Editorials represent the opinions
of the authors and *JAMA* and
not those of the American Medical Association.

What Makes a Good Predictor?

The Evidence Applied to Coronary Artery Calcium Score

John P. A. Ioannidis, MD
Ioanna Tzoulaki, PhD

EACH YEAR, RESEARCHERS IDENTIFY THOUSANDS OF potential new "tools" for predicting patients' medical futures. There is heightened interest for discovering, validating, and incorporating into clinical practice predictors that improve treatment choices and outcomes thereof.[1,2] Thousands of articles report on potential predictors. A search of PubMed clinical queries under *prognosis* (specific strategy) yields 165 746 articles for cancer, 72 354 for cardiovascular disease, and even 3749 for rheumatoid arthritis. These run the gamut, including genetic tests, biomarkers, and an increasing variety of imaging modes, lengthening the list of candidate predictors.[3] However, very few of these proposed predictors eventually change practice. Why? What makes a good predictor?

A good predictor is one that has a favorable risk-benefit ratio, reasonable cost, acceptability, and convenience. As for any intervention in health care, proper evidence ideally requires randomized trials demonstrating that using the predictor improves decision making and subsequent clinical outcomes without inordinate adverse events. It also requires formal cost-effectiveness analyses, integrating benefits, risks, and cost.[2] However, hardly any of the predictors in the literature or even those routinely adopted in clinical practice have had their effectiveness proven in randomized trials. Only a few examples of such trials exist; eg, trials evaluating the benefits of screening for abdominal aneurysms or measuring brain-type natriuretic peptide in patients with dyspnea.[2] Conversely, a comprehensive randomized trial agenda trying to evaluate every proposed predictor in each proposed disease application and population would require millions of trials, which is unrealistic. Which candidate predictors should be evaluated by randomized trials and how should they be chosen for best results?

A commonsense checklist might be to, first, preferably test predictors for diseases with major morbidity. Second, some effective treatment should be available. Third, the treatment should not be equally effective (or equally risky) for all persons. Fourth, consideration of the predictor should allow more accurate classification of individuals into categories in which treatment is or is not indicated. Fifth, the incremental prediction should be accomplished beyond what can be achieved with information already available. Sixth, there should be consensus about and standardization of established, routine predictors. Seventh, the predictor should be unambiguously defined and measured.

Most published research on predictors is irrelevant or tangential to this checklist. Almost all articles report statistically significant results,[4] but this means little. Many investigators deal with whether a predictor in isolation has any ability to predict something. This, however, does not consider that many clinical facts and routine laboratory predictors may already inform prognosis. Thus, it is often not clear whether the new test adds incremental prognostic information beyond known factors. Much of the literature is chaotic, and data dredging and selective reporting[5] abound. Strong studies with clear design, purpose, and knowledge are clearly needed.

In this issue of *JAMA*, Polonsky et al[6] present such a well-designed study addressing coronary artery calcium score (CACS) as a predictor of coronary heart disease (CHD). Is this predictor good enough? In regard to the aforementioned checklist, first, CHD indeed carries major morbidity. Second, effective lipid-lowering treatments are available for preventive purposes. Third, the absolute effectiveness of the treatments (absolute risk reduction) varies at different categories of baseline risk. Patients at greater than 20% risk of CHD over 10 years should be treated, those with less than 10% should not, and those with 10% to 20% are in the gray zone of intermediate risk.[7] Fourth, Polonsky et al suggest that CACS does allow for a better classification of patients into categories in which, seemingly, treatment is or not indicated. Fifth, this is accomplished in addition to the information available from established routine predictors,

Author Affiliations: Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, and Biomedical Research Institute, Foundation for Research and Technology–Hellas, Ioannina, Greece (Dr Ioannidis); Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Department of Medicine, Tufts University School of Medicine, and Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts (Dr Ioannidis); and Department of Epidemiology and Biostatistics, Imperial College of Medicine, London, England (Dr Tzoulaki).

Corresponding Author: John P. A. Ioannidis, MD, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina 45110, Greece (jioannid@cc.uoi.gr).

See also p 1610.

including age, sex, smoking, diabetes, systolic blood pressure, use of antihypertensive agents, and total and high-density lipoprotein cholesterol levels. Sixth, there is consensus that these are indeed established routine predictors. Seventh, CACS can be unambiguously defined and measured.

This sounds perfect, but there are still caveats.[8] First, the risk thresholds in the article by Polonsky et al[6] are calculated for 5 years, whereas the standard literature for treatment refers to 10-year risks.[7] Polonsky et al assume that less than 3%, 3% to 10%, and greater than 10% at 5 years are equivalent to less than 10%, 10% to 20%, and greater than 20% at 10 years. However, these categorizations are quite arbitrary, and less than 5%, 5% to 10%, and greater than 10% would have been closer to the standard. Second, the routine predictors overlap with those included in the well-standardized and validated Framingham Risk Score (FRS), but Polonsky et al did not calculate the FRS based on the validated coefficients of each variable. The FRS was developed based on 10-year follow-up of white populations. Polonsky et al have a mixed-ancestry population and attempt to adjust for ancestry by adding it as a covariate, which is a reasonable approach, but this is no longer a standardized FRS. Third, the selected CHD outcome is standardized and widely accepted, but it is a composite outcome that includes as disparate events as death and slight angina.[9] Fourth, CACS can be measured accurately by experienced computed tomography radiologists, and interobserver and intraobserver agreement was high in the study by Polonsky et al, but measurements may be less accurate when widely used in the community.

