CLINICAL MODELING

Thoroughly Modern Risk Prediction?

Michael J. Pencina[1]* and Ralph B. D'Agostino Sr.[2]

Researchers use a data-driven technique called statistical learning to fashion risk prediction algorithms for clinical situations with low event rates.

As one aid for treatment decisions, clinicians analyze a patient's clinical parameters using risk assessment algorithms to predict the individual's likelihood of developing particular diseases in the future. For example, risk prediction models for cardiovascular disease (CVD) use a person's age, sex, systolic blood pressure, total and high-density lipoprotein cholesterol, antihypertensive treatment, and smoking and diabetes statuses to calculate their 10-year risk of CVD; treatment recommendations are then based on this 10-year risk. Prediction models are related to diagnostic models, with one crucial distinction: In the case of the latter, disease onset has already occurred, while in the case of the former, disease may occur in the future. Thus, prediction algorithms are developed on the basis of prospective cohort studies, in which large numbers of disease-free participants submit to extensive baseline measurements and then are followed for a prespecified amount of time for the development of a disease (an "event"). Most often, statistical techniques called regressions are used to decide which baseline risk factors to include in the final risk prediction model and how to relate these factors to the risk of specific events. These regressions assume a certain, usually simple, model that links risk factors with the onset of disease.

However, the computer revolution of the last few decades has opened new possibilities for the construction of risk algorithms that do not require reliance on a model in the traditional sense. In this issue of *Science Translational Medicine*, Chia *et al.* (*1*) present one such new approach to constructing risk prediction algorithms that is based on a data-driven technique called statistical learning and is intended for situations in which the event rate is low. Here, we put their proposal in the context of risk prediction in general and attempt to address the following question: What types of questions must be answered to determine whether the new technique might replace the current approaches?

PREDICTING THE FUTURE BETTER

Risk prediction algorithms have been successfully developed in all major fields of modern medicine, including CVD, cancer, and their comorbidities. The pioneer in this area is the Framingham Heart Study (http://www.framinghamheartstudy.org), which originated in the late 1940s and recently recruited its third generation of participants, making it one of the longest running epidemiological studies in the world (*2*). The first risk prediction models for coronary heart disease (CHD) were developed in the 1960s on the basis of the Framingham data and employed a technique called linear discriminant analysis, which was replaced by logistic regression (*3*). Over the last two decades, the most popular functions for risk prediction model building have relied on survival regression techniques, which assume that the outcome and predictors are related through a function (that is, model) for which parameters can be estimated given the data. In 2001, the National Cholesterol Education Program's Third Adult Treatment Panel (NCEP ATP III) used, for their lipid treatment guidelines, the Framingham prediction model for 10-year risk of "hard" CHD (*4*). Individuals whose 10-year risk of CHD exceeded 20% were considered candidates for pharmacotherapy with cholesterol lowering agents such as statins. The newest Framingham function was developed in 2008 by D'Agostino *et al.* (*5*) and used the Cox proportional hazards model to assess the 10-year risk of broadly defined CVD.

The Framingham functions represent a major step toward improved risk assessment and management and served as examples and motivation for researchers who wished to develop similar tools for uses other than primary prevention of CVD. Most commonly, the performance of risk prediction functions is assessed by their calibration and discrimination. Calibration addresses the question of how close model-based risks are to the observed outcomes. For example, consider a group of people whose 10-year model-based risk of developing a certain

