Cardiovascular diseases	19 (22)	4 (36)	5 (25)
GI-related diseases	12 (14)	1 (9)	2 (10)
Malignancies	36 (41)	4 (36)	7 (35)
Other	21 (23)	2 (18)	6 (30)
No. of outcomes (n (%))			
One	67 (76)	8 (73)	16 (81)
Two	12 (14)	2 (18)	3 (14)
Three	6 (7)	1 (9)	1 (5)
Four	2 (2)	0	0
Five	1 (1)	0	0
Type of outcome (n (%))			
Mortality-related	32 (36)	3 (27)	9 (45)
Other	56 (64)	8 (73)	11 (55)
Sample size [median (IQR)]	445 (153-1127)	340 (173-7329)	190 (70-345)
Calibration metrics	24 (27)	3 (27)	1 (5)
Hosmer-Lemeshow statistic	15 (17)	1 (9)	1 (5)
Predicted/Observed ratio	9 (10)	2 (18)	0

Publications of validation studies by overlapping authors appeared between 1982 and 2012 (**4.5.2 and 4.5.4 Supplementary table**). The median (IQR) sample size was 340 (173 - 7,329). Also, 20 external validation studies by different authors were identified between 1987 and 2013, all of them published later than validation studies by overlapping author(s) (**4.5.2** and **4.5.4 Supplementary table**). External validation studies by different authors were usually small [median (IQR) of 190 (70 - 345)]; in 11 of 22 cases, they were at least five times smaller than the sample size of population where the model was first derived.

Probability of Model Validation

As shown in **Figure 4.2A**, the probability of a newly introduced model to be validated in subsequent publications by any author group at 2, 5, and 10 years was 13%, 25%, and 38%, respectively. Five years after the publication of the derivation study, the probability of having a validation by overlapping author(s) was 9% (**Figure 4.2B**) and the probability of having a validation by different authors was 16% (**Figure 4.2C**). No validations occurred more than 10 years after the derivation of a model. When we focused only on derivation studies published during the last decade (2003 - 2013), the probability of validation of a new model by any author group at 2, 5, and 10 years remained low (14%, 28%, and 34%, respectively). Finally, the year of derivation study publication was not associated with subsequent validation of the newly proposed model (hazard ratio 1.02 per year, P = 0.53).

Discrimination Performance in Subsequent Validations

Derivation study AUCs were given or inferred (for eight models) in 76 (60%) risk prediction models [training set (n = 29); cross-validation, split sample, or bootstrap method on the training test (n = 20); and testing set (n = 27)]. AUC metrics were available in 14 of 16 (88%; inferred for five models) for overlapping author validations and 17 of 22 (77%; inferred for six models) for different author validations. AUC was lower with subsequent validations. AUC was higher in derivation than the external validation study in 11 of 14 cases when the external validation involved overlapping author(s) (median AUC change: −0.04 lower in external validation, P = 0.009 by paired t-test); in 14 of 17 cases when only different authors were involved (median AUC change: −0.05 lower in external validation, P < 0.001 by paired t-test); and 25 of 31 cases overall (median AUC change: −0.05 lower in external validation, P < 0.001 by paired t-test). There were very sparse data on comparisons of overlapping author vs. different author validations of the same model.

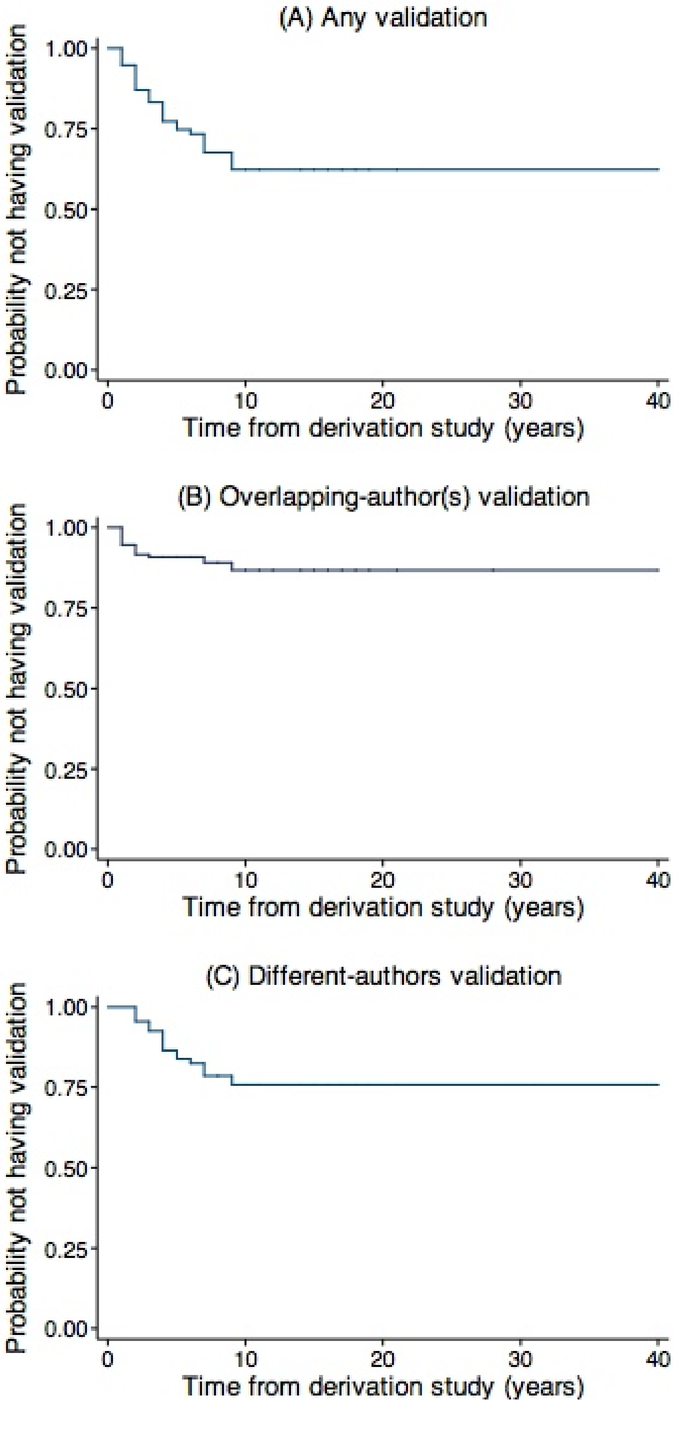

Figure 4.2: Time to validation. (A) Any validation. (B) Validation by overlapping author(s). (C) Validation by different authors.

When considering only the testing set and unbiased estimates in the training set, the AUC during external validation in a subsequent study decreased on average by 0.062 (P < 0.001 by paired t-test). The decrease in AUC from derivation to any subsequent validation exceeded 0.03 in 19 cases, whereas increase of such magnitude was not seen (P < 0.001 by the sign test). Decreases exceeding 0.05 in the AUC were seen in 15 cases, whereas no increase of such magnitude was seen in any case (P < 0.001 by the sign test). Derivation AUCs were strongly correlated with AUCs in subsequent validation steps [$\rho_{\text{derivation/overlapping author validation}} = 0.70$ (P = 0.006), $\rho_{\text{derivation/different author validation}} = 0.69$ (P = 0.002), and $\rho_{\text{derivation/any validation}} = 0.72$ (P < 0.001); **Figure 4.3**].

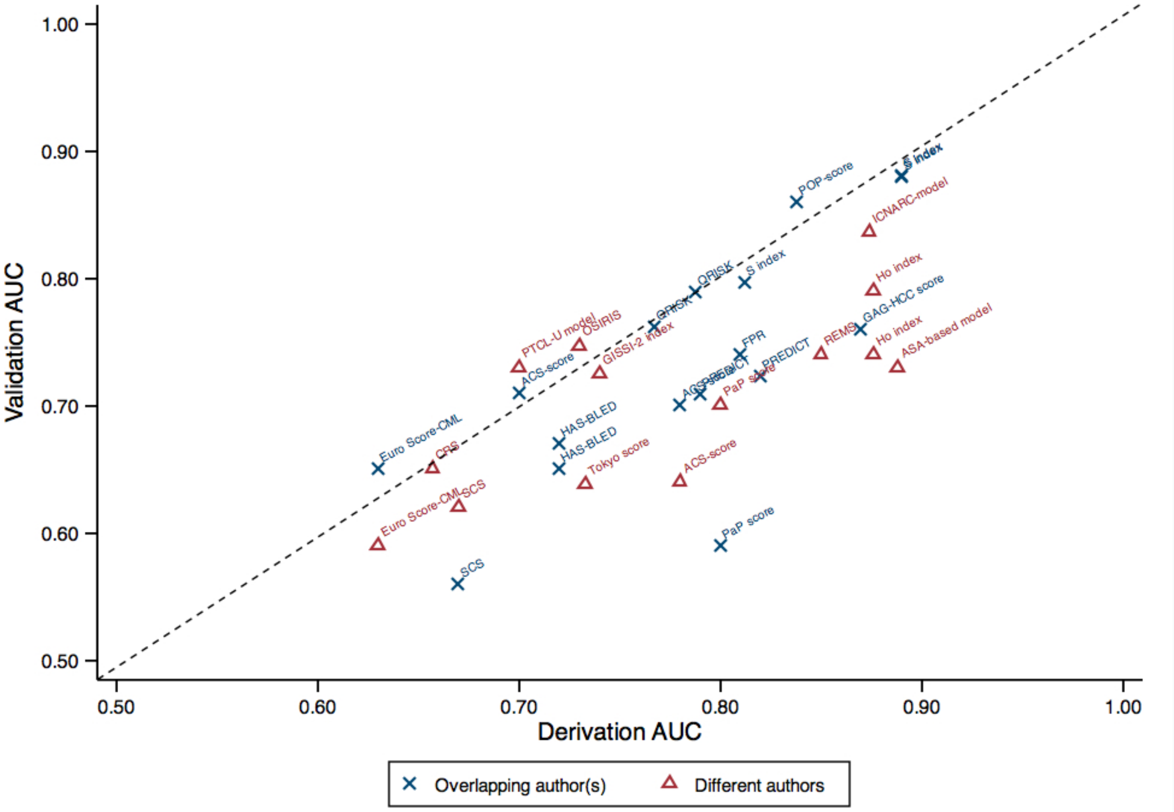

Figure 4.3: Correlation of models' predictive performance between derivation and validation studies. AUC, area under the curve; HAS-BLED, hypertension, abnormal renal/liver function, stroke, bleeding history or predisposition, labile international normalized ratio, elderly, drugs/alcohol concomitantly; FPR, florence prediction rule; GAG-HCC score, guide with age, gender, HBV DNA, core promoter mutations and cirrhosis; POP score, pancreatitis outcome prediction score; ICNARC-model, intensive care national audit & research centre model; SCS, simplified comorbidity score; ACS, acute coronary syndrome; ASA model, ASA status-based model; REMS, rapid emergency medicine score; PTCL-U, peripheral T-cell lymphoma unspecified model; GISSI-2, gruppo italiano per lo studio della sopravvivenza Nell'infarto miocardico-2; OSIRIS, osteoporosis index of risk; CRS, cardiac risk score; PaP score, palliative prognostic score; CML, chronic myelogenous leukemia.

Independent External Validation in the Derivation Study

For 58 of 127 models, some validation was included even in the same derivation study. Of those 58, 24 models had been validated in the same data set [cross-validation (n = 6), split sample (n = 13), or bootstrap method (n = 5)]. Another 34 models used a different independent sample that came from a different data set (recruited at a different time and/or location) than the training sample. AUC estimates were given for the training sample and any independent validation in the same derivation study for 18 of these 34 models.

Sensitivity Analysis

We performed a sensitivity analysis considering as external validation by overlapping author(s) any independent external validation presented that had been published in either the very same derivation study or another subsequent publication. The probability of a new model to be validated externally by either the same/overlapping or different authors at 2, 5, and 10 years was 13%, 39%, and 57%, respectively. The probability of a new model to be validated externally by same/overlapping author(s) at 2, 5, and 10 years was 9%, 25%, and 40%, respectively. In the sensitivity analysis, the AUC estimate was higher in derivation than the external validation by same/overlapping author(s) in 23 of 31 cases (median AUC change: 0.02 lower in external validation, P = 0.001 by paired t-test); it was higher in the external validation by same/overlapping author(s) than by different authors in four of seven cases (median AUC change: lower by 0.05 with different authors, P = 0.59 by paired *t*-test).

4.4 DISCUSSION

Our empirical evaluation shows that the performance of risk prediction models during external validation by overlapping or different authors is typically substantially worse than what is described when a model is first developed. Moreover, most risk prediction models never undergo some external validation in any subsequent study, and very few are externally validated by different authors than those who developed these models. Thus, their predictive ability in the literature may be overestimated.

Newly introduced risk prediction models should ideally follow a set of careful development steps from derivation to establishment and ultimately use in clinical practice.[57,109] Prospective validation of a prediction model in an independent sample other than the one where it was developed is crucial to examine the model's stability, reproducibility, and external validity.[102,109,273] Analyses in the derivation sample, even when complemented with appropriate internal validation techniques, are not sufficient.[100,111] Recently published systematic reviews of studies in prognostic research field have raised concerns about the proper validation and reporting strategies of new risk models.[212,274,275] Besides discrimination, other metrics, such as calibration or reclassification, may also need to be reported in both derivation and validation studies.[54,56,57] Poor reporting may result in difficult or even misleading interpretations.[274,276-278]

Currently, numerous prediction models have been developed for a variety of clinical conditions, settings, and outcomes[212,279-283], but few such models have been properly derived, validated, reported in medical literature, and finally implemented in clinical practice.[284,285] Moreover, even among apparently well-validated and established predictive tools, appropriate for wide clinical use, there is significant within-tool variability in predictive accuracy across different studies and clinical settings.[157] A promising initial predictive performance in the derivation sample may not be reproduced when applied in a completely independent study sample.[286,287]

Our results should also be interpreted with caution because we have studied a sample of articles that cover a 40-year span. In early years, it is understandable that methodological issues about multivariate prediction models were poorly understood and many methods were not fully developed. However, even when we limited to studies published in the last decade, the conclusions did not change and external independent validation still remained infrequent. With further methodological developments in the last few years[56,288] and sensitization to these issues, it is conceivable that further improvements may ensue in the validation practices of currently published studies, but this will require several years of follow-up to examine.

Limitations

We should acknowledge that we did not capture every single multivariate prediction model that has been published in the literature. There is no efficient search to do this, and even if there were such a search, the volume of information would have been prohibitive because it is likely that there are many hundreds and possibly several thousands of multivariate predictive models that have been proposed in the literature. There are standard searches for identifying the space of prognostic and diagnostic prediction studies in PubMed that have high sensitivity, that is, the Ingui and broad Haynes searches claim to have 98% and 96% sensitivity, respectively[289], and these can be very useful to apply in situations where searches are performed to identify studies on one or a few models, as for a topic-focused systematic review. However, the specificity of 86% and 79%, respectively, makes these searches prohibitive to search for models regardless of topic and field because one would then have to screen in detail 14% or 21%, respectively, of the entire PubMed. Our searches aimed to specifically identify models where the authors specifically describe them as new or novel. Removing the terms new and novel from the two searches yields 17,877 and 13,386 items, respectively, a number that is 30 times larger than the one that we had to screen and that is prohibitively large to screen in detail. One may speculate whether models whose authors emphasize the novelty are a selected sample with different performance than other models. However, it is unlikely that authors would emphasize novelty to camouflage shortcomings in the development and validation of the model. Moreover, it is unlikely that independent authors would have a lower propensity to try to validate a model because its original authors described it as new or novel.

Second, AUCs were sometimes not reported, and we estimated them as described previously. Inferred AUCs depend on how risk group categories are defined and (for prospective outcomes) on length of follow-up. Third, the retrieved studies had very limited reported data on other features of model performance, that is, calibration. Thus, we cannot examine whether these also deteriorate as external validation efforts are performed. Fourth, derivation and validation studies may have differed in the exact characteristics of their study populations. Although we carefully matched conditions and outcomes, such population differences are unavoidable and may explain in part the deterioration in AUC performance during external validation. It is unclear whether clinicians who might wish to apply a model may consider applying it to patients with characteristics similar to those of the participants of the derivation study or the validation study or may even extend its use to patients with very different characteristics.

Conclusions

Allowing for these caveats, our study establishes that predictive risk models may have worse performance when externally validated in other patients with the same condition and for the same outcome as they were first developed. The clinical relevance of these changes in performance needs to be considered and discussed on a case-by-case basis. For models that have clear applicability, a small or modest drop in performance may not invalidate their clinical use, whereas for others where their role is more unclear, this may be sufficient to make them clinically useless. For example, the Framingham Risk Score does not have particular high AUC but is largely used and considered useful. Moreover, the availability of other models to predict the same outcome may also be influential, if the loss of performance on external validation is such that a model is no longer competitive against other models that serve the same purpose. This is likely to represent more closely also their performance when a clinician wants to apply them to yet another external population, his/her own patients. Models may offer misleadingly high expectations of risk prediction in the absence of rigorous external independent validation.

4.5 SUPPLEMENTARY MATERIAL

4.5.1 Supplementary references

S1. Hippisley-Cox J, Coupland C, Robson J, Brindle P. Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: cohort study using QResearch database. BMJ. 2010;341:c6624. [PMID: 21148212]

S2. Stamatopoulos B, Meuleman N, De Bruyn C, Pieters K, Anthoine G, Mineur P, Bron D, Lagneaux L. A molecular score by quantitative PCR as a new prognostic tool at diagnosis for chronic lymphocytic leukemia patients. PLoS One. 2010;5(9). pii: e12780. [PMID: 20862275]

S3. Campbell HE, Gray AM, Harris AL, Briggs AH, Taylor MA. Estimation and external validation of a new prognostic model for predicting recurrence-free survival for early breast cancer patients in the UK. Br J Cancer. 2010;103(6):776-86. [PMID: 20823886]

S4. Maluenda G, Delhaye C, Gaglia MA Jr, Ben-Dor I, Gonzalez MA, Hanna NN, Collins SD, Wakabayashi K, Torguson R, Xue Z, Satler LF, Pichard AD, Waksman R. A novel percutaneous coronary intervention risk score to predict one-year mortality. Am J Cardiol. 2010;106(5):641-5. [PMID: 20723638]

S5. Chung JW, Cheon JH, Park JJ, Jung ES, Choi EH, Kim H. Development and validation of a novel prognostic scoring model for ischemic colitis. Dis Colon Rectum. 2010;53(9):1287-94. [PMID: 20706072]

S6. Drenthen W, Boersma E, Balci A, Moons P, Roos-Hesselink JW, Mulder BJ, Vliegen HW, van Dijk AP, Voors AA, Yap SC, van Veldhuisen DJ, Pieper PG; ZAHARA Investigators. Predictors of pregnancy complications in women with congenital heart disease. Eur Heart J. 2010;31(17):2124-32. [PMID: 20584777]

S7. Phillips AA, Shapira I, Willim RD, Sanmugarajah J, Solomon WB, Horwitz SM, Savage DG, Bhagat G, Soff G, Zain JM, Alobeid B, Seshan VE, O'Connor OA. A critical analysis of prognostic factors in North American patients with human T-cell lymphotropic virus type-1-associated adult T-cell leukemia/lymphoma: a multicenter clinicopathologic experience and new prognostic score. Cancer. 2010;116(14):3438-46. [PMID: 20564100]

S8. Hsu CY, Huang YH, Hsia CY, Su CW, Lin HC, Loong CC, Chiou YY, Chiang JH, Lee PC, Huo TI, Lee SD. A new prognostic model for hepatocellular carcinoma based on total tumor volume: the Taipei Integrated Scoring System. J Hepatol. 2010;53(1):108-17. [PMID: 20451283]

S9. Röllig C, Thiede C, Gramatzki M, Aulitzky W, Bodenstein H, Bornhäuser M, Platzbecker U, Stuhlmann R, Schuler U, Soucek S, Kramer M, Mohr B, Oelschlaegel U, Stölzel F, von Bonin M, Wermke M, Wandt H, Ehninger G, Schaich M; Study Alliance Leukemia. A novel prognostic model in elderly patients with acute myeloid leukemia: results of 909 patients entered into the prospective AML96 trial. Blood. 2010;116(6):971-8. [PMID: 20442365]

S10. Nowak AK, Francis RJ, Phillips MJ, Millward MJ, van der Schaaf AA, Boucek J, Musk AW, McCoy MJ, Segal A, Robins P, Byrne MJ. A novel prognostic model for malignant mesothelioma incorporating quantitative FDG-PET imaging with clinical parameters. Clin Cancer Res. 2010;16(8):2409-17. [PMID: 20371686]

S11. Elley CR, Robinson E, Kenealy T, Bramley D, Drury PL. Derivation and validation of a new cardiovascular risk score for people with type 2 diabetes: the new zealand diabetes cohort study. Diabetes Care. 2010;33(6):1347-52. [PMID: 20299482]

S12. Ananthakrishnan AN, McGinley EL, Binion DG, Saeian K. A novel risk score to stratify severity of Crohn's disease hospitalizations. Am J Gastroenterol. 2010;105(8):1799-807. [PMID: 20216534]

S13. Itoh K, Kinoshita T, Watanabe T, Yoshimura K, Okamoto R, Chou T, Ogura M, Hirano M, Asaoku H, Kurosawa M, Maeda Y, Omachi K, Moriuchi Y, Kasai M, Ohnishi K, Takayama N, Morishima Y, Tobinai K, Kaba H, Yamamoto S, Fukuda H, Kikuchi M, Yoshino T, Matsuno Y, Hotta T, Shimoyama M. Prognostic analysis and a new risk model for Hodgkin lymphoma in Japan. Int J Hematol. 2010;91(3):446-55. [PMID: 20198461]