A closer look is warranted in particular at the statistical methods that address what a new predictor achieves in addition to routine predictors. These methods address discrimination (multivariate-adjusted risk ratios, changes in area under the curve or C index, integrated discrimination index), reclassification (net reclassification improvement, risk stratification capacity), and calibration. Polonsky et al examined these aspects in an exemplary fashion. Discrimination tells whether the predictor can help further differentiate who will and who will not have an event. Adjusting for other predictors, the hazard ratio for an event increases 1.48-fold per 1-point increase in $\ln\text{CACS} + 1$. The area under the curve, a measure that captures the overall trade-off between sensitivity and specificity, improves by 0.05 (from 0.76 to 0.81). The integrated discrimination index sums the improvements in true-positive rates minus the worsening in false-positive rates and has a favorable value of 0.026. Reclassification evaluates whether patients are reclassified in different categories of risk and whether these changes are correct or wrong. The net reclassification improvement is a favorable 0.25. Finally, calibration examines goodness of fit; ie, whether the estimated risk is appropriate or systematically off target. Calibration testing showed no significant lack of fit (Hosmer-Lemeshow $P = .24$).

So, do these statistical numbers add up to prove that CACS is a successful prognostic tool? First, additional studies are needed to ensure that these favorable results are observed in different cohorts using standardized definitions and analyses. Then, ideally, these data should be synthesized with meta-analysis. This rarely happens, probably because standardization of the predictive literature is so poor. A few brave or unwise meta-analysts dare to perform these reviews and usually fall into traps of selective reporting.[5]

When the checklist is satisfied, the next step is to test whether the predictor should be used routinely. Polonsky et al cautiously acknowledge that they analyzed a prospective cohort, not a randomized intervention trial. Thus, the authors have not yet demonstrated that the added accuracy in risk stratification can actually aid clinicians in better treating patients or improving their clinical outcomes. Therefore, their findings, no matter how promising, do not suffice to recommend this marker for widespread routine use. Moreover, cost and harms may be major issues. Computed tomography costs $200 to $600[10] and routine implementation at the population level can be very expensive. The lifetime excess cancer risk due to radiation exposure from a single examination at age 40 years is 9 cancers per 100 000 men and 28 cancers per 100 000 women. This risk should be taken into account in formal risk-benefit analyses.[11]

All of these aspects require careful weighting. The evidence to date suggests that while CACS is a promising tool, the verdict is not in yet as to whether it is ready for routine use, and much more is still left to do.

Financial Disclosures: None reported.

REFERENCES

1. Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? *BMJ*. 2009;338:b375.
2. Hlatky MA, Greenland P, Arnett DK, et al; American Heart Association Expert Panel on Subclinical Atherosclerotic Diseases and Emerging Risk Factors and the Stroke Council. Criteria for evaluation of novel markers of cardiovascular risk: a scientific statement from the American Heart Association. *Circulation*. 2009;119(17):2408-2416.
3. Ioannidis JP. Limits to forecasting in personalized medicine: an overview. *Int J Forecast*. 2009;25:773-783.
4. Kyzas PA, Denaxa-Kyza D, Ioannidis JP. Almost all articles on cancer prognostic markers report statistically significant results. *Eur J Cancer*. 2007;43(17):2559-2579.
5. Kyzas PA, Loizou KT, Ioannidis JP. Selective reporting biases in cancer prognostic factor studies. *J Natl Cancer Inst*. 2005;97(14):1043-1055.
6. Polonsky TS, McClelland RL, Jorgensen NW, et al. Coronary artery calcium score and risk classification for coronary heart disease prediction. *JAMA*. 2010;303(16):1610-1616.
7. Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. 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*. 2001;285(19):2486-2497.
8. Tzoulaki I, Liberopoulos G, Ioannidis JP. Assessment of claims of improved prediction beyond the Framingham risk score. *JAMA*. 2009;302(21):2345-2352.
9. Ferreira-González I, Busse JW, Heels-Ansdell D, et al. Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. *BMJ*. 2007;334(7597):786.
10. Scan Directory: calcium scan. http://www.scandirectory.com/content/calcium_scan.asp. Accessed April 7, 2010.
11. Kim KP, Einstein AJ, Berrington de González A. Coronary artery calcification screening: estimated radiation dose and cancer risk. *Arch Intern Med*. 2009;169(13):1188-1194.