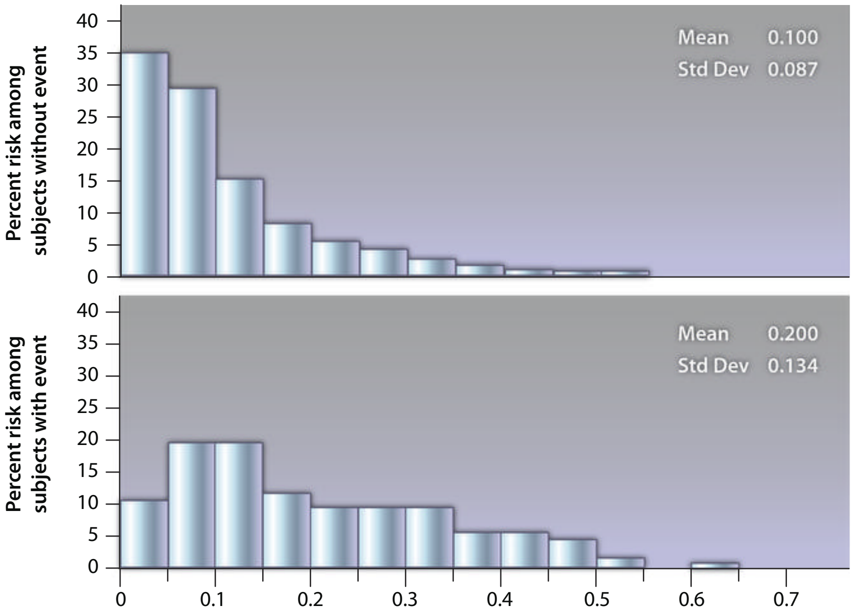

Fig. 1. Discrimination histogram. Shown are histograms of model-based risks for clinical events and nonevents. These histograms depict a model in which the probability mass is concentrated to the left for nonevents (many small risks) and is relatively flat for events. Further improvement in discrimination should be sought, especially through methods that would increase model-based risks for events. Std Dev, standard deviation.

[1]Department of Biostatistics, Boston University, Harvard Clinical Research Institute, Boston, MA 02118, USA. [2]Department of Mathematics and Statistics, Boston University, Harvard Clinical Research Institute, Boston, MA 02215, USA.
*Corresponding author. E-mail: mpencina@bu.edu

CREDIT: B. STRAUCH/SCIENCE TRANSLATIONAL MEDICINE

disease (the event) equals 10%. If the model is well calibrated and if we follow these people for 10 years, we expect that 10% of them will experience the event. Discrimination focuses on the model's ability to distinguish those who will experience the event from those who will not. For example, consider a situation in which all individuals who develop events have predicted risks of >50% and those who do not develop events have risks of <50%. Such a model displays perfect discrimination. Discrimination has traditionally been quantified using the area under the receiver operating characteristic curve (AUROC), also known as the c statistic. The AUROC has been shown to be equal to the probability that, given two subjects—one with and one without an event—the model assigns a higher risk to the one with the event. As such, the AUROC ranges from 0.5 for a random assignment to 1.0 for a perfect model. The AUROC values for the Framingham function usually fall in the range between 0.75 and 0.85, indicating good or very good discrimination. Still, there is room for improvement.

A substantial portion of biomedical research efforts and funding has been dedicated to identifying new predictors that can increase the AUROC. New blood biomarkers and genetic factors have been added to the Framingham functions, but their ability to improve discrimination proved disappointing (*6*, *7*). Research has shown that patient clinical variables with very large conditional effect sizes (standardized difference in means between those who do and do not develop events) are needed to meaningfully increase the AUROC when the baseline model has a good discrimination (*8*).

In response, Pencina and D'Agostino *et al.* (*9*) proposed two other methods with which to quantify improvement in the performance of a risk prediction model: integrated discrimination improvement (IDI) and net reclassification improvement (NRI). IDI can be interpreted as the increase in integrated (average) sensitivity plus integrated (average) specificity of the model after modifications have been made. IDI is estimated as the difference in the discrimination slopes of the two models under consideration (one before and one after modifications). The discrimination slope is defined as the difference in mean model-based risks between subjects with and without events and can be viewed as a summary statistic associated with the discrimination histograms depicted in Fig. 1, which considers model-based risks for events and nonevents separately. For models that discriminate well, the probability mass is concentrated to the left for nonevents (many small risks) and to the right for events (many high risks). The discrimination slope is the difference in means for the two distributions depicted. IDI measures how much we shifted the probability mass to the left for nonevents and to the right for events.

NRI assumes the existence of clinically meaningful risk categories. For example, the NCEP ATP III panel mentioned previously provides distinct treatment recommendations for people whose 10-year risk of CHD exceeds 20%, is less than 10%, and is in the 10 to 20% range. NRI measures the amount of correct reclassification among these categories after modification. IDI and NRI are more sensitive with respect to detecting improvement in model performance than the AUROC, but the practical conclusions derived from their applications frequently agree with those obtained when relying on the AUROC.

RETHINKING REGRESSION

A different approach to improving the performance of risk prediction models is based on the use of techniques other than regression to estimate individual risks. These approaches are driven by the data and intended to capture relationships more complex than the simple linear form imposed by the commonly used regression techniques. However, the impact of these new methods on cardiovascular risk prediction has been limited at best.