S14. Ananthakrishnan AN, McGinley EL, Binion DG, Saeian K. Simple score to identify colectomy risk in ulcerative colitis hospitalizations. Inflamm Bowel Dis. 2010;16(9):1532-40. [PMID: 20091926]

S15. Andersson B, Andersson R, Brandt J, Höglund P, Algotsson L, Nilsson J. Gastrointestinal complications after cardiac surgery - improved risk stratification using a new scoring model. Interact Cardiovasc Thorac Surg. 2010;10(3):366-70. [PMID: 19995792]

S16. Hernández D, Sánchez-Fructuoso A, González-Posada JM, Arias M, Campistol JM, Rufino M, Morales JM, Moreso F, Pérez G, Torres A, Serón D; Spanish Late Allograft Dysfunction Study Group. A novel risk score for mortality in renal transplant recipients beyond the first posttransplant year. Transplantation. 2009;88(6):803-9. [PMID: 19920780]

S17. Menza TW, Hughes JP, Celum CL, Golden MR. Prediction of HIV acquisition among men who have sex with men. Sex Transm Dis. 2009;36(9):547-55. [PMID: 19707108]

S18. Cattermole GN, Mak SK, Liow CH, Ho MF, Hung KY, Keung KM, Li HM, Graham CA, Rainer TH. Derivation of a prognostic score for identifying critically ill patients in an emergency department resuscitation room. Resuscitation. 2009;80(9):1000-5. [PMID: 19608327]

S19. Iacob S, Gheorghe L, Iacob R, Gheorghe C, Hrehoreţ D, Popescu I. MELD exceptions and new predictive score of death on long waiting lists for liver transplantation. Chirurgia (Bucur). 2009;104(3):267-73. [PMID: 19601457]

S20. Kim HK, Jeong MH, Ahn Y, Kim JH, Chae SC, Kim YJ, Hur SH, Seong IW, Hong TJ, Choi DH, Cho MC, Kim CJ, Seung KB, Chung WS, Jang YS, Rha SW, Bae JH, Cho JG, Park SJ; other Korea Acute Myocardial Infarction Registry Investigators; Korea Acute Myocardial infarction Registry (KAMIR) Study Group of Korean Circulation Society. A new risk score system for the assessment of clinical outcomes in patients with non-ST-segment elevation myocardial infarction. Int J Cardiol. 2010;145(3):450-4. [PMID: 19541376]

S21. Moore L, Lavoie A, Turgeon AF, Abdous B, Le Sage N, Emond M, Liberman M, Bergeron E. The trauma risk adjustment model: a new model for evaluating trauma care. Ann Surg. 2009;249(6):1040-6. [PMID: 19474674]

S22. Suh SY, Choi YS, Shim JY, Kim YS, Yeom CH, Kim D, Park SA, Kim S, Seo JY, Kim SH, Kim D, Choi SE, Ahn HY. Construction of a new, objective prognostic score for terminally ill cancer patients: a multicenter study. Support Care Cancer. 2010;18(2):151-7. [PMID: 19381691]

S23. Bria E, Milella M, Sperduti I, Alessandrini G, Visca P, Corzani F, Giannarelli D, Cerasoli V, Cuppone F, Cecere FL, Marchetti A, Sacco R, Mucilli F, Malatesta S, Guetti L, Vitale L, Ceribelli A, Rinaldi M, Terzoli E, Cognetti F, Facciolo F. A novel clinical prognostic score incorporating the number of resected lymph-nodes to predict recurrence and survival in non-small-cell lung cancer. Lung Cancer. 2009;66(3):365-71. [PMID: 19327866]

S24. Negassa A, Monrad ES, Srinivas VS. A simple prognostic classification model for postprocedural complications after percutaneous coronary intervention for acute myocardial infarction (from the New York State percutaneous coronary intervention database). Am J Cardiol. 2009;103(7):937-42. [PMID: 19327419]

S25. Capodanno D, Capranzano P, Bucalo R, Sanfilippo A, Ruperto C, Caggegi A, Ussia G, Galassi AR, Tamburino C. A novel approach to define risk of stent thrombosis after percutaneous coronary intervention with drug-eluting stents: the DERIVATION score. Clin Res Cardiol. 2009;98(4):240-8. [PMID: 19219391]

S26. Hernandez DJ, Han M, Humphreys EB, Mangold LA, Taneja SS, Childs SJ, Bartsch G, Partin AW. Predicting the outcome of prostate biopsy: comparison of a novel logistic regression-based model, the prostate cancer risk calculator, and prostate-specific antigen level alone. BJU Int. 2009;103(5):609-14. [PMID: 19007374]

S27. Yuen MF, Tanaka Y, Fong DY, Fung J, Wong DK, Yuen JC, But DY, Chan AO, Wong BC, Mizokami M, Lai CL. Independent risk factors and predictive score for the development of hepatocellular carcinoma in chronic hepatitis B. J Hepatol. 2009;50(1):80-8. [PMID: 18977053]

S28. Nobre SR, Cabral JE, Gomes JJ, Leitão MC. In-hospital mortality in spontaneous bacterial peritonitis: a new predictive model. Eur J Gastroenterol Hepatol. 2008;20(12):1176-81. [PMID: 18941414]

S29. Kitai S, Kudo M, Minami Y, Ueshima K, Chung H, Hagiwara S, Inoue T, Ishikawa E, Takahashi S, Asakuma Y, Haji S, Osaki Y, Oka H, Seki T, Kasugai H, Sasaki Y, Matsunaga T. A new prognostic staging system for hepatocellular carcinoma: value of the biomarker combined Japan integrated staging score. Intervirology. 2008;51 Suppl 1:86-94. [PMID: 18544953]

S30. García-Almagro FJ, Gimeno JR, Villegas M, Hurtado J, Teruel F, Cerdán MC, González-Carrillo J, Pascual D, Rodríguez-Barranco M, Valdés M. Prognostic value of the Thrombolysis in Myocardial Infarction risk score in a unselected population with chest pain. Construction of a new predictive model. Am J Emerg Med. 2008;26(4):439-45. [PMID: 18410812]

S31. Tice JA, Cummings SR, Smith-Bindman R, Ichikawa L, Barlow WE, Kerlikowske K. Using clinical factors and mammographic breast density to estimate breast cancer risk: development and validation of a new predictive model. Ann Intern Med. 2008;148(5):337-47. [PMID: 18316752]

S32. Mountzios G, Bamias A, Voulgaris Z, Rodolakis A, Vlahos G, Gourgoulis G, Papaiakovou EE, Giannopoulos A. Prognostic factors in patients treated with taxane-based chemotherapy for recurrent or metastatic endometrial cancer: proposal for a new prognostic model. Gynecol Oncol. 2008;108(1):130-5. [PMID: 17988722]

S33. Tsimberidou AM, Wen S, O'Brien S, McLaughlin P, Wierda WG, Ferrajoli A, Faderl S, Manning J, Lerner S, Mai CV, Rodriguez AM, Hess M, Do KA, Freireich EJ, Kantarjian HM, Medeiros LJ, Keating MJ. Assessment of chronic lymphocytic leukemia and small lymphocytic lymphoma by absolute lymphocyte counts in 2,126 patients: 20 years of experience at the University of Texas M.D. Anderson Cancer Center. J Clin Oncol. 2007;25(29):4648-56. [PMID: 17925562]

S34. Kim SJ, Kim BS, Choi CW, Choi J, Kim I, Lee YH, Kim JS. Ki-67 expression is predictive of prognosis in patients with stage I/II extranodal NK/T-cell lymphoma, nasal type. Ann Oncol. 2007;18(8):1382-7. [PMID: 17693651]

S35. Xu X, Ling Q, Wu J, Chen J, Gao F, Feng XN, Zheng SS. A novel prognostic model based on serum levels of total bilirubin and creatinine early after liver transplantation. Liver Int. 2007;27(6):816-24. [PMID: 17617125]

S36. Aletti GD, Santillan A, Eisenhauer EL, Hu J, Aletti G, Podratz KC, Bristow RE, Chi DS, Cliby WA.A new frontier for quality of care in gynecologic oncology surgery: multi-institutional assessment of short-term outcomes for ovarian cancer using a risk-adjusted model. Gynecol Oncol. 2007;107(1):99-106. [PMID: 17602726]

S37. Cianchi F, Messerini L, Comin CE, Boddi V, Perna F, Perigli G, Cortesini C. Pathologic determinants of survival after resection of T3N0 (Stage IIA) colorectal cancer: proposal for a new prognostic model. Dis Colon Rectum. 2007;50(9):1332-41. [PMID: 17429709]

S38. Tischendorf JJ, Hecker H, Krüger M, Manns MP, Meier PN. Characterization, outcome, and prognosis in 273 patients with primary sclerosing cholangitis: A single center study. Am J Gastroenterol. 2007;102(1):107-14. [PMID: 17037993]

S39. Senni M, Santilli G, Parrella P, De Maria R, Alari G, Berzuini C, Scuri M, Filippi A, Migliori M, Minetti B, Ferrazzi P, Gavazzi A. A novel prognostic index to determine the impact of cardiac conditions and co-morbidities on one-year outcome in patients with heart failure. Am J Cardiol. 2006;98(8):1076-82. [PMID: 17027575]

S40. Ballesteros MA, López-Hoyos M, Muñoz P, Marin MJ, Miñambres E. Apoptosis of neuronal cells induced by serum of patients with acute brain injury: a new in vitro prognostic model. Intensive Care Med. 2007;33(1):58-65. [PMID: 16964482]

S41. Park YH, Kim WS, Bang SM, Lee SI, Uhm JE, Kang HJ, Na II, Yang SH, Lee SS, Kim K, Jung CW, Park K, Ko YH, Ryoo BY. Prognostic factor analysis and proposed prognostic model for conventional treatment of high-grade primary gastric lymphoma. Eur J Haematol. 2006;77(4):304-8. [PMID: 16879609]

S42. Went P, Agostinelli C, Gallamini A, Piccaluga PP, Ascani S, Sabattini E, Bacci F, Falini B, Motta T, Paulli M, Artusi T, Piccioli M, Zinzani PL, Pileri SA. Marker expression in peripheral T-cell lymphoma: a proposed clinical-pathologic prognostic score. J Clin Oncol. 2006;24(16):2472-9. [PMID: 16636342]

S43. Williams BA, Wright RS, Murphy JG, Brilakis ES, Reeder GS, Jaffe AS. A new simplified immediate prognostic risk score for patients with acute myocardial infarction. Emerg Med J. 2006;23(3):186-92. [PMID: 16498154]

S44. Tekesin I, Eberhart LH, Schaefer V, Wallwiener D, Schmidt S. Evaluation and validation of a new risk score (CLEOPATRA score) to predict the probability of premature delivery for patients with threatened preterm labor. Ultrasound Obstet Gynecol. 2005;26(7):699-706. [PMID: 16308893]

S45. Miyake Y, Sakaguchi K, Iwasaki Y, Ikeda H, Makino Y, Kobashi H, Araki Y, Ando M, Kita K, Shiratori Y. New prognostic scoring model for liver transplantation in patients with non-acetaminophen-related fulminant hepatic failure. Transplantation. 2005;80(7):930-6. [PMID: 16249741]

S46. Alvares CL, Davies FE, Horton C, Patel G, Powles R, Sirohi B, Zuha R, Gatt A, Saso R, Treleaven JG, Dearden CE, Potter MN, Ethell ME, Morgan GJ. Long-term outcomes of previously untreated myeloma patients: responses to induction chemotherapy and high-dose melphalan incorporated within a risk stratification model can help to direct the use of novel treatments. Br J Haematol. 2005;129(5):607-14. [PMID: 15916682]

S47. Meyer S, Gottschling S, Biran T, Georg T, Ehlayil K, Graf N, Gortner L. Assessing the risk of mortality in paediatric cancer patients admitted to the paediatric intensive care unit: a novel risk score? Eur J Pediatr. 2005;164(9):563-7. [PMID: 15912385]

S48. Sadanandan S, Cannon CP, Gibson CM, Murphy SA, DiBattiste PM, Braunwald E; TIMI Study Group. A risk score to estimate the likelihood of coronary artery bypass surgery during the index hospitalization among patients with unstable angina and non-ST-segment elevation myocardial infarction. J Am Coll Cardiol. 2004;44(4):799-803. [PMID: 15312862]

S49. Skírnisdóttir I, Seidal T, Sorbe B. A new prognostic model comprising p53, EGFR, and tumor grade in early stage epithelial ovarian carcinoma and avoiding the problem of inaccurate surgical staging. Int J Gynecol Cancer. 2004;14(2):259-70. [PMID: 15086725]

S50. Haferlach T, Kern W, Schoch C, Schnittger S, Sauerland MC, Heinecke A, Büchner T, Hiddemann W; German AML Cooperative Group. A new prognostic score for patients with acute myeloid leukemia based on cytogenetics and early blast clearance in trials of the German AML Cooperative Group. Haematologica. 2004;89(4):408-18. [PMID: 15075074]

S51. Halonen KI, Leppäniemi AK, Lundin JE, Puolakkainen PA, Kemppainen EA, Haapiainen RK. Predicting fatal outcome in the early phase of severe acute pancreatitis by using novel prognostic models. Pancreatology. 2003;3(4):309-15. [PMID: 12890993]

S52. von Eyben FE, Madsen EL, Fritsche H, Suciu G, Liu F, Amato R. A new prognostic model for testicular germ cell tumours. APMIS. 2003;111(1):100-5; discussion 105. [PMID: 12752246]

S53. Freedland SJ, Wieder JA, Jack GS, Dorey F, deKernion JB, Aronson WJ. Improved risk stratification for biochemical recurrence after radical prostatectomy using a novel risk group system based on prostate specific antigen density and biopsy Gleason score. J Urol. 2002;168(1):110-5. [PMID: 12050502]

S54. Castellanos-Ortega A, Delgado-Rodríguez M, Llorca J, Sánchez Burón P, Mencía Bartolomé S, Soult Rubio A, Milano Manso G, Domínguez Sampedro P, Blanco Montero R, Rodríguez Núñez A, Zambrano Pérez E, Rey Galán C, López Negueruela N, Reig Sáenz R. A new prognostic scoring system for meningococcal septic shock in children. Comparison with three other scoring systems. Intensive Care Med. 2002;28(3):341-51. [PMID: 11904666]

S55. Smith MA, Luxton RW, Pallister CJ, Smith JG. A novel predictive model of outcome in de novo AML based on S-phase activity and proliferative response of blast cells to haemopoietic growth factors. Leuk Res. 2002;26(4):345-8. [PMID: 11839376]

S56. Josting A, Franklin J, May M, Koch P, Beykirch MK, Heinz J, Rudolph C, Diehl V, Engert A. New prognostic score based on treatment outcome of patients with relapsed Hodgkin's lymphoma registered in the database of the German Hodgkin's lymphoma study group. J Clin Oncol. 2002;20(1):221-30. [PMID: 11773173]

S57. Djoulah S, Busson M, Sasazuki T, Maillere B, Yasunaga S, Kimura A, Charron D, Hors J. A new predictive model for insulin-dependent diabetes mellitus susceptibility based on combinations of molecular HLA-DRB1 and HLA-DQB1 pockets. Tissue Antigens. 1999;54(4):341-8. [PMID: 10551417]

S58. Wong DT, Cheng DC, Kustra R, Tibshirani R, Karski J, Carroll-Munro J, Sandler A. Risk factors of delayed extubation, prolonged length of stay in the intensive care unit, and mortality in patients undergoing coronary artery bypass graft with fast-track cardiac anesthesia: a new cardiac risk score. Anesthesiology. 1999;91(4):936-44. [PMID: 10519495]

S59. Duong DH, Kolluri VR, Spittaler PJ, Sengupta RP. Risk Score Estimation: a new method to determine optimal timing of aneurysm clipping for improved management outcome. Neurol Res. 1998;20(3):218-24. [PMID: 9583582]

S60. Adler M, Verset D, Bouhdid H, Bourgeois N, Gulbis B, Le Moine O, Van de Stadt J, Gelin M, Thiry P. Prognostic evaluation of patients with parenchymal cirrhosis. Proposal of a new simple score. J Hepatol. 1997 Mar;26(3):642-9. [PMID: 9075673]

S61. Paganini EP, Halstenberg WK, Goormastic M. Risk modeling in acute renal failure requiring dialysis: the introduction of a new model. Clin Nephrol. 1996;46(3):206-11. [PMID: 8879857]

S62. Sarbia M, Bittinger F, Porschen R, Dutkowski P, Willers R, Gabbert HE. Prognostic value of histopathologic parameters of esophageal squamous cell carcinoma. Cancer. 1995;76(6):922-7. [PMID: 8625216]

S63. Ménard S, Bufalino R, Rilke F, Cascinelli N, Veronesi U, Colnaghi MI. Prognosis based on primary breast carcinoma instead of pathological nodal status. Br J Cancer. 1994;70(4):709-12. [PMID: 7917924]

S64. Moran MR, Rothenberger DA, Gallo RA, Goldberg SM, James EC. A predictive model for distant metastases in rectal cancer using DNA ploidy studies. Am J Surg. 1992;163(6):599-601. [PMID: 1595840]

S65. Altaca G, Sayek I, Onat D, Cakmakçi M, Kamiloğlu S. Risk factors in perforated peptic ulcer disease: comparison of a new score system with the Mannheim Peritonitis Index. Eur J Surg. 1992;158(4):217-21. [PMID: 1352135]

S66. Taussig LM, Kattwinkel J, Friedewald WT, Di Sant'Agnese PA. A new prognostic score and clinical evaluation system for cystic fibrosis. J Pediatr. 1973;82(3):380-90. [PMID: 4698929]

4.5.2 Supplementary table: Details of the derivation studies of newly introduced risk prediction models without any further validation studies.

First author (year)	Model	Outcome(s)	Study population	Disease / Clinical condition	Sample size	AUC (95% CI) Derivation / [Validation*]	Calibration metric
Hippisley-Cox J, et al. (2010)[S1]	QRISK	CVD	general population free of cardiovascular disease and not taking statins	CVD	2343759 (development); 1267159 (validation)	ND / [Validation: 0.828 (95% CI, 0.826-0.830)(men); 0.842 (95% CI, 0.840-0.844)(women)]	Predicted / Observed ratio
Stamatopoulos B, et al. (2010)[S2]	quantitative PCR score for CLL	treatment-free survival, overall survival	patients with chronic lymphocytic leukemia	malignancies	170	ND	ND
Campbell HE,et al. (2010)[S3]	number of positive axillary lymph nodes, tumour grade and size, age	first recurrent event	women with early breast cancer	malignancies	1844 (development); 1787 (validation)	0.745 (95% CI, 0.717-0.773) / [Validation: 0.720 (95% CI, 0.693-0.746)]	Predicted / Observed ratio
Maluenda G, et al. (2010)[S4]	Novel PCI risk score	1-year mortality	consecutive patients undergo PCI	CVD	6932 (development); 973 (validation)	0.818 (ND) / [Validation: 0.836 (ND)]	Hosmer-Lemeshow goodness-of-fit test
Chung JW, et al. (2010)[S5]	model to predict development of severe ischemic colitis	severe ischemic colitis	patients with ischemic colitis	GI-related	153	ND / [Validation: 0.91 (95% CI, 0.86-0.96)]	Hosmer-Lemeshow goodness-of-fit test

Drenthen W, et al. (2010)[S6]	modified risk score for cardiac complications during completed pregnancies	composite cardiac complications	women with congenital heart disease	CVD	714	0.762	Hosmer-Lemeshow goodness-of-fit test
Phillips AA, et al. (2010)[S7]	new prognostic model for patients with HTLV-1-associated ATLL	overall response rate, overall survival	patients with HTLV-1 associated ATLL	malignancies	89	ND	ND
Hsu CY, et al. (2010)[S8]	Taipei Integrated Scoring System	survival	patients with hepatocellular carcinoma	malignancies	2030	ND	ND
Röllig C, et al. (2010)[S9]	novel prognostic model in elderly patients with acute myeloid leukemia	disease-free survival	elderly patients with AML	malignancies	909	ND	ND
Nowak AK, et al. (2010)[S10]	prognostic nomogram for malignant pleural mesothelioma	survival	patients with malignant pleural mesothelioma	malignancies	89	0.652 (ND)	ND
Elley CR, et al. (2010)[S11]	Diabetes Cohort Study risk prediction model	first fatal or nonfatal cardiovascular event	patients with type 2 diabetes without previous CVD	other	36127 (development); 12626 (validation)	0.673 (ND) / [Validation: 0.68 (ND)]	Predicted / Observed ratio

Ananthakrishnan AN, et al. (2010)[S12]	simple quantitative risk score to measure the severity of Crohn’s disease hospitalizations	severe hospitalization course	patients with Crohn's disease	GI-related	25938 (developement); ND (validation)	0.68 (ND) / [Validation: 0.68 (ND)]	ND
Itoh K, et al. (2010)[S13]	simple prognostic model for Hodgkin Lymphoma	overall survival	patients with advanced Hodgkin lymphoma	malignancies	167	ND	ND
Ananthakrishnan AN, et al. (2010)[S14]	colectomy risk model	total colectomy	hospitalized patients with ulcerative colitis	GI-related	15142 (development); ND (validation)	ND	ND
Andersson B, et al. (2010)[S15]	GICS	gastrointestinal complications after cardiac surgery	patients undergoing cardiac surgery procedures	CVD	5593 (development); 1123 (validation)	0.81 (ND) / [Validation: 0.83 (ND)]	Hosmer-Lemeshow goodness-of-fit test
Hernández D, et al. (2009)[S16]	model for 3 year mortality in post kidney transplant patients	mortality	kidney transplantation recipients	other	2542 (development); 2476 (validation)	0.75 (95% CI, 0.72-0.78) / [Validation: 0.74 (95% CI, 0.70-0.77)]	ND