First, many of these more data-driven techniques lead to models that perform well on the development data but do not improve upon regression models when tested on validation samples. This is a key problem, because all algorithms are developed for application to new patient cohorts, not to the ones on which the models were developed. Second, many of the new methods rely on statistical analyses that are not easily performed with standard statistical packages, which limits applicability. Third, the new methods do not produce closed-form solutions (that is, simple formulas) that are simple to program and understand.

An informative overview of these and other problems has been given by Hand (*10*), who demonstrated that "when building predictive models of increasing complexity, the marginal gain from complicated models is typically small compared to the predictive power of the simple models." Despite his skepticisms, Hand expects that, with the modern computer revolution, it will become possible to develop techniques that meaningfully improve existing risk algorithms. The question herein is whether this time has arrived.

Chia *et al.* (*1*) propose a new technique to develop risk prediction algorithms, called SVM1.5, and illustrate its application using several data sets of different sizes and event rates (including some with very low event rates below 0.2%). For each clinical situation—which included assessment of risk of CVD-related and other adverse outcomes in patients with acute coronary syndrome or undergoing various surgical procedures—a development and training data set was used to produce the final algorithm, which was then evaluated on a validation data set. Validation is performed to mitigate the problem of the new algorithm appearing to perform better than it truly would ("overoptimism"), which is likely to result when performance is assessed on the same data used to develop the algorithm. This reuse of data is a particular concern for all data-driven techniques, which thus far have tended to perform well only on development data. But the results obtained by Chia *et al.* did not confirm this general trend. The new algorithm outperformed logistic regression in a majority of cases.

To decipher the extent to which risk prediction was improved with their new SVM1.5 algorithm, Chia *et al.* (*1*) contrasted its performance with that of logistic regression and more traditional versions of data-driven algorithms. The authors presented their results in terms of AUROCs and IDIs, the two most commonly used metrics for improvement in discrimination. Chia *et al.* also attached *P*-values to their comparisons, but we believe these to be of limited value as the focus should be on the changes in magnitude of the measures of improvement in discrimination rather than on their statistical significance.

So how meaningful were the improvements in risk prediction achieved by Chia *et al.*? The results seem to vary from application to application. For example, in table 1 in the Chia *et al.* paper (*1*), we see that with the DISPERSE and MERLIN cohorts, for the end point of cardiovascular death, the SVM1.5 algorithm yielded an AUROC of 0.76 compared to the logistic regression's AUROC of 0.72, a meaningful increase. On the other hand, for the outcome of myocardial infarction (MI), the AUROCs were much closer: 0.57 for logistic regression and 0.58 for SVM1.5. The performance for both models for MI is surprisingly weak, and one might speculate that the competing risk of death might play a role here. Some of the most substantial increases in AUROC and IDI were observed for cases in which the numbers of events (<60) and their rates (0.1 to 0.2%) were small. For example, the algorithms that predicted the risk of coma in the National Surgical Quality Improvement Program's cohort (event rate of 0.2%) had an AUROC of 0.62 when developed using logistic regression and 0.76 when developed with SVM1.5 (table 2 of Chia *et al.*), which is considered by the field to be a substantial improvement.

Another spectacular gain was observed for the Blue Cross Blue Shield of Michigan Cardiovascular Consortium (BMC2) cohort, for which algorithms that assessed the risk of mortality were constructed; logistic regression produced an AUROC of 0.82, and SVM1.5 had an impressive AUROC of 0.93 (table 3 of Chia *et al.*). The IDI equaled 0.010 (table 6 of Chia *et al.*), which suggests a very large improvement in separation of events from nonevents given that the event rate equaled just 0.001. But the small number of events (n = 19) puts the credibility of this result in question. On the other hand, the authors argued that their method is particularly well suited for cohorts with low event rates, and the results presented seem to confirm this. Because no clinically meaningful risk categories exist in many of the settings considered, Chia *et al.* do not present NRIs. We can speculate, however, that many of the improvements in the AUROCs and IDIs would translate into meaningful increments in NRIs. It is also important to note that the calibration of the proposed method appears to be at least as good as the calibration of the logistic regression models.