Menza TW, et al. (2009)[S17]	full and simple HIV acquisition model - 1 and 4 years	HIV acquisition	men who have sex with men who tested for HIV	other	1903 (development); 2081 (validation)	ND / [Validation: Full-1 0.68 (95% CI, 0.62-0.75); Full-4 0.67 (95% CI, 0.62-0.71); Simple-1 0.67 (95% CI, 0.60-0.75); Simple-4 0.66 (0.61-0.71)	Predicted / Observed ratio
Cattermole GN, et al. (2009) [S18]	PEDS	death or admission to the ICU within 7 days of ED attendance; 30-day mortality	critically ill patients in ED	other	330	0.909 (95% CI, 0.872-0.938); 0.898 (95% CI, 0.860-0.928)	ND
Iacob S, et al. (2009)[S19]	model to predict death on transplant list prior to receiving liver transplant	death	adults patients with end-stage liver disease	GI-related	372	0.85 (ND)	ND
Kim HK, et al. (2010)[S20]	simple assessment tool for better early bedside risk stratification for both short- and long-term clinical outcomes	death from any cause	patients with NSTEMI	CVD	2148	0.815 (95% CI, 0.79-0.84)	ND

Moore L, et al. (2009)[S21]	Trauma Risk Adjustment Model (TRAM)	hospital mortality	adult patients with blunt trauma	other	72527 (development); 178377 (validation)	0.944 (ND) / [Validation: 0.942 (ND)]	Hosmer-Lemeshow goodness-of-fit test
Suh SY, et al. (2010)[S22]	Objective Prognostic Score (OPS)	survival	terminally ill cancer patients	malignancies	209	ND	ND
Bria E, et al. (2009)[S23]	clinical prognostic score in non-small-cell lung cancer	overall survival, cancer-specific survival, disease-free survival	resected non-small-cell lung cancer patients	malignancies	415 (development); 297 (validation)	ND / [Validation: ND]	ND
Negassa A, et al. (2009)[S24]	tree-structured prognostic classification for postprocedural complications after PCI for aMI	postprocedural complications (in-hospital death, stroke, or CABG surgery)	patients who underwent emergency PCI	CVD	5385 (development); 7414 (validation)	ND / [Validation: 0.78 (95% CI, 0.75-0.80)]	ND
Capodanno D, et al. (2009)[S25]	DERIVATION score	stent thrombosis	patients who underwent PCI with DES as treatment of symptomatic CAD	CVD	1377	ND	ND
Hernandez DJ, et al. (2009)[S26]	logistic-regression based model	prostate cancer, high-grade prostate cancer (Gleason score≥7)	men scheduled for prostate biopsy	malignancies	1108	71.2 (ND)	ND

Yuen MF, et al. (2009)[S27]	GAG-HCC score (5- and 10-years)	hepatocellular carcinoma	patients with chronic HBV infection	GI-related	820	Full model-5 years: 0.88 (95% CI, 0.82-0.93); Full model-10 years: 0.89 (95% CI, 0.85-0.93); Simple model-10 years: 0.88 (95% CI, 0.82-0.92)	ND
Nobre SR, et al. (2008)[S28]	predictive model for in-hospital mortality for patient with end-stage liver disease	in-hospital mortality	cirrhotic patients with spontaneous bacterial peritonitis	other	73	0.88 (95% CI, 0.78-0.97)	ND
Kitai S, et al. (2008)[S29]	The biomarker combined Japan Integrated Staging (bm-JIS) score	overall survival	patients with hepatocellular carcinoma	malignancies	1924	ND	ND
García-Almagro FJ, et al. (2008)[S30]	improved TIMI risk score with diabetes and EF	cardiac events (MI, revascularization, cardiac death)	patients with chest pain without ST-segment elevation	CVD	711	ND	ND
Tice JA, et al. (2008)[S31]	Breast Cancer Surveillance Consortium breast density model algorithm	breast cancer (invasive cancer and ductal carcinoma in situ)	women undergoing mammography who had no previous diagnosis of breast cancer	malignancies	377 440 (derivation); 251 789 (validation)	0.657 (95% CI, 0.65-0.67) / [Validation: 0.66 (95% CI, 0.651-0.669)]	Predicted / Observed ratio

Mountzios G, et al. (2008)[S32]	prognostic factors in RMEC after taxane-based chemotherapy	mortality	patients who had received paclitaxel-based chemotherapy for recurrent or metastatic endometrial adenocarcinoma	malignancies	110	ND	ND
Tsimberidou AM, et al. (2007)[S33]	new prognostic score (3 different modesl)	survival	patients with CLL and SLL	malignancies	1417 (derivation); 710 (validation)	ND	ND
Kim SJ, et al. (2007)[S34]	new prognostic score	overall and disease free survival	patients with localized extranodal natural killer (NK)/T-cell lymphoma	malignancies	50	ND	ND

Xu X, et al. (2007)[S35]	post-transplant model	3, 6 months and 1 year post-transplantation mortality	patients who underwent liver transplantation for benign end-stage liver diseases	other	161	0.876 (ND)(3 mo);
						0.878 (ND)(6 mo);
						0.849 (ND)(1 y)
							Hosmer-Lemeshow goodness-of-fit test

Aletti GD, et al. (2007)[S36]	risk-adjusted, multicenter outcomes model for ovarian cancer surgery	30 days morbidity, 3 months mortality, chemotherapy non-feasible	gynecologic cancer surgery	malignancies	564	ND	ND

Cianchi F, et al. (2007)[S37]	model to identify patients with high-risk Stage IIA colorectal cancer	mortality	patients with Stage II colorectal cancer	malignancies	238	ND	ND
Tischendorf JJ, et al. (2007)[S38]	Primary sclerosing cholangitis (PSC) score	survival	patients with primary sclerosing cholangitis	GI-related	182 (derivation); 91 (validation)	ND	ND
Senni M, et al. (2006)[S39]	CardioVascular Medicine Heart Failure (CVM-HF) index	mortality	patients with stable heart failure	CVD	292 (derivation); 515 (validation)	0.844 (95% CI, 0.779-0.89) / [Validation: 0.812 (95% CI, 0.76-0.86)]	Hosmer-Lemeshow goodness-of-fit test
Ballesteros MA, et al. (2007)[S40]	model for early apoptosis rate	6-months mortality	patients who had suffered from acute brain injury requiring intensive care	other	70	ND	Hosmer-Lemeshow goodness-of-fit test
Park YH, et al. (2006)[S41]	high-grade primary gastric lymphoma (HG-PGL) risk model	overall survival	patients with high-grade primary gastric lymphoma	malignancies	214	ND	ND
Went P, et al. (2006)[S42]	modified Prognostic Index for T-cell lymphoma	Disease-specific survival	patients with peripheral T-cell lymphomas/unspecified (PTCLs/U)	malignancies	93	ND	ND

Williams BA, et al. (2006)[S43]	a new simplified immediate prognostic risk score	1-month mortality	patients with aMI	CVD	809 (derivation); 403 (validation)	0.78 (ND) / [Validation: 0.81 (ND)]	ND
Tekesin I, et al. (2005)[S44]	CLEOPATRA I, CLEOPATRA II	preterm delivery	pregnant women at risk of preterm delivery	other	85 (derivation); 85 (validation	ND / [Validation: 0.692 (95% CI, 0.562-0.821) (Cleopatra I); 0.811 (95% CI, 0.694-0.928) (Cleopatra II)]	ND
Miyake Y, et al. (2005)[S45]	prognostic scoring model for liver transplantation	2-weeks fatal outcome	patients with fulminant hepatic failure	other	80 (derivation); 26 (validation)	ND	ND
Alvares CL, et al. (2005)[S46]	multiparametric risk-adapted model	overall survival	patients with previously untreated multiple myeloma	malignancies	383	ND	ND
Meyer S, et al. (2005)[S47]	mortality risk score for paediatric cancer patients admitted to ICU	non-survival	paediatric cancer patients	malignancies	32	ND	ND
Sadanandan S, et al. (2004)[S48]	Coronary Artery Bypass Graft (CABG) risk score	in-hospital CABG surgery	patients with UA and NSTEMI	CVD	2220 (derivation); 3722 (validation)	ND / [Validation: 0.61 (ND)]	ND
Skírnisdóttir I, et al. (2004)[S49]	prognostic model for early stage epithelial ovarian carcinoma	disease-free survival	patients with ovarian carcinomas in FIGO stages IA-IIC	malignancies	226	ND	ND

Haferlach T, et al. (2004)[S50]	cytogenetically based risk score	complete remission, overall survival, event-free survival, relapse-free survival	patients with de novo AML	malignancies	321	ND	ND
Halonen KI, et al. (2003)[S51]	Severe Acute Pancreatitis (SAP) score	mortality	patients in the early phase of severe acute pancreatitis	GI-related	234 (development); 60 (validation)	ND / [Validation: 0.862 (0.761-0.962)]	ND
von Eyben FE, et al. (2003)[S52]	risk model for survival prediction in patients with testicular germ cell tumor	survival	patients with metastatic testicular germ cell tumours	malignancies	81	0.77 (ND)	ND
Freedland SJ, et al. (2002)[S53]	risk stratification model after radical prostatectomy	adverse pathological features or biochemical recurrence	patients who underwent radical prostatectomy	malignancies	325 (development); 490 (validation)	ND	ND
Castellanos-Ortega A, et al. (2002)[S54]	new prognostic scoring system for meningococcal septic shock in children	mortality	children admitted to the pediatric ICU with presumed or confirmed meningococcal septic shock	other	192 (development); 158 (validation)	[Development: 0.91 (ND)] / [Validation: 0.88 (ND)]	Hosmer-Lemeshow goodness-of-fit test
Smith MA, et al. (2002)[S55]	novel predictive model of outcome in de novo AML	complete remission	AML patients with de novo disease	other	30	ND	ND

Josting A, et al. (2002)[S56]	German Hodgkin's Lymphoma Study Group (GHSG) database risk score	freedom from second failure	patients with relapsed Hodgkin's disease	malignancies	422	0.59 (ND)[†]	ND
Djoulah S, et al. (1999)[S57]	model based on the characteristics of the relevant pockets of HLA-DR and -DQ molecules	risk susceptibility prediction of insulin-dependent type I diabetes	healthy controls	CVD	337	ND	Predicted / Observed ratio
Wong DT, et al. (1999)[S58]	Cardiac Risk Scores (CRS)	prolonged ICU length of stay; delayed extubation	patients undergoing CABG surgery with fast-track cardiac anesthesia	CVD	885 (development); 885 (validation)	[Development: prolonged ICU 0.889 (ND); delayed extubation 0.728 (ND)] / [Validation: prolonged ICU 0.851 (ND); delayed extubation 0.707 (ND)]	Predicted / Observed ratio
Duong DH, et al. (1998)[S59]	risk score for unfavorable outcome of surgery for cebrebral aneurysm	patient outcome at discharge	patients undegoingr aneurysm surgery	CVD	703	ND	ND
Adler M, et al. (1997)[S60]	Erasme score	1-year liver related mortality	patients with parenchymal cirrhosis	GI-related	63 (development); 46 (validation)	0.936 (ND) / [Validation: ND]	ND

Paganini EP, et al. (1996)[S61]	mortality prediction model for patients requiring hemodialysis in ICU	mortality	patients with acute renal failure	critical illness	512 (development); 88 (validation1); 35 (valdiation2)	ND	Hosmer-Lemeshow goodness-of-fit test
Sarbia M, et al. (1995)[S62]	New prognostic score of histopathologic al parameters	survival	patients who underwent potentially curative resection for squamous cell carcinoma of the esophagus	malignancies	138	ND	ND
Ménard S, et al. (1994)[S63]	New prognostic score based on primary breast carcinoma	overall survival	surgically treated breast cancer patients	malignancies	467	ND	ND
Moran MR, et al. (1992)[S64]	DNA ploidy based model	development of distant metastases	patients who underwent surgery for rectal cancer	malignancies	188	ND	ND
Altaca G, et al. (1992)[S65]	score to predict survival after perforated duodenal ulcer	mortality	patients with peptic ulcers	GI-related	173	ND	ND
Taussig LM, et al. (1973)[S66]	clinical score to predict mortality in cystic fibrosis	mortality	patiets with cystic fibrosis	other	73	ND	ND

* Any validation of the newly proposed model in the same derivation study publication through some unbiased procedure (cross-validation, split-sample, bootstrap) on the training set, or in a totally different testing set.
† AUC was not given by the authors and was estimated as described in the Methods.

AUC area under the curve; CI, confidence interval; CVD, caridovascular disease; ND, no data; GICS, gastrointestinal complication score; HIV, human immunodeficiency virus; PEDS, Prince of Wales Emergency Department Score; ICU, intensive care unit; ED, emergency department; PCR, polymerase chain reaction; CLL, chronic lymphocytic leukemia; SLL, small lymphocytic lymphoma; PCI, percutaneous coronary intervention; GI, gastrointestinal; HTLV-1, human T-cell lymphotropic virus type-1; ATLL, adult T-cell leukemia/lymphoma; AML, acute myeloid leukemia; CV, cardiovascular; MI, myocardial infarction; CAD, coronary artery disease; CABG, coronary artery bypass graft; UA, unstable angina; NSTEMI, non-ST-elevation myocardial infarction; DES, drug-eluting stent; HLA, human leukocyte antigen; DNA, deoxyribonucleic acid.

4.5.3 Supplementary table: Details of the derivation studies of newly introduced risk prediction models that were further validated.

First author (year)	Model	Outcome(s)	Study population	Disease/ Clinical area	Sample size	AUC (95% CI) Derivation / [Validation*]	Calibration metric
Zhou K, et al. (2010)[218]	S index	significant liver fibrosis; advanced liver fibrosis; cirrhosis	patients with chronic HBV	GI-related	386 (development); 416 (validation)	0.686 (ND); 0.698 (ND); 0.762 (ND) / [Validation: 0.812 (ND); 0.89 (ND); 0.89(ND)]	ND
Röllig C, et al. (2010)[219]	novel prognostic model in elderly patients with acute myeloid leukemia	overall survival	elderly patients with acute mteloid leukemia	malignancies	909	ND	ND
Pisters R, et al. (2010)[12]	HAS-BLED	major bleeding	patients with atrial fibrillation	CVD	3963	0.72 (95% CI, 0.65-0.79)	ND
Conti A, et al. (2010)[220]	Florence prediction rule	6-month composite endpoint (CV death, nonfatal MI, revascularization)	patients with acute chest without known CAD	CVD	1106 (development); 1127 (validation)	0.83 (95% CI, 0.77-0.88) / [Validation: 0.81 (ND)]	ND
Wishart GC, et al. (2010)[221]	PREDICT	all-cause mortality, breast-cancer specific mortality	patients with invasive breast cancer	malignancies	5694 (development); 5468 (validation)	0.81 (ND); 0.84 (ND) / [Validation: 0.79 (ND); 0.82 (ND)]	Hosmer-Lemeshow goodness-of-fit test

Yuen MF, et al. (2009)[290]	GAG-HCC score (5-year) ("Guide with Age, ender, HBV DNA, Core promoter mutations and Cirrhosis")	Hepatocellular carcinoma	patients with chronic HBV infection	GI-related	820	0.87 (95% CI, 0.82-0.93)	ND
Yau T, et al. (2008)[223]	Advanced Liver Cancer Prognostic System (ALCPS)	3-month survival	patients with advanced hepatocellular carcinoma	malignancies	1109 (development); 361 (validation) / [ND]	ND / [Validation: 0.77 (ND)]	ND
Hippisley-Cox J, et al. (2007)[31]	QRISK	CVD (MI, CHD, stroke, TIAs)	UK primary care population (QRESEARCH cohort)	other	1283174 (development)(men: 636753, women: 646421); 614553 (validation) (men: 305140; women: 309413)	ND / [Validation: 0.7674 (ND) (men); 0.7879 (ND) (women)]	Predicted / Observed ratio
Harrison DA, et al. (2007)[224]	Pancreatitis Outcome Prediction (POP) Score	mortality	patients admitted to ICUs with severe acute pancreatitis	GI-related	2462 (1494 (development); 968 (validation))	ND / [Validation: 0.8381 (ND)]	Hosmer-Lemeshow goodness-of-fit test
Harrison DA, et al. (2007)[291]	Intensive Care National Audit & Research Centre (ICNARC) model	mortality	critical care admissions	other	137100 (development); 79526 (validation)	0.863 (ND) / [Validation: 0.874 (ND)]	Hosmer-Lemeshow goodness-of-fit test

Colinet B, et al. (2005)[226]	Simplified Comorbidity Score (SCS)	survival	non-small-cell lung cancer patients	cancer	735 (development); 136 (validation)	0.67 (ND)[†] / [Validation: ND]	ND
Sanchis J, et al. (2005)[227]	New Risk Score for Patients With Acute Chest Pain, Non–ST-Segment Deviation, and Normal Troponin Concentrations	mortality or MI (primary end point); mortality, MI, or urgent revascularization (secondary end point)	patients with chest pain, non–ST-segment deviation ECG, and normal troponin levels	CVD	646	0.78 (ND) (primary end point); 0.70 (ND) (secondary end point)	ND
Tateishi R, et al. (2005)[228]	Tokyo score	death	patients with naïve hepatocellular carcinoma received medical ablation (development) and underwent hepatectomy (validation)	cancer	403 (development); 203 (validation)	0.737 (ND) / [Validation: 0.733 (ND)]	ND

Donati A, et al. (2004)[229]	ASA status-based model	death or survival at hospital discharge	patients who underwent any type of elective or emergency surgical procedure (excluding cardiac surgery or Caesarean delivery)	other	1936 (development); 1849 (validation)	0.881 (95% CI, 0.833-0.930) / [Validation: 0.888 (95% CI, 0.838-0.937)]	Hosmer-Lemeshow goodness-of-fit test
Ho GT, et al. (2004)[230]	Ho index	response (no colectomy) or non-response to medical therapy (colectomy)	patients with severe ulcerative colitis	GI-related	167	0.876 (ND)	ND
Olsson T, et al. (2004)[231]	Rapid Emergency Medicine score (REMS)	in-hospital mortality	nonsurgical adult entries to the emergency department	other	11751	0.852 (95% CI, 0.838-0.866) / [Validation: 0.85 (ND)]	Hosmer-Lemeshow goodness-of-fit test
Gallamini A, et al. (2004)[232]	a new prognostic model for Peripheral T-cell lymphoma unspecified (PTCL-U)	overall survival	patients with peripheral T-cell non-Hodgkin lymphomas	cancer	385	0.70 (ND)[†]	ND
Villella M, et al. (2003)[233]	GISSI-2 index	mortality	STEMI patients	CVD	6251	0.74 (ND)	ND
Sedrine WB, et al. (2002)[234]	Osteoporosis Index of Risk (OSIRIS)	risk of osteoporosis	postmenopausal women	other	1303	0.73 (ND)[†]	ND

Josting A, et al. (2002)[235]	German Relapsed Hodgkin Prognostic Score (GRHS)	overall survival (primary end point), freedom from second failure (secondary end point)	patients with relapsed Hodgkin's disease	malignancies	422	0.69 (ND)†	ND
LeMaire SA, et al. (2001)[236]	preoperative risk factors to predict an adverse outcome after elective thoracoabdominal aortic aneurysm repair	composite end-point: death within 30 days, death before discharge from the hospital, paraplegia, paraparesis, stroke, or acute renal failure requiring dialysis	patients underwent elective graft repair of thoracoabdominal aortic aneurysm	CVD	1108	ND	ND
Wong DT, et al. (1999)[237]	Cardiac Risk Scores (preoperative and postoperative risk factors)	mortality	patients undergoing CABG surgery with fast-track cardiac anesthesia	CVD	885 (development); 885 (validation)	0.725 (ND) (mortality) / [Validation: 0.657 (ND) (mortality)]	Predicted / Observed ratio
Pirovano M, et al. (1999)[292]	Palliative Prognostic Score (PaP Score)	all-cause mortality	terminally ill cancer patients with advanced solid tumors	malignancies	519	0.80 (ND)†	ND
Hasford J, et al. (1998)[239]	The Euro Score for CML	survival	patients with early CML	malignancies	981 (development); 322 (validation)	0.63 (ND)† / [Validation: ND]	ND

Roberts AB, et al. (1985)[240]	Prediction model using ultrasound parameters to predict fetal birth weight	fetal weight	fetuses that delivered within 48h of the ultrasound measurements	other	50	0.64 (ND)[†]	ND
Lenstrup C, et al. (1982)[241]	CTG scoring system	intra-uterine growth	pregnant women in the 35-36th week of pregnancy	other	88	ND	ND

[*] Any validation of the newly proposed model in the same derivation study publication through some unbiased procedure (cross-validation, split-sample, bootstrap) on the training set, or in a totally different testing set.
[†] AUC was not given by the authors and was estimated as described in the Methods.

AUC, area under the curve; CI, confidence interval; CVD, cardiovascular disease; NA, not applicable; ND, no data; HAS-BLED, Hypertension, Abnormal renal/liver function, Stroke, Bleeding history or predisposition, Labile international normalized ratio, Elderly (>65 years), Drugs/alcohol concomitantly; GI, gastrointestinal; ECG, electrocardiogram; STEMI, ST-elevation myocardial infarction; CABG, coronary artery bypass graft; CML, chronic myeloid leukemia; CTG, cardiotocography; MACCE, major adverse cardiovascular events; MI, myocardial infarction; CHD, coronary heart disease; TIA, transient ischemic attack; ICU, intensive care unit.

4.5.4 **Supplementary table:** Details of the subsequent validation studies by overlapping and/or different authors.

Derivation studies	**Validation studies by overlapping authors**						**Validation studies by different authors**					
First author (year); [Model]	**First author (year)**	**Outcome(s)**	**Study population / [Clinical condition]**	**Sample size**	**AUC (95% CI)**	**Calibration metrics**	**First author (year)**	**Outcome(s)**	**Study population / [Clinical condition]**	**Sample size**	**AUC (95% CI)**	**Calibration metrics**
Zhou K, et al. (2010)[218]; [**S index**]	Qu Y, et al. (2012)	Significant liver fibrosis; advanced liver fibrosis; cirrhosis	Chronic HBV/ [CI-related]	128	0.797 (ND); 0.881 (ND); 0.88 (ND)	ND	No	NA	NA	NA	NA	NA
Röllig C, et al. (2010)[219]; [**New AML model**]	No	NA	NA	NA	NA	NA	Djunic I, et al. (2013)[247]	overall survival	Elderly patients with AML / [malignancies]	120	ND	ND
Pisters R, et al. (2010)[12]; [**HAS-BLED**]	Lip GY, et al. (2011)[261]	major bleeding	atrial fibrillation / [CVD]	7329	0.65 (95% CI, 0.61-0.68)	Hosmer-Lemeshow statistic	Smith J, et al. (2012)[267]	major bleeding	ACS patients/ [CVD]	318	0.67 (95% CI, 0.54-0.79)	ND
Conti A, et al. (2010)[220]; [**Florence prediction rule**]	Conti A, et al. (2012)[242]	coronary stenosis at angiography or CV death, MI, angina, revasc.	Patients with chest pain / [CVD]	291	0.74 (ND)	ND	No	NA	NA	NA	NA	NA
Wishart GC, et al. (2010)[221]; [**PREDICT**]	Wishart GC, et al. (2011)[244]	all-cause mortality, breast-cancer specific mortality	Stage I or II invasive breast cancer / [cancer]	3140	all-cause mortality: 0.709 (ND); breast-cancer specific mortality: 0.723 (ND)	Predicted / Observed ratio	No	NA	NA	NA	NA	NA

Yuen MF, et al. (2009)[290] [**GAG-HCC score (5-year)**]	No	NA	NA	NA	NA	NA	Wong GL, et al. (2013)[246]	Hepatocellular carcinoma	Chronic HBV patients / [GI-related]	1531	0.76 (95% CI, 0.70-0.82)	ND
Yau T, et al. (2008)[223]; [**Advanced Liver Cancer Prognostic System**]	No	NA	NA	NA	NA	NA	Lin ZZ, et al. (2012)[248]	overall survival	Patients with hepatocellular carcinoma / [malignancies]	156	ND	ND
Hippisley-Cox J, et al. (2007)[31]; [**QRISK**]	Hippisley-Cox J, et al. (2008)[30]	CVD (MI, CHD, stroke, TIAs)	UK primary care population (QRESEARCH and THIN cohort) / [other]	QRESEARCH: 607733 (301622 (men), 306111 (women)); THIN: 1072800 (529813 (men), 542987 (women))	QRESEARCH: [men: 0.77 (95% CI, 0.7667-0.7733)][women: 0.7879 (95% CI, 0.7842-0.7915)]; THIN: [men: 0.7619 (95% CI, 0.7592-0.7646)][women: 0.7888 (95% CI, 0.7858-0.7918)]	Predicted / Observed ratio	No	NA	NA	NA	NA	NA
Harrison DA, et al. (2007)[224]; [**POP Score**]	No	NA	NA	NA	NA	NA	Juneja D, et al. (2010)[250]	severity and 30-day mortality	patients admitted to ICUs with severe acute pancreatitis / [GI-related]	55	AP severity: 0.73 (95% CI, 0.56-0.90); 30-day mortality: 0.86 (95% CI, 0.75-0.88)	ND

Harrison DA, et al. (2007)[291]; [**ICNARC model**]	Nolan JP, et al. (2007)[263]	hospital mortality	mechanically ventilated patients admitted to ICUs after cardiac arrest / [CVD]	24132	ND	ND	Chen JC, et al. (2012)[268]	mortality	acute dialysis patients / [other]	529	0.836 (0.797-0.874)	Hosmer-Lemeshow statistic
Colinet B, et al. (2005)[226]; [**SCS**]	Jacot W, et al. (2008)[262]	survival	non-small-cell lung cancer patients / [cancer]	301	0.56 (ND)*	ND	Kuo YW, et al. (2011)[269]	overall survival	non-small cell lung cancer patients / [cancer]	172	0.62 (ND)*	ND
Sanchis J, et al. (2005)[227]; [**New Risk Score**]	Sanchis J, et al. (2006)[265]	primary end point: mortality or MI; secondary end point: mortality, MI, or urgent revascularization	patients with acute chest pain of possible coronary origin, with normal ECG and troponin levels / [CVD]	340	Mortality or MI: 0.70 (ND)*; MACCE: 0.71 (ND)*	ND	Manini AF, et al. (2009)[270]	mortality or MI	patients with acute chest pain, with normal ECG and troponin levels / [CVD]	148	0.64 (ND)*	ND
Tateishi R, et al. (2005)[228]; [**Tokyo score**]	No	NA	NA	NA	NA	NA	Kondo K, et al. (2007)[252]	survival	patients with HCC underwent hepatectomy / [cancer]	235	0.638 (ND)	ND
Donati A, et al. (2004)[229]; [**ASA status-based model**]	No	NA	NA	NA	NA	NA	Haga Y, et al. (2011)[249]	in-hopsital mortality	patients underwent elective procedures / [other]	5272	0.73 (95% CI, 0.63-0.83)	ND

Ho GT, et al. (2004)[230]; [**Ho index**]	No	NA	NA	NA	NA	NA	Aceituno M, et al. (2008)[251]	need of early surgery	patients with steroid-refractory ulcerative colitis / [GI-related]	34 and 38	0.79 (95% CI, 0.59-0.99) and 0.74 (95% CI, 0.53-0.96)	ND
Olsson T, et al. (2004)[231]; [**REMS**]	No	NA	NA	NA	NA	NA	Goodacre S, et al. (2006)[257]	in-hospital mortality	Patients brought to the ED / [other]	5583	0.74 (95% CI, 0.70-0.78)	ND
Gallamini A, et al. (2004)[232]; [**PTCL-U model**]	No	NA	NA	NA	NA	NA	Yang DH, et al. (2009)[254]	overall survival, progression-free survival	peripheral T-cell lymphoma / [malignancies]	64	0.73 (ND)*	ND
Villella M, et al. (2003)[233]; [**GISSI-2 index**]	No	NA	NA	NA	NA	NA	Valeur N, et al. (2007)[256]	mortality, reinfraction and/or death	STEMI patients treated with primary PCI or fibrinolysis / [CVD]	1117	mortality: 0.725 (ND)	ND
Sedrine WB, et al. (2002)[234]; [**OSIRIS**]	No	NA	NA	NA	NA	NA	Cook RB, et al. (2005)[258]	osteoporosis	Postmenopaus al women / [CVD]	208	0.747	ND
Josting A, et al. (2002)[235]; [**GRHS**]	No	NA	NA	NA	NA	NA	Smeltzer JP, et al. (2011)[253]	event-free survival, overall survival	relapsed or refractory Hodgkin lymphoma / [cancer]	46	ND	ND

LeMaire SA, et al. (2001)[236]; [**model for elective TAAA repair**]	No	NA	NA	NA	NA	NA	Siegenthaler MP, et al. (2008)[255]	mortality	patients underwent elective or emergency endovascular repair of TAAA / [CVD]	21	ND	ND
Wong DT, et al. (1999[237]; [**Cardiac Risk Scores**]	No	NA	NA	NA	NA	NA	Yende S, et al. (2002)[259]	mortality, prolonged mechanical ventilation, time to wean	patients undergoing conventional CABG or off-pump CABG / [CVD]	345	0.65 (ND)*	ND
Pirovano M, et al. (1999)[292]; [**PaP Score**]	Tassinari D, et al. (2008)[264]	all-cause mortality	advanced cancer patients / [cancer]	173	0.59 (ND)*	ND	Glare P, et al. (2001)[271]	all-cause mortality	patients with advanced disease (cancer and non-cancer) / [other]	100	0.70 (ND)*	ND
Hasford J, et al. (1998)[239]; [**The Euro Score for CML**]	Hasford J, et al. (2005)[266]	survival	patients with Philadelphia chromosome-positive CML / [cancer]	773	0.65 (ND)*	ND	Bonifazi F, et al. (2000)[272]	survival	patients with Philadelphia (Ph)-positive CML / [cancer]	272	0.59 (ND)*	ND
Roberts AB, et al. (1985)[240]; [**model for prediction fetal birth weight**]	No	NA	NA	NA	NA	NA	Simon NV, et al. (1987)[260]	fetal weight	pregnant women who delivered within 72 h of an ultrasound evaluation / [other]	259	ND	ND

Lenstrup C, et al. (1982)[241]; [**CTG scoring system**]	Lenstrup C.. (1982)[245]	intra-uterine growth	women in 28th-30th week of pregnancy received weekly CTGs until delivery / [other]	111	ND	ND	No	NA	NA	NA	NA	NA

* AUC was not given by the authors and was estimated as described in the Methods.

AUC, area under the curve; CI, confidence interval; MI, myocardial infarction; CAD, coronary artery disease; CVD, cardiovascular disease; ACS, acute coronary syndrome; TLF, target lesion failure; AML, acute myeloid leukemia; CV, cardiovascular; STEMI, ST-elevation myocardial infarction; PCI, percutaneous coronary intervention; CABG, coronary artery bypass graft; ND, no data; HAS-BLED, Hypertension, Abnormal renal/liver function, Stroke, Bleeding history or predisposition, Labile international normalized ratio, Elderly (>65 years), Drugs/alcohol concomitantly; POP score, Pancreatitis Outcome Prediction Score; ICNARC model, Intensive Care National Audit & Research Centre model; SCS, Simplified Comorbidity Score; REMS, Rapid Emergency Medicine score; PTCL-U, Peripheral T-cell lymphoma unspecified model; GRHS; German Relapsed Hodgkin Prognostic Score; PaP score, Palliative Prognostic Score; POCE, patient-oriented composite endpoint; NA, not applicable; CHD, coronary heart disease; TIAs, transient ischemic attack; ICU, intensive care unit; AP, acute pancreatitis; HCC, hepatocellular carcinoma; TAAA, thoraco-abdominal aneurysm; CML, chronic myeloid leukemia; CTG, cardiotocography.

4.5.5 Supplementary table: Characteristics of the newly developed risk prediction models.

Derivation study	Model	Number of included variables	Set of included variables	Outcome(s)
Models without Further Validation				
Hippisley-Cox J, et al. (2010)[S1]	QRISK	10	age, area-based index of deprivation, body-mass index, diabetes, family history, hypertension-related medications, sex, systolic blood pressure, smoking, total cholesterol/HDL ratio	CVD
Stamatopoulos B, et al. (2010)[S2]	quantitative PCR score for CLL	3	ZAP70, LPL, microRNA-29c	treatment-free survival, overall survival
Campbell HE,et al. (2010)[S3]	number of positive axillary lymph nodes, tumour grade and size, age	4	number of positive axillary lymph nodes, tumour grade, tumour size, patient age	first recurrent event
Maluenda G, et al. (2010)[S4]	Novel PCI risk score	8	TIMI grade 3 flow after PCI, history of heart failure, left main coronary artery disease, chronic renal failure, diabetes mellitus, hematocrit decrease after PCI and hematocrit at baseline, age	1-year mortality
Chung JW, et al. (2010)[S5]	model to predict development of severe ischemic colitis	3	tachycardia, shock within 24 hours, the presence of ulcer	severe ischemic colitis
Drenthen W, et al. (2010)[S6]	modified risk score for cardiac complications during completed pregnancies in women with congenital heart disease	8	histoy of arrhythmias, cardiac medication before pregnancy, NYHA class prior to pregnancy, left heart obstruction, systemic atrioventricular valve regurgitation, pulmonary atrioventricular regurgitation, mechanical valve prosthesis, cyanotic heart disease (correcte/uncorrected)	composite cardiac complications

Phillips AA, et al. (2010)[S7]	new prognostic model for patients with HTLV-1-associated ATLL	ND	new risk model (International Prognostic Index (IPI) and The Prognostic Index for peripheral T-cell lymphoma unspecified (PTCL-U) (PIT))	overall response rate, overall survival
Hsu CY, et al. (2010)[S8]	Taipei Integrated Scoring System	7	TTV based model (total tumor volume (TTV) – Child–Turcotte–Pugh (CTP) – a-fetoprotein (AFP) model)	survival
Röllig C, et al. (2010)[S9]	novel prognostic model in elderly patients with acute myeloid leukemia	6	karyotype, age, NPM1 mutation status, white blood cell count, lactate dehydrogenase, CD34expression	disease-free survival
Nowak AK, et al. (2010)[S10]	prognostic nomogram for malignant pleural mesothelioma	3	total glycolytic volume on FDG-PET scan, weight loss, pleurodesis	survival
Elley CR, et al. (2010)[S11]	Diabetes Cohort Study risk prediction model	9	age, sex, duration of known diabetes, systolic blood pressure, smoking status, total cholesterol/HDL ratio, ethnicity, A1C, albumin/creatinine ratio	first fatal or nonfatal cardiovascular event
Ananthakrishnan AN, et al. (2010)[S12]	simple quantitative risk score to measure the severity of Crohn's disease hospitalizations	10	disease phenotype (inflammatory, obstructing, fistulizing), anemia, requirement blood transfusion, malnutrition, total parenteral nutrition, volume depletion, Clostridium difficile infection, admission to a teaching hospital, interhospital transfer	severe hospitalization course
Itoh K, et al. (2010)[S13]	simple prognostic model for Hodgkin Lymphoma	2	male, elevated serum lactate dehydrogenase	overall survival
Ananthakrishnan AN, et al. (2010)[S14]	colectomy risk score	6	anemia, requirement for blood transfusion, malnutrition, total parenteral nutrition, transfer from outside hospital, admission to teaching hospital	total colectomy

Andersson B, et al. (2010)[S15]	Gastrointestinal Complication Score (GICS)	9	age>80 y, active smoker, inotropic support, NYHA class III-IV, cardiopulmonary bypass time>150 min, postoperative atrial fibrillation, postoperative heart failure, reoperation due to bleeding, postoperative vascular complication	gastrointestinal complications after cardiac surgery; postoperative myocardial infarction; neurological dysfunction; infection
Hernández D, et al. (2009)[S16]	model for 3 year mortality in post kidney transplant patients	8	age, pretransplant diabetes, positive hepatits C virus (HCV) antibodies, new onset of diabetes after transplantantion at the first year, serum creatinine at the first year, proteinuria >1g at the first year, use of tacrolimus at the first year, use of mycophenolate mofetil at the first year	mortality
Menza TW, et al. (2009)[S17]	Full and simple HIV acquisition model for 1- and 4-years	6 and 4 (age and race/ethnicity excluded)	age<40 years, non-white and non Asian/Pacific Islander race/ethnicity, current laboratory diagnosis of a bacterial STD (gonorrhea, Chlamydia, or early syphilis) or of having ever had a bacterial STD, use of methamphetamine or inhaled nitrites in the prior 6 months, >10 male sex partners in the prior year, and unprotected anal intercourse with a partner of unknown or positive HIV status (nonconcordant unprotected anal intercourse, UAI) in the prior year	HIV acquisition
Cattermole GN, et al. (2009)[S18]	Prince of Wales Emergency Department Score (PEDS)	6	systolic BP, GCS, glucose, HCO3, WBC, metastatic cancer history	30-day mortality

Iacob S, et al. (2009)[S19]	model to predict death on transplant list prior to receiving liver transplant	3	refractory ascites, hepatorenal syndrome, Model for End Stage Renal (MELD) score	death
Kim HK, et al. (2010)[S20]	simple assessment tool for better early bedside risk stratification for both short- and long-term clinical outcomes	3	TIMI risk index, Killip class, serum creatinine	death from any cause
Moore L, et al. (2009)[S21]	Trauma Risk Adjustment Model (TRAM)	6	Abbreviated Injury Scale (AIS) code, Glasgow coma score (GCS), systolic blood pressure, respiratory rate, age, number of comorbidities	hospital mortality
Suh SY, et al. (2010)[S22]	Objective Prognostic Score (OPS)	7	anorexia, resting dyspnea, low performance status, leukocytosis, elevated serum bilirubin, elevated serum creatinine, elevated serum LDH	survival
Bria E, et al. (2009)[S23]	clinical prognostic score in non-small-cell lung cancer	2	number of resected lymph-nodes, node ratio (ratio between thenumber of positive nodes andnumber of removed nodes)	overall survival, cancer-specific survival, disease-free survival
Negassa A, et al. (2009)[S24]	tree-structured prognostic classification for postprocedural complications after PCI for aMI	4	cardiogenic shock, congestive heart failure, age, diabetes	postprocedural complications (in-hospital death, stroke, or CABG surgery)
Capodanno D, et al. (2009)[S25]	DERIVATION score	5	baseline left ventricular ejection fraction <50%, angioplasty in the setting of acute coronary syndromes, bifurcation lesion, left anterior descending as target vessel, multiple stenting	stent thrombosis

Hernandez DJ, et al. (2009)[S26]	logistic-regression based model	8	African-American race, age, ln(PSA), family history, digital rectal examination, number of biopsy cores, previous negative biopsy, %fPSA	prostate cancer, high-grade prostate cancer (Gleason score≥7)
Yuen MF, et al. (2009)[S27]	GAG-HCC score (5- and 10-year) ("Guide with Age, Gender, HBV DNA, Core promoter mutations and Cirrhosis")	5 and 4 (excluding core promoter mutations)	age, gender, HBV DNA levels, core promoter mutations, cirrhosis	hepatocellular carcinoma
Nobre SR, et al. (2008)[S28]	predictive model for in-hopsital mortality for patient with end-stage liver disease	2	Model for end-stage liver disease (MELD), age	in-hospital mortality
Kitai S, et al. (2008)[S29]	The biomarker combined Japan Integrated Staging (bm-JIS) score	4	a-fetoprotein (AFP), Lens culinaris agglutinin-reactive AFP, des-carboxy prothrombin, conventional JIS score	overall survival
García-Almagro FJ, et al. (2008)[S30]	improved TIMI risk score with diabetes and EF	3	The Thrombolysis inMyocardial Infarction (TIMI) risk score, left ventricular ejection fraction, diabetes	cardiac events (MI, revascularization, cardiac death)
Tice JA, et al. (2008)[S31]	Breast Cancer Surveillance Consortium breast density model algorithm	5	age, race or ethnicity, breast density, history of breast cancer in a first-degree relative, history of a breast biopsy	breast cancer (invasive cancer and ductal carcinoma in situ)
Mountzios G, et al. (2008)[S32]	prognostic factors in RMEC after taxane-based chemotherapy	2	performance status, relapse within the field of previous RT	mortality

Tsimberidou AM, et al. (2007)[S33]	New prognostic score	5 and 6	Deletion 17p or 6q with or without other cytogenetic abnormalities, age >60 years, relative, history of a breast bio3.5 g/dL, creatinine>1.6 mg/dL, +/- RAI stage OR +/- BINET stage	survival
Kim SJ, et al. (2007)[S34]	New prognostic score	2	Ki-67 expression, the primary site of involvement	overall and disease-free survival
Xu X, et al. (2007)[S35]	post-transplant model	5	total bilirubin, aspartate aminotransferase, international normalized ratio, serum creatinine and blood urea nitrogen	3, 6 months and 1 year post-transplantation mortality
Aletti GD, et al. (2007)[S36]	risk-adjusted, multicenter outcomes model for ovarian cancer surgery	7	age, ASA, albumin, surgical complexity score, stage/age, ASA/ASA surgical complexity score	30 days morbidity, 3 months mortality, chemotherapy non-feasible
Cianchi F, et al. (2007)[S37]	model to identify patients with high-risk Stage IIA colorectal cancer	2	tumor growth pattern, extent of tumor spread beyond muscularis propria	mortality
Tischendorf JJ, et al. (2007)[S38]	PSC (Primary sclerosing cholangitis) score	7	age, low albumin, persistent bilirubin elevation longer than 3 months, hepatomegaly, splenomegaly, dominant bile duct stenosis, and intra- and extrahepatic ductal changes at the time of diagnosis	survival

Senni M, et al. (2006)[S39]	CardioVascular Medicine Heart Failure (CVM-HF) index	13	age, anemia, hypertension, chronic obstructive pulmonary disease, complicated diabetes mellitus, moderate to severe kidney dysfunction, metastatic cancer, no b-blockers, no angiotensin-converting enzyme inhibitors/ angiotensin receptor blockers, New York Heart Association class III/IV, left ventricular ejection fraction <20%, severe valvular heart disease, atrial fibrillation	mortality
Ballesteros MA, et al. (2007)[S40]	model for early apoptosis rate	4	early apoptosis rate induced with regional serum, GCS, APACHE II score and pupil abnormalities	6-months mortality
Park YH, et al. (2006)[S41]	HG-PGL risk model	4	advanced age, male gender, higher LDH, ascites	overall survival
Went P, et al (2006)[S42]	modified Prognostic Index for T-cell lymphoma (including patient and tumor-specific factors)	4	age (>60 years), high lactate dehydrogenase, poor performance status, and Ki-67>= 80%)	disease-specific survival
Williams BA, et al. (2006)[S43]	a new simplified immediate prognostic risk score	8	age, sex, systolic blood pressure, admission serum creatinine, extent of ST segment depression, QRS duration, Killip class, and infarct location	1-month mortality
Tekesin I, et al. (2005)[S44]	CLEOPATRA I, CLEOPATRA II	2	cervical length (≤ 2.5 cm), previous preterm delivery/ fetal fibronectin, previous preterm delivery	preterm delivery
Miyake Y, et al. (2005)[S45]	prognostic scoring model for liver transplantation	4	cause of fulminant hepatic failure (hepatitis B virus or indeterminate), hepatic coma grade (III or IV), systemic inflammatory response syndrome (yes) and ratio of total to direct bilirubin (>2.0)	2-weeks survival and death