Are we then ready to recommend SVM1.5 to Framingham and other investigators for routine use in the development of risk prediction algorithms? On one hand, the potential improvements in model performance are encouraging. If the event rate is low, the method used by Chia *et al.* seems likely to produce an algorithm that will outperform the one based on a usual regression. Risk prediction algorithms are common for diseases with low incidence or prevalence. Predicting different forms of cancer or individual components of CVD are obvious examples for a possible application of SVM1.5. The gain in performance when using SVM1.5 comes at a very low cost from the clinical perspective: We do not need to measure an expensive biomarker to meaningfully increase the AUROC or IDI; instead, we make more efficient use of the predictors that are already being measured. This is very appealing from cost-analysis, development time, patient burden, and risk perspectives.

On the other hand, the evidence presented by Chia *et al.* (*1*) is anecdotal. Examples are sufficient to disprove a claim but never to prove it. Without formal proof of the superiority of SVM1.5, we cannot know whether the results will hold in settings different from the ones considered. Will the SVM1.5-based algorithms validate as well as they did for Chia *et al.* in other clinical settings? Furthermore, the application of SVM1.5 seems easy when reading the elegant work of Chia *et al.*, but how difficult the methods prove to be in practice remains to be seen by those willing to try. Another question that remains is how easy it will be to translate SVM1.5 results into a clinically usable tool. Can the results be expressed in terms of predicted risk in a given time horizon? Can the results be translated into "heart age" (*5*) or another easily interpretable tool for communication of risk to patients? Can the results be extended to apply to time-to-event outcomes? Will the extra effort necessary to employ the new algorithm be worth the gains in model performance? These are serious concerns that mandate that the promising results of Chia *et al.* be replicated in other settings before the new method for generating risk prediction algorithms can be confidently applied. Still, these concerns are not sufficient to discount the new method without further investigation. It definitely seems worth a try.

REFERENCES AND NOTES

1. C. -C. Chia, I. Rubinfeld, B. M. Scirica, S. McMillan, H. S. Gurm, Z. Syed, Looking beyond historical patient outcomes to improve clinical models. *Sci. Transl. Med. 4*, 131ra49 (2012).
2. T. R. Dawber, G. F. Meadors, F. E. Moore Jr., Epidemiologic approaches to heart disease: The Framingham study. *Am. J. Public Health* **41**, 279–286 (1951).
3. S. H. Walker, D. B. Duncan, Estimation of the probability of an event as a function of several independent variables. *Biometrika* **54**, 167–179 (1967).
4. 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* **285**, 2486–2497 (2001).
5. R. B. D'Agostino Sr., R. S. Vasan, M. J. Pencina, P. A. Wolf, M. R. Cobain, J. M. Massaro, W. B. Kannel, General cardiovascular risk profile for use in primary care: The Framingham Heart Study. *Circulation* **117**, 743–753 (2008).
6. T. J. Wang, P. Gona, M. G. Larson, G. H. Tofler, D. Levy, C. Newton-Cheh, P. F. Jacques, N. Rifai, J. Selhub, S. J. Robins, E. J. Benjamin, R. B. D'Agostino, R. S. Vasan, Multiple biomarkers for the prediction of first major cardiovascular events and death. *N. Engl. J. Med.* **355**, 2631–2639 (2006).
7. N. P. Paynter, D. I. Chasman, G. Paré, J. E. Buring, N. R. Cook, J. P. Miletich, P. M. Ridker, Association between a literature-based genetic risk score and cardiovascular events in women. *JAMA* **303**, 631–637 (2010).
8. J. H. Ware, The limitations of risk factors as prognostic tools. *N. Engl. J. Med.* **355**, 2615–2617 (2006).
9. M. J. Pencina, R. B. D'Agostino Sr., R. B. D'Agostino Jr., R. S. Vasan, Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Stat. Med.* **27**, 157–172, discussion 207–212 (2008).
10. D. J. Hand, Classifier technology and the illusion of progress. *Stat. Sci.* **21**, 1–14 (2006).

Funding: This work was supported by the U.S. National Heart, Lung, and Blood Institute's Framingham Heart Study (contract N01-HC-25195).

10.1126/scitranslmed.3004127

Citation: M. J. Pencina, Ralph B. D'Agostino Sr., Thoroughly modern risk prediction? *Sci. Transl. Med.* **4**, 131fs10 (2012).

Science Translational Medicine

Thoroughly Modern Risk Prediction?

Michael J. Pencina and Ralph B. D'Agostino, Sr.

Sci Transl Med **4**, 131fs10131fs10.
DOI: 10.1126/scitranslmed.3004127