Alvares CL, et al. (2005)[S46]	multiparametric risk-adapted model	2	response to treatment at baseline and after 3m	overall survival
Meyer S, et al. (2005)[S47]	mortality risk score for paediatric cancer patients admitted to ICU	6	non-solid tumour, number of organ failures (n>2), neutropenia, septic shock, mechanical ventila- tion, and inotropic medication	non-survival
Sadanandan S, et al. (2004)[S48]	CABG risk score	6	history of CABG, troponin(+), prior angina, ST-segment deviation >0.5 mm, history of peripheral arterial disease, male gender	in-hospital coronary artery bypass graft surgery (CABG)
Skírnisdóttir I, et al. (2004)[S49]	prognostic model for early stage epithelial ovarian carcinoma	3	tumor grade, p53 status, and EGFR status	disease-free survival
Haferlach T, et al. (2004)[S50]	cytogenetically based risk score	3	AML with t(8;21), inv(16), or t(16;16); the unfavorable-prognosis group contained AML with aberrations of chromosomes 5 or 7, aberra- tions of 11q23, 12p, or 17p, inv(3), t(3;3), or with a complex aberrant karyotype (i.e. 3 chromosomes involved)	complete remission, overall survival, event-free survival, relapse-free survival
Halonen KI, et al. (2003)[S51]	Severe Acute Pancreatitis (SAP) score	4	age, highest serum creatinine value within 60-72 h from primary admission, need for mechanical ventilation, chronic health status	mortality
von Eyben FE, et al. (2003)[S52]	risk model for survival prediction in patients with testicular germ cell tumor	2	International Germ Cell Consensus Classification (IGCCC) and isoenzyme 1 catalytic concentration (S-LD-1)	survival
Freedland SJ, et al. (2002)[S53]	risk stratification model after radical prostatectomy	2	PSA density and biopsy Gleason score	adverse pathological features or biochemical recurrence

Castellanos-Ortega A, et al. (2002)[S54]	New risk score	7	cyanosis, Glasgow coma scale <8, refractory hypotension, oliguria, leukocytes <4000/mm3, partial thromboplastin time more than 150% of control value, base deficit >10 mmol/l	mortality
Smith MA, et al. (2002)[S55]	novel predictive model of outcome in de novo AML	3	SFM S-phase activity, S-phase activity following exposure to G+GM-CSF and karyotype	complete remission
Josting A, et al. (2002)[S56]	German Hodgkin's Lymphoma Study Group (GHSG) database risk score	3	time to recurrence, clinical stage at relapse, anemia at relapse	freedom from second failure
Djoulah S, et al. (1999)[S57]	model based on the characteristics of the relevant pockets of HLA-DR and -DQ molecules	2	HLA- DRB1 and -DQB1 – P4 and P9 pockets	risk susceptibility prediction of insulin-dependent type I diabetes
Wong DT, et al. (1999)[S58]	Cardiac Risk Score (preoperative and postoperative risk factors)	3	time to recurrence, clinical stage at relapse, anemia at relapse	prolonged intensice care unit stay; delayed extubation
Duong DH, et al. (1998)[S59]	risk score for unfavorable outcome of surgery for cebrebral aneurysm	5	vasospasm, clinical status, age, associated medical problems, technical difficulties	patient outcome at discharge
Adler M, et al. (1997)[S60]	Erasme score	5	encephalopathy, alkaline phosphatase, bilirubin, cholinesterase, bile acids	1-year liver related mortality
Paganini EP, et al. (1996)[S61]	mortality prediction model for patients requiring hemodialysis in ICU	8	male gender, respiratory failure, haematologic dysfunction, bilirubin, surgery, creatinine, failed organ, BUN	mortality
Sarbia M, et al. (1995)[S62]	New prognostic score	2	pattern of invasion, inflammatory response	survival

Ménard S, et al. (1994)[S63]	New prognostic score	4	tumour size, grading, laminin receptor, c-erbB-2 overexpression	overall survival
Moran MR, et al. (1992)[S64]	DNA ploidy based model	3	>3 positive lymph nodes, invasion of the tumor through the wall, percentage of cells in 4C peak (GzM phase)-diploid nuclei	development of distant metastases
Altaca G, et al. (1992)[S65]	score to predict survival after perforated duodenal ulcer	4	coexisting medical illness, male sex, white cell count, acute renal failure	mortality
Taussig LM, et al. (1973)[S66]	clinical score to predict mortality in cystic fibrosis	15	set of variables divided as pulmonary, general, blood gases, gastointestinal, miscalleneous	mortality
Further Validated Models				
Zhou K, et al.[218]	S index	3	γ-glutamyltransferase, platelets, albumin	significant liver fibrosis; advanced liver fibrosis; cirrhosis
Röllig C, et al. (2010)[219]	novel prognostic model in elderly patients with acute myeloid leukemia	6	karyotype, age, NPM1 mutation status, white blood cell count, lactate dehydrogenase, CD34expression	overall survival
Pisters R, et al.[12]	HAS-BLED	7	hypertension, abnormal renal/liver function, stroke, bleeding history or predisposition, Labile international normalized ratio, elderly (> 65 years), drugs/alcohol concomitantly	major bleeding
Conti A, et al. (2010)[220]	Florence prediction rule	5	chest pain score >6, metabolic syndrome, age >50y, diabetes mellitus, gender (male)	6-month composite endpoint (CV death, nonfatal MI, revascularization)

Wishart GC, et al.[221]	PREDICT	6	number of positive nodes, tumor size, tumor grade, detection by screening, chemotherapy, hormone therapy	all-cause mortality, breast-cancer specific mortality
Yuen MF, et al.[290]	GAG-HCC score (5-year) ("Guide with Age, Gender, HBV DNA, Core promoter mutations and Cirrhosis")	4	age, gender, HBV DNA levels, cirrhosis	Hepatocellular carcinoma
Yau T, et al. (2008)[223]	Advanced Liver Cancer Prognostic System (ALCPS)	11	ascites, abdominal pain, weight loss, Child-Pugh grade, alkaline phosphatase, total bilirubin, alpha-fetal protein, urea, portal vein thrombosis, tumor size, presence of lung metastases	3-month survival
Hippisley-Cox J, et al.[31]	QRISK	11	age, area-based index of deprivation, BMI, diabetes, family history, HTN-related medications, sex, SBP, smoking, TC/HDL ratio	CVD (MI, CHD, stroke, TIAs)
Harrison DA, et al.[224]	Pancreatitis Outcome Prediction (POP) Score	6	arterial pH, age, serum urea, mean arterial pressure, PaO2/FIO2 ratio, total serum calcium	mortality
Harrison DA, et al.[291]	Intensive Care National Audit & Research Centre (ICNARC) model	16	highest heart rate, lowest systolic BP, highest temperature, lowest respiratory rate, PaO2/FIO2 ratio (ventilation), lowest arterial pH, highest serum urea, highest serum creatinine, highest serum sodium, urine output (24 hrs), lowest WBC, sedated/paralyzed or GCS, age, diagnostic category coefficients and interactions with the physiology score, cardiopulmonary resuscitation within 24 hrs before admission, source of admission	mortality

Colinet B, et al.[226]	Simplified Comorbidity Score (SCS)	7	tobacco consumption, diabetes mellitus and renal insufficiency, respiratory, neoplastic and cardiovascular comorbidities, alcoholism	survival
Sanchis J, et al.[227]	New Risk Score for Patients With Acute Chest Pain, Non–ST-Segment Deviation, and Normal Troponin Concentrations	5	chest pain score>10 points, >2 chest pain episodes in last 24 h, age>67 years, insulin-dependent diabetes mellitus, prior PTCA	mortality or MI (primary end point); mortality, MI, or urgent revascularization (secondary end point)
Tateishi R, et al.[228]	Tokyo score	4	serum albumin, bilirubin, size of tumours, number of tumours	death
Donati A, et al.[229]	ASA status-based model	4	ASA status, age, type of surgery (elective, urgent, emergency), degree of surgery (minor, moderate or major)	death or survival at hospital discharge
Ho GT, et al.[230]	Ho index	3	mean stool frequency, colonic dilatation within the first 3 days, hypo-albuminaemia	response (no colectomy) or non-response to medical therapy (colectomy)
Olsson T, et al.[231]	Rapid Emergency Medicine score (REMS)	6	coma, respiratory frequency, oxygen saturation, blood pressure, pulse rate, age	in-hospital mortality
Gallamini A, et al.[232]	a new prognostic model for Peripheral T-cell lymphoma unspecified (PTCL-U)	4	age, performance status equal to or more than 2, LDH level, bone marrow involvement	overall survival
Villella M, et al.[233]	GISSI-2 index	6	low work capacity (<100 watts on the bicycle ergometer or <6 minutes on the treadmill), exercise-induced symptomatic ST-segment depression, low double product (product of peak heart rate and SBP <=21,700), early left ventricular failure, recovery phase left ventricular dysfunction, electrical instability	mortality

Sedrine WB, et al.[234]	Osteoporosis Index of Risk (OSIRIS)	4	age, body weight, current hormone replacement therapy use, history of previous low impact fracture	risk of osteoporosis
Josting A, et al.[235]	German Relapsed Hodgkin Prognostic Score (GRHS)	3	time to recurrence, clinical stage at relapse, anemia at relapse	overall survival
LeMaire SA, et al.[236]	preoperative risk factors to predict an adverse outcome after elective thoracoabdominal aortic aneurysm repair: preoperative renal insufficiency, increasing age, symptomatic aneurysms, extent II aneurysms	4	preoperative renal insufficiency, increasing age, symptomatic aneurysms, extent II aneurysms	composite end-point: death within 30 days, death before discharge from the hospital, paraplegia, paraparesis, stroke, or acute renal failure requiring dialysis
Wong DT, et al.[237]	Cardiac Risk Score (preoperative and postoperative risk factors)	3	left ventricle grade 4, emergency surgery, female gender	mortality
Pirovano M, et al.[292]	Palliative Prognostic Score (PaP Score)	6	Clinical Prediction of Survival (CPS), Karnofsky Performance Status (KPS), anorexia, dyspnea, total white blood count (WBC), lymphocyte percentage	all-cause mortality
Hasford J, et al.[239]	The Euro Score for CML	6	age, spleen size, blast count, platelet count, eosinophil count, basophil count	survival
Roberts AB, et al.[240]	Prediction model using ultrasound parameters to predict fetal birth weight	3	head circumference, abdominal circumference, femur length	fetal weight

Lenstrup C.[241]	CTG scoring system (Lenstrup)	6	reduced variability, reduced variability/silent pattern, bradycardia or tachycardia present, early or uncharacteristic decelerations present, variable declerations present, late declerations present	intra-uterine growth

CVD, caridovascular disease; PCR, polymerase chain reaction; CLL, chronic lymphocytic leukemia; PCI, percutaneous coronary intervention; GI, gastrointestinal; HTLV-1, human T-cell lymphotropic virus type-1; ATLL, adult T-cell leukemia/lymphoma; AML, acute myeloid leukemia; CV, cardiovascular; MI, myocardial infarction; CAD, coronary artery disease; NSTEMI, non-ST-elevation myocardial infarction; CABG, coronary artery bypass graft; DES, drug-eluting stent; ICU, intensive care unit; HLA, human leukocyte antigen; MACCE, major adverse cardiac and cerebrovascular event; BMI, body mass index; HTN, hypertension; SBP, systolic blood pressure; TC, total cholesterol; HDL, high-density lipoprotein; CHD, coronary heart disease; TIAs, transient ischemic attacks; BP, blood pressure; GCS, Glasgow Coma Scale; CML, chronic myelogenous leukemia; CTG, cardiotocography.

Section 5.
Summary

Section 5.1 Summary and conclusions

More people now live with disease and conditions that impair health than at any other time in history; prognosis research provides crucial evidence for translating findings from the laboratory to humans, and from clinical research to clinical practice. Prognosis research seeks to understand and improve future outcomes in people with a given disease or health condition.

Although typically in medical terms prognosis refers to the most likely clinical course of a diseased patient, the term is also applied to the prediction of future risk in a normal population. Except in rare instances, both of these settings include a stochastic element, one that is subject to chance.[294] Prognostication and prediction involve estimating risk, or the probability of a future event or state. The outcome not only is unknown, but does not yet exist, distinguishing this task from diagnosis. Therefore, prognostic models, the core-tool of prognostication, add the element of time.[293] Clinically, prognostic models are often used for risk stratification, or for assigning levels of risk, such as high, intermediate, or low, which may then form the basis of treatment decisions.

Models for prognostic risk prediction have been widely used in the cardiovascular field to predict risk of future events or to stratify apparently healthy individuals into risk categories. Appropriate model assessment is critical to the determination of clinical impact and to guideline development. Prediction tools are useful only when they are easily accessible at the point of care, which is why for most of them there is also designed an online calculator. Such calculators are implemented in electronic patient records, electronic order entry systems, or smartphone or tablet applications. Overall, prediction models that include age, sex, symptoms, and risk factors allow for accurate estimation of the probability of coronary artery disease in low prevalence populations. The addition of single predictors in previously established models requires specific statistical approach and verification. Implementation of such updated models can improve clinical outcomes, but need further evaluation in individual models' level.

The country-specific predictions for estimated 10 year cardiovascular disease burden are striking, particularly areas with large proportion of high-risk individuals. A next step would be to quantify the positive effects on a population level if these prediction models and subsequent risk based preventative management were used in these countries. By use of so-called population-level linked-evidence models, estimates of country-specific 10 year cardiovascular disease-risk groups can be combined with known effect sizes from randomised trials of various treatments (eg, lipid-lowering and blood-pressure-lowering drugs), supplemented with treatment adherence figures, to quantify the expected decrease in cardiovascular disease burden per country within 10 years. These predictions might further help, and indeed convince, decision-makers across the world to decide on wide-scale introduction of risk-based management for cardiovascular disease.

Prognostic tools should be evaluated in several sequential stages: initial model performance (model development), prospective validation in independent cohorts (external validation of a model), impact on patient management and outcome and cost-effectiveness. However, even for established and widely used prognostic tools, many of these steps suffer from methodological limitations and in many cases are missing. Moreover, it is imporantant to highlight the paucity of evidence around their impact on patient management and clinical outcomes. Such important evidence would ideally come from randomized control trials (RCTs), which compare the outcomes of patients whose management is guided by the proposed prognostic tool with the outcomes of patients who are managed without it. However, there are so many prognostic tools, that it is impossible to evaluate all of them in RCTs. Efforts should focus around those with most promising results. In selecting which models to test in randomized trials, one may wish to consider not only satisfactory, validated discriminating ability, but also what is the respective change in disease management that can be anticipated; how effective are the available preventive or treatment interventions for the disease and how much room exists for improvement; what is the expected cost to get the information required for building the model, and to implement it in practice; and how likely it is that the model can be used widely by non-expert health practitioners. Going through such a checklist is likely to eliminate the large majority of proposed prognostic models. Nevertheless, there are currently no randomized trials assessing the implementation of any cardiovascular prediction models. Such studies should be encouraged. A more through and systematic research agenda would be useful to build surrounding late implementation issues, including ease of use, and impact on resources in diverse settings.

The bottom line is that the best test of a prediction model is not accuracy but improved clinical outcomes. Compared with clinician judgment, a prediction model might improve diagnostic accuracy, reduce costs and harms, and lead to improved health outcomes. Documenting this benefit requires RCTs in which providers are randomized to use the proposed prediction model or not, and the outcome is improved health. Very few models have been tested in this way.[46,98] Prediction of risk is not enough—we need evidence that prediction can lead to actions that reduce risk beyond what would occur without the prediction rule.

Section 5.2 Περίληψη

Η Ιατρική αναπτύχθηκε με βάση τη φιλοσοφία και υπέστη τον αυστηρό έλεγχο της λογικής, ενώ η θεραπευτική τέχνη εξελίχθηκε με βάση τον ορθολογισμό. Η διδασκαλία της Ιατρικής πράξης έδινε ιδιαίτερη έμφαση στην άμεση παρατήρηση της φύσης και την επαγωγική λογική. Η διδασκαλία βασιζόταν στη διάγνωση και στην εξέταση του ασθενούς, ενώ οι ιατροί ενδιαφέρονταν όχι τόσο για την διερεύνηση των αιτιών των νόσων, όσο για την πρόγνωση της νόσου κάθε επιμέρους ασθενούς. Για τον ιπποκρατικό γιατρό η πρόγνωση ήταν πολύ σημαντική. Κάθε μεταβολή καταγραφόταν προσεκτικά. Για παράδειγμα, το "ιπποκρατικό προσωπείο", όπως εξακολουθεί και σήμερα να ονομάζεται, αποτελούσε βαρύ προγνωστικό σημείο. «Τον ιητρόν δοκέει μοι άριστον είναι πρόνοιαν επιτηδεύειν», δηλαδή για έναν γιατρό το πιο σημαντικό, κατά τη γνώμη μου είναι να μπορεί να κάνει προγνώσεις. Έτσι αρχίζει "το προγνωστικόν" ένα από τα έργα της Ιπποκρατικής συλλογής και συνεχίζει: Πρόγνωση είναι κάθε διαδικασία που αποσκοπεί στην εκτίμηση της απόλυτης ή σχετικής πιθανότητας ιατρικών ή άλλων εκβάσεων στο μέλλον με βάση γνωστές παραμέτρους που αφορούν άτομα ή πληθυσμούς.

Η πρόγνωση είναι ακρογωνιαία έννοια στην ιατρική σκέψη και πράξη για την υγεία των ατόμων και ακολούθως πληθυσμών. Τα προγνωστικά δεδομένα χρησιμοποιούνται για τον προσδιορισμό του κινδύνου συγκεκριμένων ατόμων με ή χωρίς θεραπεία, για την κατάταξη των ασθενών σε ομάδες με διαφορετικά επίπεδα κινδύνου, για την εκτίμηση ενδείξεων για την έναρξη της θεραπείας (με βάση κάποιο οριακό κίνδυνο πάνω από τον οποίο η θεραπεία ενδείκνυται), για την αξιολόγηση αλληλεπιδράσεων προγνωστικών παραγόντων με διάφορες θεραπείες (και κατά συνέπεια τη χρήση συγκεκριμένων εξατομικευμένων θεραπευτικών σχημάτων για διαφορετικούς ασθενείς), καθώς και το σχεδιασμό μελλοντικών μελετών με συνεκτίμηση των γνωστών προγνωστικών παραγόντων. Η εκτίμηση της πρόγνωσης είναι στην ουσία μια προσπάθεια για τον καλύτερο και πιο αποδοτικό καθορισμό και υπολογισμό των σχέσεων μεταξύ παραγόντων κινδύνου και έκβασης σε έναν πληθυσμό.

Η πρόγνωση είναι ιδιαίτερα σημαντική σε διάφορους τομείς (π.χ. περιβαλλοντικές συνθήκες, οικονομία) και σήμερα γίνεται ασφαλέστερα με την συμβολή διαφορετικών επιστημών, αξιοποιώντας έναν τεράστιο όγκο πληροφοριών. Η ασφαλής πρόγνωση αποδίδει πολυτρόπως θετικά αποτελέσματα, συνεπώς πρέπει να γίνεται προσεκτικά αναζητώντας τα ανάλογα τεκμήρια. Τότε η σημασία της δεν πρέπει σε καμία περίπτωση να υποτιμάται, αλλά να λαμβάνεται σοβαρά υπόψη. Αν, λοιπόν, είναι καθοριστική σε διάφορους τομείς όπως οι περιβαλλοντικές συνθήκες, πόσο μάλλον, αν πρόκειται για τον τομέα της υγείας, για την πορεία της ασθένειας, για την ποιότητα ζωής του κάθε ασθενούς. Περισσότεροι άνθρωποι ζουν σήμερα ενώ ασθενούν ή ζουν σε συνθήκες που επηρεάζουν την υγεία τους περισσότερο από κάθε άλλη στιγμή της ιστορίας. Η έρευνα στην οποία βασίζεται η πρόγνωση παρέχει σημαντικά τεκμήρια για την μετάβαση των όποιων ευρημάτων από το ερευνητικό εργαστήριο στους ανθρώπους, και από την κλινική έρευνα στην κλινική πράξη. Η έρευνα στον χώρο της πρόγνωσης στοχεύει στην κατανόηση και βελτίωση των μελλοντικών εκβάσεων σε ανθρώπους με συγκεκριμένες νόσους ή ανθρώπους που βρίσκονται σε συγκεκριμένη κατάσταση υγείας.

Η πρόγνωση στις περισσότερες περιπτώσεις γίνεται μέσω απλών δεικτών ή εργαλείων που ονομάζονται προγνωστικά μοντέλα. Τα προγνωστικά μοντέλα συνδυάζουν πολλαπλά χαρακτηριστικά με στόχο τον υπολογισμό της πιθανότητας εμφάνισης μίας συγκεκριμένης έκβασης στο μέλλον. Έτσι, τα προγνωστικά μοντέλα προσθέτουν το στοιχείο του χρόνου. Η πληροφορία αυτή δίνει την δυνατότητα στους κλινικούς ιατρούς να εξατομικεύουν τις θεραπευτικές τους παρεμβάσεις με βάση το προφίλ κινδύνου του κάθε ασθενούς. Αν και τυπικά όσο αφορά την πρόγνωση στην ιατρική, αυτή αναφέρεται στην κλινική πορεία/έκβαση ενός ασθενούς, ο όρος μπορεί ακόμη να χρησιμοποιηθεί για την πρόγνωση του μελλοντικού κινδύνου εμφάνισης ασθενειών σε φυσιολογικούς κατα τα άλλα πληθυσμούς - σε άτομα που την στιγμή της διερεύνησης με την εφαρμογή του προγνωστικού μοντέλου δεν εμφανίζουν κάποιο σύμπτωμα. Με εξαίρεση σπάνιες περιπτώσεις, και οι δύο προαναφερθέντες περιπτώσεις περιλαμβάνουν ένα στοχαστικό στοιχείο, το οποίο υπόκειται στην τύχη. Η διαδικασία της πρόγνωσης περιλαμβάνει τον υπολογισμό του κινδύνου ή της πιθανότητας ενός μελλοντικού συμβάμαντος ή κατάστασης. Επιπλέον η έκβαση όχι μόνο δεν είναι γνωστή αλλά ακόμη δεν υπάρχει, κάτι που διαφοροποιεί *την πρόγνωση από την διάγνωση. Στην κλινική πράξη ωστόσο, τα προγνωστικά μοντέλα χρησιμοποιούνται κυρίως για την ταξινόμηση των ατόμων σε διαφορετικά επίπεδα κινδύνου για μια συγκεκριμένη έκβαση (χαμηλού, μέτριου, υψηλού κινδύνου), τα οποία μπορεί να καθορισούν την βάση για την μετέπειτα εφαρμογή συγκεκριμένων θεραπειών και λήψη ιατρικών αποφάσεων.*

Τα προγνωστικά μοντέλα για την πρόγνωση συγκεκριμένων εκβάσεων έχουν ευρέως χρησιμοποιηθεί στο πεδίο των καρδιοαγγειακών νοσημάτων για την πρόγνωση μελλοντικών εκβάσεων ή για την κατηγοροποπίηση κατά τα άλλα υγειών ατόμων την στιγμή της κλινικής εκτίμησης σε συγκεκριμένες ομάδες κινδύνου για μελλοντικές εκβάσεις. Ιδιαίτερης σημασίας είναι η κατάλληλη αξιολόγηση των μοντέλων για τον προσδιορισμό του αντικύπου τους στην κλινική πράξη και την περαιτέρω εξέλιξη τους. Τα προγνωστικά μοντέλα είναι χρήσιμα μόνο όταν μπορούν εύκολα να εφαρμοστούν στην κλινική πράξη, γι'αυτό τον λόγο για τα περισσότερα από τα διαθέσιμα μοντέλα έχουν αναπτυχθεί εύχρηστες - φιλικές προς τον χρήστη υπολογιστικές μηχανές. Συγκεκριμένα προγνωστικά μοντέλα έχουν αναπτυχθεί που περιλαμβάνουν την ηλικία, το φύλο, τα συμπτώματα, και παράγοντες κινδύνου, και επιτρέπουν την ακριβή εκτίμηση της πιθανότητας για την εμφάνιση στεφανιαίας νόσου σε πληθυσμούς που βρίσκονται σε χαμηλο επιπολασμό. Η προσθήκη νέων μεταβλητών στα ήδη θεμελιωμένα προγνωστικά μοντέλα για την πρόγνωση εμφάνισης καρδιοαγγειακής νόσου απαιτεί συγκεκρίμενη μεθοδολογική προσέγγιση και κατάλληλη αξιολόγηση, πριν την εφαρμογή στην κλινική πράξη. Επιπλέον, οι εξατομικευμένες για κάθε χώρα προβλέψεις όσο αφορά τον υπολογιζόμενο δεκαετές φορτίο καρδιοαγγειακής νοσηρότητας ειναι εντυπωσιακές, ιδιαίτερα όσο αφορά γεωγραφικές περιοχές με υψηλό ποσοστό ατόμων υψηλού κινδύνου για καρδιοαγγειακή νόσο.

Το επόμενο στάδιο αφορά την ποσοτικοποίηση του όποιου θετικού αποτελέσματος σε επίπεδο πληθυσμού, εάν κάποιο προγνωστικό μοντέλο και ακολούθως η περαιτέρω διαχείριση του πληθυσμού βασιστεί σε προληπτικές παρεμβάσεις ανάλογα με τον προβλεπόμενο κίνδυνο σε αυτόν τον πληθυσμό. Ο συνδυασμός της πληροφορίας από προγνωστικά μοντέλα με την αποτελεσματικότητα *συγκεκριμένων θεραπευτικών παρεμβάσεων, μπορεί να συμβάλλει σημαντικά στην υιοθέτηση αποτελεσματικών παρεμβάσεων σε επίπεδο πληθυσμών για την διαχείριση της καρδιοαγγειακής νόσου.*

Όμως, ποια προγνωστικά εργαλεία πληρούν τις προϋποθέσεις για ευρεία εφαρμογή σε πληθυσμιακό επίπεδο; Τα προγνωστικά μοντέλα θα πρέπει να αξιολογηθούν σε διαδοχικά στάδια: η αρχική διακριτική ικανότητα του μοντέλου (ανάπτυξη του μοντέλου), προοπτική επικύρωση σε ανεξάρτητες κοόρτες (εξωτερική επικύρωση του μοντέλου), αξιολόγηση του αντίκτυπου όσο αφορά τον χειρισμό των ασθενών, κλινικές εκβάσεις, και κόστους-αποτελεσματικότητας από την χρήση των προγνωστικών μοντέλων. Ωστόσο, ακόμα και για εδραιωμένα και ευρέως χρησιμοποιούμενα προγνωστικά μοντέλα, πολλά από τα προαναφερθέντα στάδια αξιολόγησης έχουν μεθοδολογικούς περιορισμούς και σε αρκετές περιπτώσεις απουσιάζουν.

Ιδιαίτερης σημασίας είναι η έλλειψη ισχυρών τεκμηρίων όσο αφορά την αποτελεσματικότητα και τον αντίκτυπο της χρήσης των προγνωστικών μοντέλων στον χειρισμό των ασθενών (ή υγιών ατόμων που βρίσκονται σε κίνδυνο για κάποια συγκεκριμένη έκβαση με βάση το προγνωστικό μοντέλο) και σε κλινικές εκβάσεις. Τέτοια τεκμήρια θα μπορούσαν ιδανικά να προέρχονται από τυχαιοποιημένες κλινικές δοκιμές, οι οποίες θα συνέκριναν συγκεκριμένες εκβάσεις μεταξύ ασθενών των οποίων η θεραπευτικές παρεμβάσεις καθοδηγήθηκαν με βάση τον υπολογιζόμενο κίνδυνο από τη χρήση προγνωστικών μοντέλων, και ασθενών για τους οποίους δεν χρησιμοποιήθηκε η πληροφορία από τα προγνωστικά μοντέλα. Ωστόσο, υπάρχουν τόσα διαφορετικά προγνωστικά μοντέλα, τα οποία δεν μπορούν όλα να αξιολογηθούν μέσω τυχαιοποιημένων κλινικών δοκιμών. Οι προσπάθειες θα πρέπει να επικεντρωθούν στα πιο υποσχόμενα μοντέλα. Η διαδικασία επιλογής αυτών των μοντέλων που θα πρέπει να αξιολογηθούν σε επίπεδο τυχαιοποιημένων κλινικών δοκιμών, δεν πρέπει να βασιστεί μόνο στην διακριτική ικανότητα των μοντέλων ανεξαρτήτως επικύρωσης σε διαφορετικές μελέτες επικύρωσης, αλλά επιπλέον θα πρέπει κάποιος να συνυπολογίσει εάν υπάρχει η δυνατότητα και ποιες θα είναι οι πιθανές αλλαγές στην περαιτέρω θεραπευτική αντιμετώπιση που θα επιφέρει η χρήση κάποιου προγνωστικού μοντέλου; πόσο αποτελεσματικές είναι οι *παρούσες προληπτικές ή θεραπευτικές παρεμβάσεις και εάν υπάρχουν περιθώρια βελτίωσης; ποιο είναι το αναμενόμενο κόστος για την εφαρμογή του μοντέλου στην κλινική πράξη; και τέλος πόσο πιθανό είναι το συγκεκριμένο μοντέλο να μπορεί να χρησιμοποιηθεί ευρέως από μη-ειδικούς. Εφαρμόζοντας μία σειρά από τέτοια ερωτήματα είναι πολύ πιθανό να αποκλειστεί η μεγαλύτερη πλειοψηφία προγνωστικών μοντέλων. Παρόλα αυτά, δεν υπάρχουν μέχρι στιγμής τυχαιοποιημένες κλινικές δοκιμές που να αξιολογούν συγκεκριμένα την χρήση προγνωστικών μοντέλων για την εκτίμηση του καρδιαγγειακού κινδύνου. Εν κατακλείδι, το πιο αξιόπιστο μέτρο εκτίμησης ενός προγνωστικού μοντέλου δεν είναι η διακριτική ικανότητα του μοντέλου, αλλά η βελτίωση μέσω της εφαρμογής του των κλινικών εκβάσεων. Η πρόγνωση του κινδύνου δεν είναι αρκετή, απαιτούνται τεκμήρια ότι η πρόγνωση μπορεί να οδηγήσει σε ενέργειες που θα μειώσουν τον κίνδυνο σε μεγαλύτερο βαθμό σε σχέση με την μη εφαρμογή του.*

Section 6.
Appendix - Bibliography

1. Mozaffarian D, Benjamin EJ, Go AS, et al. Heart disease and stroke statistics--2015 update: a report from the American Heart Association. *Circulation.* Jan 27 2015;131(4):e29-322.

2. Murray CJ, Lopez AD. Measuring the global burden of disease. *N Engl J Med.* Aug 1 2013;369(5):448-457.

3. Excellence. NIfHaC. Lipid modification: cardiovascular risk assessment and the modification of blood lipids for the primary and secondary prevention of cardiovascular disease. *2008.*http://guidance.nice.org.uk/CG67*.*

4. Hippocrates. On airs, waters and places. In: Adams F, ed. The genuine works of Hippocrates. . *Baltimore: Wilkins and Wilkins,.* 1939.

5. Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? *BMJ.* 2009;338:b375.

6. Dorresteijn JA, Visseren FL, Ridker PM, et al. Estimating treatment effects for individual patients based on the results of randomised clinical trials. *BMJ.* 2011;343:d5888.

7. Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. *JAMA.* Sep 12 2007;298(10):1209-1212.

8. Hayward RA, Kent DM, Vijan S, Hofer TP. Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. *BMC Med Res Methodol.* 2006;6:18.

9. Kattan MW, Vickers AJ. Incorporating predictions of individual patient risk in clinical trials. *Urol Oncol.* Jul-Aug 2004;22(4):348-352.

10. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med.* Jan 6 2015;162(1):W1-73.

11. Kannel WB, McGee D, Gordon T. A general cardiovascular risk profile: the Framingham Study. *Am J Cardiol.* Jul 1976;38(1):46-51.

12. Pisters R, Lane DA, Nieuwlaat R, de Vos CB, Crijns HJ, Lip GY. A novel user-friendly score (HAS-BLED) to assess 1-year risk of major bleeding in patients with atrial fibrillation: the Euro Heart Survey. *Chest.* Nov 2010;138(5):1093-1100.

13. Galea MH, Blamey RW, Elston CE, Ellis IO. The Nottingham Prognostic Index in primary breast cancer. *Breast Cancer Res Treat.* 1992;22(3):207-219.

14. Gartman EJ, Casserly BP, Martin D, Ward NS. Using serial severity scores to predict death in ICU patients: a validation study and review of the literature. *Curr Opin Crit Care.* Dec 2009;15(6):578-582.

15. Le Gall JR, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *JAMA.* Dec 22-29 1993;270(24):2957-2963.

16. Knaus WA, Wagner DP, Draper EA, et al. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest.* Dec 1991;100(6):1619-1636.

17. Fisher B, Costantino JP, Wickerham DL, et al. Tamoxifen for prevention of breast cancer: report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study. *J Natl Cancer Inst.* Sep 16 1998;90(18):1371-1388.

18. Winzer KJ, Sauer R, Sauerbrei W, et al. Radiation therapy after breast-conserving surgery; first results of a randomised clinical trial in patients with low risk of recurrence. *Eur J Cancer.* May 2004;40(7):998-1005.

19. The CRIB (clinical risk index for babies) score: a tool for assessing initial neonatal risk and comparing performance of neonatal intensive care units. The International Neonatal Network. *Lancet.* Jul 24 1993;342(8865):193-198.

20. Wessler BS, Lai Yh L, Kramer W, et al. Clinical Prediction Models for Cardiovascular Disease: Tufts Predictive Analytics and Comparative Effectiveness Clinical Prediction Model Database. *Circ Cardiovasc Qual Outcomes.* Jul 7 2015.

21. Matheny M MM, Glasser A, Mercaldo N, Weaver RB, Jerome RN, Walden R, McKoy JN, Pritchett J, Tsai C. Systematic Review of Cardiovascular Disease Risk Assessment Tools. . *Evidence Synthesis No. 85. AHRQ Publication No. 11-05155-EF-1. Rockville, MD: Agency for Healthcare Research and Quality.* 2011.

22. Siontis GC, Tzoulaki I, Castaldi PJ, Ioannidis JP. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol.* Jan 2015;68(1):25-34.

23. WHO. Global status report on noncommunicable diseases 2010. . *Geneva: World Health Organization.* 2011.

24. Anderson KM, Odell PM, Wilson PW, Kannel WB. Cardiovascular disease risk profiles. *Am Heart J.* Jan 1991;121(1 Pt 2):293-298.

25. Ridker PM, Paynter NP, Rifai N, Gaziano JM, Cook NR. C-reactive protein and parental history improve global cardiovascular risk prediction: the Reynolds Risk Score for men. *Circulation.* Nov 25 2008;118(22):2243-2251, 2244p following 2251.

26. Ridker PM, Buring JE, Rifai N, Cook NR. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds Risk Score. *JAMA.* Feb 14 2007;297(6):611-619.

27. Collins GS, Altman DG. An independent and external validation of QRISK2 cardiovascular disease risk score: a prospective open cohort study. *BMJ.* 2010;340:c2442.

28. Collins GS, Altman DG. An independent external validation and evaluation of QRISK cardiovascular risk prediction: a prospective open cohort study. *BMJ.* 2009;339:b2584.

29. Hippisley-Cox J, Coupland C, Vinogradova Y, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ.* Jun 28 2008;336(7659):1475-1482.

30. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Brindle P. Performance of the QRISK cardiovascular risk prediction algorithm in an independent UK sample of patients from general practice: a validation study. *Heart.* Jan 2008;94(1):34-39.

31. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *BMJ.* Jul 21 2007;335(7611):136.

32. DeFrances CJ, Lucas CA, Buie VC, Golosinskiy A. 2006 National Hospital Discharge Survey. *Natl Health Stat Report.* Jul 30 2008(5):1-20.

33. Patel MR, Peterson ED, Dai D, et al. Low diagnostic yield of elective coronary angiography. *N Engl J Med.* Mar 11 2010;362(10):886-895.

34. Min JK, Gilmore A, Budoff MJ, Berman DS, O'Day K. Cost-effectiveness of coronary CT angiography versus myocardial perfusion SPECT for evaluation of patients with chest pain and no known coronary artery disease. *Radiology.* Mar 2010;254(3):801-808.

35. Ladapo JA, Jaffer FA, Hoffmann U, et al. Clinical outcomes and cost-effectiveness of coronary computed tomography angiography in the evaluation of patients with chest pain. *J Am Coll Cardiol.* Dec 15 2009;54(25):2409-2422.

36. Genders TS, Meijboom WB, Meijs MF, et al. CT coronary angiography in patients suspected of having coronary artery disease: decision making from various perspectives in the face of uncertainty. *Radiology.* Dec 2009;253(3):734-744.

37. Brotman DJ, Walker E, Lauer MS, O'Brien RG. In search of fewer independent risk factors. *Arch Intern Med.* Jan 24 2005;165(2):138-145.

38. Jarman B, Pieter D, van der Veen AA, et al. The hospital standardised mortality ratio: a powerful tool for Dutch hospitals to assess their quality of care? *Qual Saf Health Care.* Feb 2010;19(1):9-13.

39. Riley RD, Heney D, Jones DR, et al. A systematic review of molecular and biological tumor markers in neuroblastoma. *Clin Cancer Res.* Jan 1 2004;10(1 Pt 1):4-12.

40. Linton SJ, Hallden K. Can we screen for problematic back pain? A screening questionnaire for predicting outcome in acute and subacute back pain. *Clin J Pain.* Sep 1998;14(3):209-215.

41. Hingorani AD, Windt DA, Riley RD, et al. Prognosis research strategy (PROGRESS) 4: stratified medicine research. *BMJ.* 2013;346:e5793.

42. Roozenbeek B, Maas AI, Lingsma HF, et al. Baseline characteristics and statistical power in randomized controlled trials: selection, prognostic targeting, or covariate adjustment? *Crit Care Med.* Oct 2009;37(10):2683-2690.

43. Hernandez AV, Steyerberg EW, Habbema JD. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. *J Clin Epidemiol.* May 2004;57(5):454-460.

44. Lingsma HF, Roozenbeek B, Li B, et al. Large between-center differences in outcome after moderate and severe traumatic brain injury in the international mission on prognosis and clinical trial design in traumatic brain injury (IMPACT) study. *Neurosurgery.* Mar 2011;68(3):601-607; discussion 607-608.

45. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. . *Springer.* 2009.

46. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Ann Intern Med.* Feb 7 2006;144(3):201-209.

47. McShane LM, Altman DG, Sauerbrei W, et al. Reporting recommendations for tumor marker prognostic studies. *J Clin Oncol.* Dec 20 2005;23(36):9067-9072.

48. Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JD. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat Med.* Aug 30 2004;23(16):2567-2586.

49. Concato J. Challenges in prognostic analysis. *Cancer.* Apr 15 2001;91(8 Suppl):1607-1614.

50. Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med.* Feb 29 2000;19(4):453-473.

51. Randolph AG, Guyatt GH, Calvin JE, Doig G, Richardson WS. Understanding articles describing clinical prediction tools. Evidence Based Medicine in Critical Care Group. *Crit Care Med.* Sep 1998;26(9):1603-1612.

52. Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. *JAMA.* Feb 12 1997;277(6):488-494.

53. Braitman LE, Davidoff F. Predicting clinical states in individual patients. *Ann Intern Med.* Sep 1 1996;125(5):406-412.

54. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med.* Feb 28 1996;15(4):361-387.

55. Laupacis A, Wells G, Richardson WS, Tugwell P. Users' guides to the medical literature. V. How to use an article about prognosis. Evidence-Based Medicine Working Group. *JAMA.* Jul 20 1994;272(3):234-237.

56. Royston P, Moons KG, Altman DG, Vergouwe Y. Prognosis and prognostic research: Developing a prognostic model. *BMJ.* 2009;338:b604.

57. Steyerberg EW, Moons KG, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med.* 2013;10(2):e1001381.

58. Riley RD, Abrams KR, Sutton AJ, et al. Reporting of prognostic markers: current problems and development of guidelines for evidence-based practice in the future. *Br J Cancer.* Apr 22 2003;88(8):1191-1198.

59. Damen JA. Prognostic models for cardiovascular disease risk in the general population: a systematic review. *Cochrane Colloquim* 2015.

60. Canet J, Gallart L, Gomar C, et al. Prediction of postoperative pulmonary complications in a population-based surgical cohort. *Anesthesiology.* Dec 2010;113(6):1338-1350.

61. Nashef SA, Roques F, Sharples LD, et al. EuroSCORE II. *Eur J Cardiothorac Surg.* Apr 2012;41(4):734-744; discussion 744-735.

62. Schulze MB, Hoffmann K, Boeing H, et al. An accurate risk score based on anthropometric, dietary, and lifestyle factors to predict the development of type 2 diabetes. *Diabetes Care.* Mar 2007;30(3):510-515.

63. Hippisley-Cox J, Coupland C, Robson J, Sheikh A, Brindle P. Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore. *BMJ.* 2009;338:b880.

64. D'Agostino RB, Sr., Vasan RS, Pencina MJ, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation.* Feb 12 2008;117(6):743-753.

65. North RA, McCowan LM, Dekker GA, et al. Clinical risk prediction for pre-eclampsia in nulliparous women: development of model in international prospective cohort. *BMJ.* 2011;342:d1875.

66. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *J Clin Epidemiol.* Dec 1995;48(12):1503-1510.

67. Concato J, Peduzzi P, Holford TR, Feinstein AR. Importance of events per independent variable in proportional hazards analysis. I. Background, goals, and general strategy. *J Clin Epidemiol.* Dec 1995;48(12):1495-1501.

68. Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. *Am J Epidemiol.* Mar 15 2007;165(6):710-718.

69. Boriani G, Botto GL, Padeletti L, et al. Improving stroke risk stratification using the CHADS2 and CHA2DS2-VASc risk scores in patients with paroxysmal atrial fibrillation by continuous arrhythmia burden monitoring. *Stroke.* Jun 2011;42(6):1768-1770.

70. Eagle KA, Lim MJ, Dabbous OH, et al. A validated prediction model for all forms of acute coronary syndrome: estimating the risk of 6-month postdischarge death in an international registry. *JAMA.* Jun 9 2004;291(22):2727-2733.

71. Jarman B, Gault S, Alves B, et al. Explaining differences in English hospital death rates using routinely collected data. *BMJ.* Jun 5 1999;318(7197):1515-1520.

72. Haybittle JL, Blamey RW, Elston CW, et al. A prognostic index in primary breast cancer. *Br J Cancer.* Mar 1982;45(3):361-366.

73. Mackway-Jones K, Ryan B. Triage labels: choosing the national standard. *BMJ.* Sep 2 1989;299(6699):620.

74. Moons KG, Kengne AP, Woodward M, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart.* May 2012;98(9):683-690.

75. Moons KG, van Klei W, Kalkman CJ. Preoperative risk factors of intraoperative hypothermia in major surgery under general anesthesia. *Anesth Analg.* Jun 2003;96(6):1843-1844.

76. Biesheuvel CJ, Vergouwe Y, Oudega R, Hoes AW, Grobbee DE, Moons KG. Advantages of the nested case-control design in diagnostic research. *BMC Med Res Methodol.* 2008;8:48.

77. Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. *Br J Cancer.* Jun 1994;69(6):979-985.

78. Harrell FE, Jr. Regression modeling strategies with applications to linear models, logistic regression, and survival analysis. *New York: Springer.* 2001.

79. Steyerberg EW, Eijkemans MJ, Harrell FE, Jr., Habbema JD. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. *Med Decis Making.* Jan-Feb 2001;21(1):45-56.

80. Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *J Clin Epidemiol.* Aug 1996;49(8):907-916.

81. Little R.J.A. RDB. Statistical analysis with missing data. 2nd ed. . *New York: John Wiley.* 2002.

82. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol.* Oct 2006;59(10):1087-1091.

83. Altman DG, Royston P. The cost of dichotomising continuous variables. *BMJ.* May 6 2006;332(7549):1080.

84. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med.* Jan 15 2006;25(1):127-141.

85. Blettner M, Sauerbrei W. Influence of model-building strategies on the results of a case-control study. *Stat Med.* Jul 30 1993;12(14):1325-1338.

86. Sauerbrei W. The use of resampling methods to simplify regression models in medical statistics. . *Appl Stat.* 1999;48:313-329.

87. Sauerbrei W, Royston P, Look M. A new proposal for multivariable modelling of time-varying effects in survival data based on fractional polynomial time-transformation. *Biom J.* Jun 2007;49(3):453-473.

88. Royston PSW. Multivariable model-building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. . *Chichester: John Wiley.* 2008.

89. Rosenberg PS, Katki H, Swanson CA, Brown LM, Wacholder S, Hoover RN. Quantifying epidemiologic risk factors using non-parametric regression: model selection remains the greatest challenge. *Stat Med.* Nov 15 2003;22(21):3369-3381.

90. Boucher KM, Slattery ML, Berry TD, Quesenberry C, Anderson K. Statistical methods in epidemiology: a comparison of statistical methods to analyze dose-response and trend analysis in epidemiologic studies. *J Clin Epidemiol.* Dec 1998;51(12):1223-1233.

91. Hosmer DW, Lemeshow, S. Applied logistic regression. *2nd ed. New York: Wiley.* 2000.

92. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology.* Apr 1982;143(1):29-36.

93. Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ.* 2009;338:b606.

94. A D. Explained varianvce in logistic regression. A Monte Carlo study of proposed measures. *Social Methods Research.* 2002;31:27-74.

95. Grobbee DE, Hoes AW. Confounding and indication for treatment in evaluation of drug treatment for hypertension. *BMJ.* Nov 1 1997;315(7116):1151-1154.

96. van Walraven C, Davis D, Forster AJ, Wells GA. Time-dependent bias was common in survival analyses published in leading clinical journals. *J Clin Epidemiol.* Jul 2004;57(7):672-682.

97. McGinn TG, Guyatt GH, Wyer PC, Naylor CD, Stiell IG, Richardson WS. Users' guides to the medical literature: XXII: how to use articles about clinical decision rules. Evidence-Based Medicine Working Group. *JAMA.* Jul 5 2000;284(1):79-84.

98. Toll DB, Janssen KJ, Vergouwe Y, Moons KG. Validation, updating and impact of clinical prediction rules: a review. *J Clin Epidemiol.* Nov 2008;61(11):1085-1094.

99. Hernandez AV, Vergouwe Y, Steyerberg EW. Reporting of predictive logistic models should be based on evidence-based guidelines. *Chest.* Nov 2003;124(5):2034-2035; author reply 2035.

100. Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. *J Clin Epidemiol.* May 2003;56(5):441-447.

101. Steyerberg EW, Eijkemans MJ, Harrell FE, Jr., Habbema JD. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Stat Med.* Apr 30 2000;19(8):1059-1079.

102. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med.* Mar 16 1999;130(6):515-524.

103. Janssen KJ, Moons KG, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol.* Jan 2008;61(1):76-86.

104. Mackillop WJ, Quirt CF. Measuring the accuracy of prognostic judgments in oncology. *J Clin Epidemiol.* Jan 1997;50(1):21-29.

105. Moons KG, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart.* May 2012;98(9):691-698.

106. Miller ME, Hui SL, Tierney WM. Validation techniques for logistic regression models. *Stat Med.* Aug 1991;10(8):1213-1226.

107. Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J Clin Epidemiol.* May 2005;58(5):475-483.

108. Ioannidis JP, Khoury MJ. Improving validation practices in "omics" research. *Science.* Dec 2 2011;334(6060):1230-1232.

109. Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. *BMJ.* 2009;338:b605.

110. Taylor JM, Ankerst DP, Andridge RR. Validation of biomarker-based risk prediction models. *Clin Cancer Res.* Oct 1 2008;14(19):5977-5983.

111. Steyerberg EW, Harrell FE, Jr., Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol.* Aug 2001;54(8):774-781.

112. Oudega R, Hoes AW, Moons KG. The Wells rule does not adequately rule out deep venous thrombosis in primary care patients. *Ann Intern Med.* Jul 19 2005;143(2):100-107.

113. Kengne AP, Patel A, Colagiuri S, et al. The Framingham and UK Prospective Diabetes Study (UKPDS) risk equations do not reliably estimate the probability of cardiovascular events in a large ethnically diverse sample of patients with diabetes: the Action in Diabetes and Vascular Disease: Preterax and Diamicron-MR Controlled Evaluation (ADVANCE) Study. *Diabetologia.* May 2010;53(5):821-831.

114. Harrison DA, Rowan KM. Outcome prediction in critical care: the ICNARC model. *Curr Opin Crit Care.* Oct 2008;14(5):506-512.

115. Siontis GC, Ioannidis JP. Response to letter: more rigorous, not less, external validation is needed. *J Clin Epidemiol.* Jan 31 2015.

116. Siontis GC, Tzoulaki I, Siontis KC, Ioannidis JP. Comparisons of established risk prediction models for cardiovascular disease: systematic review. *BMJ.* 2012;344:e3318.

117. Ivanov J, Borger MA, David TE, Cohen G, Walton N, Naylor CD. Predictive accuracy study: comparing a statistical model to clinicians' estimates of outcomes after coronary bypass surgery. *Ann Thorac Surg.* Jul 2000;70(1):162-168.

118. Loeb M, Walter SD, McGeer A, Simor AE, McArthur MA, Norman G. A comparison of model-building strategies for lower respiratory tract infection in long-term care. *J Clin Epidemiol.* Dec 1999;52(12):1239-1248.

119. Moons KG. Criteria for scientific evaluation of novel markers: a perspective. *Clin Chem.* Apr 2010;56(4):537-541.

120. Ivanov J, Tu JV, Naylor CD. Ready-made, recalibrated, or Remodeled? Issues in the use of risk indexes for assessing mortality after coronary artery bypass graft surgery. *Circulation.* Apr 27 1999;99(16):2098-2104.

121. Hlatky MA, Greenland P, Arnett DK, et al. Criteria for evaluation of novel markers of cardiovascular risk: a scientific statement from the American Heart Association. *Circulation.* May 5 2009;119(17):2408-2416.

122. Steyerberg EW, Mushkudiani N, Perel P, et al. Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. *PLoS Med.* Aug 5 2008;5(8):e165; discussion e165.

123. Vickers AJ, Cronin AM. Traditional statistical methods for evaluating prediction models are uninformative as to clinical value: towards a decision analytic framework. *Semin Oncol.* Feb 2010;37(1):31-38.

124. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology.* Jan 2010;21(1):128-138.

125. Pencina MJ, D'Agostino RB, Sr., D'Agostino RB, Jr., Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med.* Jan 30 2008;27(2):157-172; discussion 207-112.

126. Cook NR. Assessing the Incremental Role of Novel and Emerging Risk Factors. *Curr Cardiovasc Risk Rep.* Mar 1 2010;4(2):112-119.

127. Vickers AJ, Cronin AM, Begg CB. One statistical test is sufficient for assessing new predictive markers. *BMC Med Res Methodol.* 2011;11:13.

128. Pencina MJ, D'Agostino RB, Sr., Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med.* Jan 15 2011;30(1):11-21.

129. Melander O, Newton-Cheh C, Almgren P, et al. Novel and conventional biomarkers for prediction of incident cardiovascular events in the community. *JAMA.* Jul 1 2009;302(1):49-57.

130. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol.* May 1 2004;159(9):882-890.

131. Janssen KJ, Vergouwe Y, Kalkman CJ, Grobbee DE, Moons KG. A simple method to adjust clinical prediction models to local circumstances. *Can J Anaesth.* Mar 2009;56(3):194-201.

132. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *BMJ.* Apr 2 2005;330(7494):765.

133. Michie S, Johnston M. Changing clinical behaviour by making guidelines specific. *BMJ.* Feb 7 2004;328(7435):343-345.

134. Campbell MK, Piaggio G, Elbourne DR, Altman DG, Group C. Consort 2010 statement: extension to cluster randomised trials. *BMJ.* 2012;345:e5661.

135. Foy R, Penney GC, Grimshaw JM, et al. A randomised controlled trial of a tailored multifaceted strategy to promote implementation of a clinical guideline on induced abortion care. *BJOG.* Jul 2004;111(7):726-733.

136. Marrie TJ, Lau CY, Wheeler SL, Wong CJ, Vandervoort MK, Feagan BG. A controlled trial of a critical pathway for treatment of community-acquired pneumonia. CAPITAL Study Investigators. Community-Acquired Pneumonia Intervention Trial Assessing Levofloxacin. *JAMA.* Feb 9 2000;283(6):749-755.

137. Brown CA, Lilford RJ. The stepped wedge trial design: a systematic review. *BMC Med Res Methodol.* 2006;6:54.

138. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials.* Feb 2007;28(2):182-191.

139. Mdege ND, Man MS, Taylor Nee Brown CA, Torgerson DJ. Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. *J Clin Epidemiol.* Sep 2011;64(9):936-948.

140. Price HC, Griffin SJ, Holman RR. Impact of personalized cardiovascular disease risk estimates on physical activity-a randomized controlled trial. *Diabet Med.* Mar 2011;28(3):363-372.

141. Harrell FE, Jr., Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA.* May 14 1982;247(18):2543-2546.

142. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. *Stat Med.* Sep 15-30 1999;18(17-18):2529-2545.

143. D'Agostino RB, Sr., Grundy S, Sullivan LM, Wilson P, Group CHDRP. Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. *JAMA.* Jul 11 2001;286(2):180-187.

144. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. *Stat Med.* May 15 1997;16(9):965-980.

145. Ware JH. The limitations of risk factors as prognostic tools. *N Engl J Med.* Dec 21 2006;355(25):2615-2617.

146. Greenland P, O'Malley PG. When is a new prediction marker useful? A consideration of lipoprotein-associated phospholipase A2 and C-reactive protein for stroke risk. *Arch Intern Med.* Nov 28 2005;165(21):2454-2456.

147. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation.* Feb 20 2007;115(7):928-935.

148. Wang TJ, Gona P, Larson MG, et al. Multiple biomarkers for the prediction of first major cardiovascular events and death. *N Engl J Med.* Dec 21 2006;355(25):2631-2639.

149. Cook NR, Buring JE, Ridker PM. The effect of including C-reactive protein in cardiovascular risk prediction models for women. *Ann Intern Med.* Jul 4 2006;145(1):21-29.

150. Pencina MJ, D'Agostino RB, Sr., Demler OV. Novel metrics for evaluating improvement in discrimination: net reclassification and integrated discrimination improvement for normal variables and nested models. *Stat Med.* Jan 30 2012;31(2):101-113.

151. Brehaut JC, Stiell IG, Graham ID. Will a new clinical decision rule be widely used? The case of the Canadian C-spine rule. *Acad Emerg Med.* Apr 2006;13(4):413-420.

152. Brehaut JC, Stiell IG, Visentin L, Graham ID. Clinical decision rules "in the real world": how a widely disseminated rule is used in everyday practice. *Acad Emerg Med.* Oct 2005;12(10):948-956.

153. Sinuff T, Adhikari NK, Cook DJ, et al. Mortality predictions in the intensive care unit: comparing physicians with scoring systems. *Crit Care Med.* Mar 2006;34(3):878-885.

154. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation.* May 12 1998;97(18):1837-1847.

155. Kamath PS, Kim WR, Advanced Liver Disease Study G. The model for end-stage liver disease (MELD). *Hepatology.* Mar 2007;45(3):797-805.

156. Halpern EJ, Fischman D, Savage MP, Koka AR, DeCaro M, Levin DC. Decision analytic model for evaluation of suspected coronary disease with stress testing and coronary CT angiography. *Acad Radiol.* May 2010;17(5):577-586.

157. Siontis GC, Tzoulaki I, Ioannidis JP. Predicting death: an empirical evaluation of predictive tools for mortality. *Arch Intern Med.* Oct 24 2011;171(19):1721-1726.

158. Christakis NA, Lamont EB. Extent and determinants of error in doctors' prognoses in terminally ill patients: prospective cohort study. *BMJ.* Feb 19 2000;320(7233):469-472.

159. Chow E, Harth T, Hruby G, Finkelstein J, Wu J, Danjoux C. How accurate are physicians' clinical predictions of survival and the available prognostic tools in estimating survival times in terminally ill cancer patients? A systematic review. *Clin Oncol (R Coll Radiol).* 2001;13(3):209-218.

160. Glare P, Virik K, Jones M, et al. A systematic review of physicians' survival predictions in terminally ill cancer patients. *BMJ.* Jul 26 2003;327(7408):195-198.

161. Kruger S, Ewig S, Giersdorf S, et al. Cardiovascular and inflammatory biomarkers to predict short- and long-term survival in community-acquired pneumonia: Results from the German Competence Network, CAPNETZ. *Am J Respir Crit Care Med.* Dec 1 2010;182(11):1426-1434.

162. Timsit JF, Fosse JP, Troche G, et al. Calibration and discrimination by daily Logistic Organ Dysfunction scoring comparatively with daily Sequential Organ Failure Assessment scoring for predicting hospital mortality in critically ill patients. *Crit Care Med.* Sep 2002;30(9):2003-2013.

163. Fine MJ, Auble TE, Yealy DM, et al. A prediction rule to identify low-risk patients with community-acquired pneumonia. *N Engl J Med.* Jan 23 1997;336(4):243-250.

164. Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. *Clin Chem.* Jan 2008;54(1):17-23.

165. Zou KH, O'Malley AJ, Mauri L. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation.* Feb 6 2007;115(5):654-657.

166. Tzoulaki I, Liberopoulos G, Ioannidis JP. Assessment of claims of improved prediction beyond the Framingham risk score. *JAMA.* Dec 2 2009;302(21):2345-2352.

167. Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Validity of prognostic models: when is a model clinically useful? *Semin Urol Oncol.* May 2002;20(2):96-107.

168. Hosmer DW, Hjort NL. Goodness-of-fit processes for logistic regression: simulation results. *Stat Med.* Sep 30 2002;21(18):2723-2738.

169. Cook NR, Ridker PM. Advances in measuring the effect of individual predictors of cardiovascular risk: the role of reclassification measures. *Ann Intern Med.* Jun 2 2009;150(11):795-802.

170. Janes H, Pepe MS, Gu W. Assessing the value of risk predictions by using risk stratification tables. *Ann Intern Med.* Nov 18 2008;149(10):751-760.

171. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med.* Jun 15 2002;21(11):1539-1558.

172. Sigounas DE, Tatsioni A, Christodoulou DK, Tsianos EV, Ioannidis JP. New prognostic markers for outcome of acute pancreatitis: overview of reporting in 184 studies. *Pancreas.* May 2011;40(4):522-532.

173. McShane LM, Altman DG, Sauerbrei W, et al. Reporting recommendations for tumor marker prognostic studies (REMARK). *J Natl Cancer Inst.* Aug 17 2005;97(16):1180-1184.

174. Lee KP, Schotland M, Bacchetti P, Bero LA. Association of journal quality indicators with methodological quality of clinical research articles. *JAMA.* Jun 5 2002;287(21):2805-2808.

175. Mihaescu R, van Zitteren M, van Hoek M, et al. Improvement of risk prediction by genomic profiling: reclassification measures versus the area under the receiver operating characteristic curve. *Am J Epidemiol.* Aug 1 2010;172(3):353-361.

176. Diamond GA. What price perfection? Calibration and discrimination of clinical prediction models. *J Clin Epidemiol.* Jan 1992;45(1):85-89.

177. Mosca L, Banka CL, Benjamin EJ, et al. Evidence-based guidelines for cardiovascular disease prevention in women: 2007 update. *J Am Coll Cardiol.* Mar 20 2007;49(11):1230-1250.

178. National Cholesterol Education Program Expert Panel on Detection E, Treatment of High Blood Cholesterol in A. Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. *Circulation.* Dec 17 2002;106(25):3143-3421.

179. Pearson TA, Blair SN, Daniels SR, et al. AHA Guidelines for Primary Prevention of Cardiovascular Disease and Stroke: 2002 Update: Consensus Panel Guide to Comprehensive Risk Reduction for Adult Patients Without Coronary or Other Atherosclerotic Vascular Diseases. American Heart Association Science Advisory and Coordinating Committee. *Circulation.* Jul 16 2002;106(3):388-391.

180. Cooney MT, Dudina A, D'Agostino R, Graham IM. Cardiovascular risk-estimation systems in primary prevention: do they differ? Do they make a difference? Can we see the future? *Circulation.* Jul 20 2010;122(3):300-310.

181. Berger JS, Jordan CO, Lloyd-Jones D, Blumenthal RS. Screening for cardiovascular risk in asymptomatic patients. *J Am Coll Cardiol.* Mar 23 2010;55(12):1169-1177.

182. Expert Panel on Detection E, Treatment of High Blood Cholesterol in A. Executive Summary of The Third Report of The National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, And Treatment of High Blood Cholesterol In Adults (Adult Treatment Panel III). *JAMA.* May 16 2001;285(19):2486-2497.

183. Woodward M, Brindle P, Tunstall-Pedoe H, estimation Sgor. Adding social deprivation and family history to cardiovascular risk assessment: the ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC). *Heart.* Feb 2007;93(2):172-176.

184. Conroy RM, Pyorala K, Fitzgerald AP, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *Eur Heart J.* Jun 2003;24(11):987-1003.

185. Assmann G, Cullen P, Schulte H. Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the prospective cardiovascular Munster (PROCAM) study. *Circulation.* Jan 22 2002;105(3):310-315.

186. Prevention of cardiovascular disease: guidelines for assessment and management of cardiovascular risk. World Health Organization. . 2007.

187. Barroso LC, Muro EC, Herrera ND, Ochoa GF, Hueros JI, Buitrago F. Performance of the Framingham and SCORE cardiovascular risk prediction functions in a non-diabetic population of a Spanish health care centre: a validation study. *Scand J Prim Health Care.* Dec 2010;28(4):242-248.

188. de la Iglesia B, Potter JF, Poulter NR, Robins MM, Skinner J. Performance of the ASSIGN cardiovascular disease risk score on a UK cohort of patients from general practice. *Heart.* Mar 2011;97(6):491-499.

189. Pandya A, Weinstein MC, Gaziano TA. A comparative assessment of non-laboratory-based versus commonly used laboratory-based cardiovascular disease risk scores in the NHANES III population. *PLoS One.* 2011;6(5):e20416.

190. van der Heijden AA, Ortegon MM, Niessen LW, Nijpels G, Dekker JM. Prediction of coronary heart disease risk in a general, pre-diabetic, and diabetic population during 10 years of follow-up: accuracy of the Framingham, SCORE, and UKPDS risk functions: The Hoorn Study. *Diabetes Care.* Nov 2009;32(11):2094-2098.

191. Chen L, Tonkin AM, Moon L, et al. Recalibration and validation of the SCORE risk chart in the Australian population: the AusSCORE chart. *Eur J Cardiovasc Prev Rehabil.* Oct 2009;16(5):562-570.

192. Woodward M, Tunstall-Pedoe H, Rumley A, Lowe GD. Does fibrinogen add to prediction of cardiovascular disease? Results from the Scottish Heart Health Extended Cohort Study. *Br J Haematol.* Aug 2009;146(4):442-446.

193. Scheltens T, Verschuren WM, Boshuizen HC, et al. Estimation of cardiovascular risk: a comparison between the Framingham and the SCORE model in people under 60 years of age. *Eur J Cardiovasc Prev Rehabil.* Oct 2008;15(5):562-566.

194. Mainous AG, 3rd, Koopman RJ, Diaz VA, Everett CJ, Wilson PW, Tilley BC. A coronary heart disease risk score based on patient-reported information. *Am J Cardiol.* May 1 2007;99(9):1236-1241.

195. Stork S, Feelders RA, van den Beld AW, et al. Prediction of mortality risk in the elderly. *Am J Med.* Jun 2006;119(6):519-525.

196. Cooper JA, Miller GJ, Humphries SE. A comparison of the PROCAM and Framingham point-scoring systems for estimation of individual risk of coronary heart disease in the Second Northwick Park Heart Study. *Atherosclerosis.* Jul 2005;181(1):93-100.

197. Ferrario M, Chiodini P, Chambless LE, et al. Prediction of coronary events in a low incidence population. Assessing accuracy of the CUORE Cohort Study prediction equation. *Int J Epidemiol.* Apr 2005;34(2):413-421.

198. Dunder K, Lind L, Zethelius B, Berglund L, Lithell H. Evaluation of a scoring scheme, including proinsulin and the apolipoprotein B/apolipoprotein A1 ratio, for the risk of acute coronary events in middle-aged men: Uppsala Longitudinal Study of Adult Men (ULSAM). *Am Heart J.* Oct 2004;148(4):596-601.

199. Empana JP, Ducimetiere P, Arveiler D, et al. Are the Framingham and PROCAM coronary heart disease risk functions applicable to different European populations? The PRIME Study. *Eur Heart J.* Nov 2003;24(21):1903-1911.

200. Marcin JP, Romano PS. Size matters to a model's fit. *Crit Care Med.* Sep 2007;35(9):2212-2213.

201. Bertolini G, D'Amico R, Nardi D, Tinazzi A, Apolone G. One model, several results: the paradox of the Hosmer-Lemeshow goodness-of-fit test for the logistic regression model. *J Epidemiol Biostat.* 2000;5(4):251-253.

202. Tzoulaki I, Liberopoulos G, Ioannidis JP. Use of reclassification for assessment of improved prediction: an empirical evaluation. *Int J Epidemiol.* Aug 2011;40(4):1094-1105.

203. Ioannidis JP. Perfect study, poor evidence: interpretation of biases preceding study design. *Semin Hematol.* Jul 2008;45(3):160-166.

204. Chalmers I, Matthews R. What are the implications of optimism bias in clinical research? *Lancet.* Feb 11 2006;367(9509):449-450.

205. Liu J, Hong Y, D'Agostino RB, Sr., et al. Predictive value for the Chinese population of the Framingham CHD risk assessment tool compared with the Chinese Multi-Provincial Cohort Study. *JAMA.* Jun 2 2004;291(21):2591-2599.

206. Hurley LP, Dickinson LM, Estacio RO, Steiner JF, Havranek EP. Prediction of cardiovascular death in racial/ethnic minorities using Framingham risk factors. *Circ Cardiovasc Qual Outcomes.* Mar 2010;3(2):181-187.

207. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology.* Sep 1983;148(3):839-843.

208. Brindle P, Beswick A, Fahey T, Ebrahim S. Accuracy and impact of risk assessment in the primary prevention of cardiovascular disease: a systematic review. *Heart.* Dec 2006;92(12):1752-1759.

209. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making.* Nov-Dec 2006;26(6):565-574.

210. Ioannidis JP, Tzoulaki I. What makes a good predictor?: the evidence applied to coronary artery calcium score. *JAMA.* Apr 28 2010;303(16):1646-1647.

211. Perel P, Edwards P, Wentz R, Roberts I. Systematic review of prognostic models in traumatic brain injury. *BMC Med Inform Decis Mak.* 2006;6:38.

212. Noble D, Mathur R, Dent T, Meads C, Greenhalgh T. Risk models and scores for type 2 diabetes: systematic review. *BMJ.* 2011;343:d7163.

213. Lemeshow S, Teres D, Klar J, Avrunin JS, Gehlbach SH, Rapoport J. Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients. *JAMA.* Nov 24 1993;270(20):2478-2486.

214. Adrie C, Francais A, Alvarez-Gonzalez A, et al. Model for predicting short-term mortality of severe sepsis. *Crit Care.* 2009;13(3):R72.

215. Rifai N, Altman DG, Bossuyt PM. Reporting bias in diagnostic and prognostic studies: time for action. *Clin Chem.* Jul 2008;54(7):1101-1103.

216. Ioannidis JP, Tzoulaki I. Minimal and null predictive effects for the most popular blood biomarkers of cardiovascular disease. *Circ Res.* Mar 2 2012;110(5):658-662.

217. Ioannidis JP, Panagiotou OA. Comparison of effect sizes associated with biomarkers reported in highly cited individual articles and in subsequent meta-analyses. *JAMA.* Jun 1 2011;305(21):2200-2210.

218. Zhou K, Gao CF, Zhao YP, et al. Simpler score of routine laboratory tests predicts liver fibrosis in patients with chronic hepatitis B. *J Gastroenterol Hepatol.* Sep 2010;25(9):1569-1577.

219. Rollig C, Thiede C, Gramatzki M, et al. A novel prognostic model in elderly patients with acute myeloid leukemia: results of 909 patients entered into the prospective AML96 trial. *Blood.* Aug 12 2010;116(6):971-978.

220. Conti A, Vanni S, Taglia BD, et al. A new simple risk score in patients with acute chest pain without existing known coronary disease. *Am J Emerg Med.* Feb 2010;28(2):135-142.

221. Wishart GC, Azzato EM, Greenberg DC, et al. PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. *Breast Cancer Res.* 2010;12(1):R1.

222. Yuen MF, Tanaka Y, Fong DY, et al. Independent risk factors and predictive score for the development of hepatocellular carcinoma in chronic hepatitis B. *J Hepatol.* Jan 2009;50(1):80-88.

223. Yau T, Yao TJ, Chan P, Ng K, Fan ST, Poon RT. A new prognostic score system in patients with advanced hepatocellular carcinoma not amendable to locoregional therapy: implication for patient selection in systemic therapy trials. *Cancer.* Nov 15 2008;113(10):2742-2751.

224. Harrison DA, D'Amico G, Singer M. The Pancreatitis Outcome Prediction (POP) Score: a new prognostic index for patients with severe acute pancreatitis. *Crit Care Med.* Jul 2007;35(7):1703-1708.

225. Harrison DA, Parry GJ, Carpenter JR, Short A, Rowan K. A new risk prediction model for critical care: the Intensive Care National Audit & Research Centre (ICNARC) model. *Crit Care Med.* Apr 2007;35(4):1091-1098.

226. Colinet B, Jacot W, Bertrand D, et al. A new simplified comorbidity score as a prognostic factor in non-small-cell lung cancer patients: description and comparison with the Charlson's index. *Br J Cancer.* Nov 14 2005;93(10):1098-1105.

227. Sanchis J, Bodi V, Nunez J, et al. New risk score for patients with acute chest pain, non-ST-segment deviation, and normal troponin concentrations: a comparison with the TIMI risk score. *J Am Coll Cardiol.* Aug 2 2005;46(3):443-449.

228. Tateishi R, Yoshida H, Shiina S, et al. Proposal of a new prognostic model for hepatocellular carcinoma: an analysis of 403 patients. *Gut.* Mar 2005;54(3):419-425.

229. Donati A, Ruzzi M, Adrario E, et al. A new and feasible model for predicting operative risk. *Br J Anaesth.* Sep 2004;93(3):393-399.

230. Ho GT, Mowat C, Goddard CJ, et al. Predicting the outcome of severe ulcerative colitis: development of a novel risk score to aid early selection of patients for second-line medical therapy or surgery. *Aliment Pharmacol Ther.* May 15 2004;19(10):1079-1087.

231. Olsson T, Terent A, Lind L. Rapid Emergency Medicine score: a new prognostic tool for in-hospital mortality in nonsurgical emergency department patients. *J Intern Med.* May 2004;255(5):579-587.

232. Gallamini A, Stelitano C, Calvi R, et al. Peripheral T-cell lymphoma unspecified (PTCL-U): a new prognostic model from a retrospective multicentric clinical study. *Blood.* Apr 1 2004;103(7):2474-2479.

233. Villella M, Villella A, Santoro L, et al. Ergometric score systems after myocardial infarction: prognostic performance of the Duke Treadmill Score, Veterans Administration Medical Center Score, and of a novel score system, GISSI-2 Index, in a cohort of survivors of acute myocardial infarction. *Am Heart J.* Mar 2003;145(3):475-483.

234. Sedrine WB, Chevallier T, Zegels B, et al. Development and assessment of the Osteoporosis Index of Risk (OSIRIS) to facilitate selection of women for bone densitometry. *Gynecol Endocrinol.* Jun 2002;16(3):245-250.

235. Josting A, Franklin J, May M, et al. New prognostic score based on treatment outcome of patients with relapsed Hodgkin's lymphoma registered in the database of the German Hodgkin's lymphoma study group. *J Clin Oncol.* Jan 1 2002;20(1):221-230.

236. LeMaire SA, Miller CC, 3rd, Conklin LD, Schmittling ZC, Koksoy C, Coselli JS. A new predictive model for adverse outcomes after elective thoracoabdominal aortic aneurysm repair. *Ann Thorac Surg.* Apr 2001;71(4):1233-1238.

237. Wong DT, Cheng DC, Kustra R, et al. Risk factors of delayed extubation, prolonged length of stay in the intensive care unit, and mortality in patients undergoing coronary artery bypass graft with fast-track cardiac anesthesia: a new cardiac risk score. *Anesthesiology.* Oct 1999;91(4):936-944.

238. Maltoni M, Nanni O, Pirovano M, et al. Successful validation of the palliative prognostic score in terminally ill cancer patients. Italian Multicenter Study Group on Palliative Care. *J Pain Symptom Manage.* Apr 1999;17(4):240-247.

239. Hasford J, Pfirrmann M, Hehlmann R, et al. A new prognostic score for survival of patients with chronic myeloid leukemia treated with interferon alfa. Writing Committee for the Collaborative CML Prognostic Factors Project Group. *J Natl Cancer Inst.* Jun 3 1998;90(11):850-858.

240. Roberts AB, Lee AJ, James AG. Ultrasonic estimation of fetal weight: a new predictive model incorporating femur length for the low-birth-weight fetus. *J Clin Ultrasound.* Oct 1985;13(8):555-559.

241. Lenstrup C. Predictive value of a single unstressed antepartum cardiotocogram in apparently uncomplicated pregnancy. Introduction of a new cardiotocography score. *Acta Obstet Gynecol Scand.* 1982;61(2):177-182.

242. Conti A, Poggioni C, Viviani G, et al. Risk scores prognostic implementation in patients with chest pain and nondiagnostic electrocardiograms. *Am J Emerg Med.* Nov 2012;30(9):1719-1728.

243. Qu Y, Gao CF, Zhou K, Zhao YP, Xu MY, Lu LG. Serum N-glycomic markers in combination with panels improves the diagnosis of chronic hepatitis B. *Ann Hepatol.* Mar-Apr 2012;11(2):202-212.

244. Wishart GC, Bajdik CD, Azzato EM, et al. A population-based validation of the prognostic model PREDICT for early breast cancer. *Eur J Surg Oncol.* May 2011;37(5):411-417.

245. Lenstrup C. Significance of unstressed antepartum cardiotocography performed weekly in last trimester of apparently normal pregnancy. *Acta Obstet Gynecol Scand.* 1982;61(5):397-402.

246. Wong GL, Chan HL, Chan HY, et al. Accuracy of risk scores for patients with chronic hepatitis B receiving entecavir treatment. *Gastroenterology.* May 2013;144(5):933-944.

247. Djunic I, Suvajdzic-Vukovic N, Virijevic M, et al. Prognostic risk score for the survival of elderly patients with acute myeloid leukaemia comprising comorbidities. *Med Oncol.* Mar 2013;30(1):394.

248. Lin ZZ, Hsu C, Hu FC, et al. Factors impacting prognosis prediction in BCLC stage C and Child-Pugh class A hepatocellular carcinoma patients in prospective clinical trials of systemic therapy. *Oncologist.* 2012;17(7):970-977.

249. Haga Y, Ikejiri K, Wada Y, et al. A multicenter prospective study of surgical audit systems. *Ann Surg.* Jan 2011;253(1):194-201.

250. Juneja D, Gopal PB, Ravula M. Scoring systems in acute pancreatitis: which one to use in intensive care units? *J Crit Care.* Jun 2010;25(2):358 e359-358 e315.

251. Aceituno M, Garcia-Planella E, Heredia C, et al. Steroid-refractory ulcerative colitis: predictive factors of response to cyclosporine and validation in an independent cohort. *Inflamm Bowel Dis.* Mar 2008;14(3):347-352.

252. Kondo K, Chijiiwa K, Nagano M, et al. Comparison of seven prognostic staging systems in patients who undergo hepatectomy for hepatocellular carcinoma. *Hepatogastroenterology.* Jul-Aug 2007;54(77):1534-1538.

253. Smeltzer JP, Cashen AF, Zhang Q, et al. Prognostic significance of FDG-PET in relapsed or refractory classical Hodgkin lymphoma treated with standard salvage chemotherapy and autologous stem cell transplantation. *Biol Blood Marrow Transplant.* Nov 2011;17(11):1646-1652.

254. Yang DH, Kim WS, Kim SJ, et al. Prognostic factors and clinical outcomes of high-dose chemotherapy followed by autologous stem cell transplantation in patients with peripheral T cell lymphoma, unspecified: complete remission at transplantation and the prognostic index of peripheral T cell lymphoma are the major factors predictive of outcome. *Biol Blood Marrow Transplant.* Jan 2009;15(1):118-125.

255. Siegenthaler MP, Weigang E, Brehm K, et al. Endovascular treatment for thoracoabdominal aneurysms: outcomes and results. *Eur J Cardiothorac Surg.* Oct 2008;34(4):810-819.

256. Valeur N, Clemmensen P, Grande P, Saunamaki K, Investigators D-. Prognostic evaluation by clinical exercise test scores in patients treated with primary percutaneous coronary intervention or fibrinolysis for acute myocardial infarction (a Danish Trial in Acute Myocardial Infarction-2 Sub-Study). *Am J Cardiol.* Oct 1 2007;100(7):1074-1080.

257. Goodacre S, Turner J, Nicholl J. Prediction of mortality among emergency medical admissions. *Emerg Med J.* May 2006;23(5):372-375.

258. Cook RB, Collins D, Tucker J, Zioupos P. Comparison of questionnaire and quantitative ultrasound techniques as screening tools for DXA. *Osteoporos Int.* Dec 2005;16(12):1565-1575.

259. Yende S, Wunderink R. Validity of scoring systems to predict risk of prolonged mechanical ventilation after coronary artery bypass graft surgery. *Chest.* Jul 2002;122(1):239-244.

260. Simon NV, Levisky JS, Shearer DM, O'Lear MS, Flood JT. Influence of fetal growth patterns on sonographic estimation of fetal weight. *J Clin Ultrasound.* Jul-Aug 1987;15(6):376-383.

261. Lip GY, Frison L, Halperin JL, Lane DA. Comparative validation of a novel risk score for predicting bleeding risk in anticoagulated patients with atrial fibrillation: the HAS-BLED (Hypertension, Abnormal Renal/Liver Function, Stroke, Bleeding History or Predisposition, Labile INR, Elderly, Drugs/Alcohol Concomitantly) score. *J Am Coll Cardiol.* Jan 11 2011;57(2):173-180.

262. Jacot W, Colinet B, Bertrand D, et al. Quality of life and comorbidity score as prognostic determinants in non-small-cell lung cancer patients. *Ann Oncol.* Aug 2008;19(8):1458-1464.

263. Nolan JP, Laver SR, Welch CA, Harrison DA, Gupta V, Rowan K. Outcome following admission to UK intensive care units after cardiac arrest: a secondary analysis of the ICNARC Case Mix Programme Database. *Anaesthesia.* Dec 2007;62(12):1207-1216.

264. Tassinari D, Montanari L, Maltoni M, et al. The palliative prognostic score and survival in patients with advanced solid tumors receiving chemotherapy. *Support Care Cancer.* Apr 2008;16(4):359-370.

265. Sanchis J, Bodi V, Nunez J, et al. Usefulness of early exercise testing and clinical risk score for prognostic evaluation in chest pain units without preexisting evidence of myocardial ischemia. *Am J Cardiol.* Mar 1 2006;97(5):633-635.

266. Hasford J, Pfirrmann M, Shepherd P, et al. The impact of the combination of baseline risk group and cytogenetic response on the survival of patients with chronic myeloid leukemia treated with interferon alpha. *Haematologica.* Mar 2005;90(3):335-340.

267. Smith JG, Wieloch M, Koul S, et al. Triple antithrombotic therapy following an acute coronary syndrome: prevalence, outcomes and prognostic utility of the HAS-BLED score. *EuroIntervention.* Oct 2012;8(6):672-678.

268. Chen JC, Wang WJ, Wu VC, et al. The ICNARC model is predictive of hospital mortality in critically ill patients supported by acute dialysis. *Clin Nephrol.* May 2012;77(5):392-399.

269. Kuo YW, Jerng JS, Shih JY, Chen KY, Yu CJ, Yang PC. The prognostic value of the simplified comorbidity score in the treatment of small cell lung carcinoma. *J Thorac Oncol.* Feb 2011;6(2):378-383.

270. Manini AF, Dannemann N, Brown DF, et al. Limitations of risk score models in patients with acute chest pain. *Am J Emerg Med.* Jan 2009;27(1):43-48.

271. Glare P, Virik K. Independent prospective validation of the PaP score in terminally ill patients referred to a hospital-based palliative medicine consultation service. *J Pain Symptom Manage.* Nov 2001;22(5):891-898.

272. Bonifazi F, De Vivo A, Rosti G, et al. Testing Sokal's and the new prognostic score for chronic myeloid leukaemia treated with alpha-interferon. Italian Cooperative Study Group on Chronic Myeloid Leukaemia. *Br J Haematol.* Nov 2000;111(2):587-595.

273. Bleeker SE, Moll HA, Steyerberg EW, et al. External validation is necessary in prediction research: a clinical example. *J Clin Epidemiol.* Sep 2003;56(9):826-832.

274. Bouwmeester W, Zuithoff NP, Mallett S, et al. Reporting and methods in clinical prediction research: a systematic review. *PLoS Med.* 2012;9(5):1-12.

275. Yourman LC, Lee SJ, Schonberg MA, Widera EW, Smith AK. Prognostic indices for older adults: a systematic review. *JAMA.* Jan 11 2012;307(2):182-192.

276. Abbasi A, Peelen LM, Corpeleijn E, et al. Prediction models for risk of developing type 2 diabetes: systematic literature search and independent external validation study. *BMJ.* 2012;345:e5900.

277. van Dieren S, Beulens JW, Kengne AP, et al. Prediction models for the risk of cardiovascular disease in patients with type 2 diabetes: a systematic review. *Heart.* Mar 2012;98(5):360-369.

278. Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Med.* 2011;9:103.

279. Wlodzimirow KA, Eslami S, Chamuleau RA, Nieuwoudt M, Abu-Hanna A. Prediction of poor outcome in patients with acute liver failure-systematic review of prediction models. *PLoS One.* 2012;7(12):e50952.

280. Spelt L, Andersson B, Nilsson J, Andersson R. Prognostic models for outcome following liver resection for colorectal cancer metastases: A systematic review. *Eur J Surg Oncol.* Jan 2012;38(1):16-24.

281. Minne L, Ludikhuize J, de Jonge E, de Rooij S, Abu-Hanna A. Prognostic models for predicting mortality in elderly ICU patients: a systematic review. *Intensive Care Med.* Aug 2011;37(8):1258-1268.

282. Leushuis E, van der Steeg JW, Steures P, et al. Prediction models in reproductive medicine: a critical appraisal. *Hum Reprod Update.* Sep-Oct 2009;15(5):537-552.

283. Minne L, Abu-Hanna A, de Jonge E. Evaluation of SOFA-based models for predicting mortality in the ICU: A systematic review. *Crit Care.* 2008;12(6):R161.

284. Mallett S, Royston P, Waters R, Dutton S, Altman DG. Reporting performance of prognostic models in cancer: a review. *BMC Med.* 2010;8:21.

285. Mallett S, Royston P, Dutton S, Waters R, Altman DG. Reporting methods in studies developing prognostic models in cancer: a review. *BMC Med.* 2010;8:20.

286. Rathore SS, Weinfurt KP, Foody JM, Krumholz HM. Performance of the Thrombolysis in Myocardial Infarction (TIMI) ST-elevation myocardial infarction risk score in a national cohort of elderly patients. *Am Heart J.* Sep 2005;150(3):402-410.

287. Charlson ME, Ales KL, Simon R, MacKenzie CR. Why predictive indexes perform less well in validation studies. Is it magic or methods? *Arch Intern Med.* Dec 1987;147(12):2155-2161.

288. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J.* Aug 1 2014;35(29):1925-1931.

289. Geersing GJ, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, Moons KG. Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews. *PLoS One.* 2012;7(2):e32844.

290. Yuen MF, Tanaka Y, Shinkai N, et al. Risk for hepatocellular carcinoma with respect to hepatitis B virus genotypes B/C, specific mutations of enhancer II/core promoter/precore regions and HBV DNA levels. *Gut.* Jan 2008;57(1):98-102.

291. Harrison DA, D'Amico G, Singer M. Case mix, outcome, and activity for admissions to UK critical care units with severe acute pancreatitis: a secondary analysis of the ICNARC Case Mix Programme Database. *Crit Care.* 2007;11 Suppl 1:S1.

292. Pirovano M, Maltoni M, Nanni O, et al. A new palliative prognostic score: a first step for the staging of terminally ill cancer patients. Italian Multicenter and Study Group on Palliative Care. *J Pain Symptom Manage.* Apr 1999;17(4):231-239.

293. Windeler J. Prognosis - what does the clinician associate with this notion? *Stat Med.* Feb 29 2000;19(4):425-430.

294. Coggon DI, Martyn CN. Time and chance: the stochastic nature of disease causation. *Lancet.* Apr 16-22 2005;365(9468):1434-1437.