VIEWPOINT

The Role of Physicians in the Era of Predictive Analytics

Allan D. Sniderman, MD
Royal Victoria Hospital, McGill University, Montreal, Quebec, Canada.

Ralph B. D'Agostino Sr, PhD
Department of Mathematics and Statistics, Boston University, Boston, Massachusetts.

Michael J. Pencina, PhD
Duke Clinical Research Institute, Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina.

Corresponding Authors: Michael J. Pencina, PhD, Duke Clinical Research Institute, Department of Biostatistics and Bioinformatics, Duke University, 2400 Pratt St, Durham, NC 27705 (michael.pencina@duke.edu) and Allan D. Sniderman, MD, McGill University Health Centre, 1001 Decarie Blvd, Royal Victoria Hospital, Glen Site C04.2458.2, Montreal, Quebec, Canada H4A 3J1 (allan.sniderman@mcgill.ca).

Every day, more information becomes available about factors that affect the risk of a clinical event. Predictive analytics incorporate this information into prognostic models that estimate the likelihood of this event for an individual patient. The Framingham Heart Study pioneered this approach[1] and such estimates have become core elements of clinical care and guideline recommendations, such as the recent ACC/AHA guidelines for managing blood cholesterol for primary prevention of cardiovascular disease risk.[2] Advances in predictive analytics and precision medicine have and will continue to change the practice of medicine.

However, to apply these risk estimates wisely, a physician must understand that the concept of risk is more subtle and complex than is generally appreciated. Specifically, it is important to distinguish between 2 meanings of risk: risk in the epidemiological sense (the risk of a group of individuals) vs risk in the clinical sense (the risk for individual members of the group). But how accurately does the overall risk for the group reflect the risk of each of the individuals who make up the group? This distinction is critical to appreciate because it underlies the essential role that physicians need to play in the era of predictive analytics. In this Viewpoint, cardiovascular risk will be used as an example, but the same concepts can be applied to risk prediction in any context.

Conditional Nature of Algorithms Assessing Individual Risk

What does a 10% risk of an event within the next decade mean to the individual for whom it was generated? Contrary to what is thought, this risk level is not that person's personal risk because probability is not meaningful in an individual context.[3] However, if a risk function accurately predicts the number of events for a group of 10 individuals at a 10% risk, on average, one person will experience an event. The question then becomes were all 10 individuals equally likely to experience this event? The answer matters because only if this is the case will individual risk be equal to group risk.

Hypothetically, if all past, present, and future predictors and processes that contribute to future events were known and quantifiable, algorithms could be constructed that produce perfect risk estimates for individuals—that is, they would predict with perfect accuracy whether an event would occur or not in every individual. This is not possible at present because only a fraction of the information that explains variation in outcome has been identified. Accordingly, estimates of risk are incomplete and conditional on the information that was included in the risk calculation.

To illustrate this point, consider the following example from the Multi-Ethnic Study of Atherosclerosis.[4] Based on the conventional risk factors, participants were first classified into 3 levels of cardiovascular risk: *low risk* (<3% likelihood of an event in 5 years), *intermediate risk* (3%-<10% likelihood of an event in 5 years), and *high risk* (≥10% likelihood of an event in 5 years). The observed event rates (1.4% for low, 5.5% for intermediate, and 12.1% for high risk) were within these boundaries in all groups and therefore confirmed the conventional estimation of risk. However, this does not establish that all the individuals within the group were at the same risk or even that the risks of all were within the range of the group to which they were assigned. The construction of these groups was based on the risk factors included in the model. Therefore, the precision of the estimates is conditional on these risk factors.

What happens to individuals if more information is added from other important risk factors not included in the conventional model, for example, coronary artery calcium? When the risk of each of the participants was recalculated based on the additional information from coronary artery calcification, 11.6% of low-risk individuals were reclassified to a higher grade of risk, 54.4% of intermediate-risk individuals were reclassified to a lower or higher grade of risk, and 35.8% of high-risk individuals were assigned to lower grades of risk. Overall, 26% were reclassified into low-, intermediate-, and high-risk groups. The model improved after the addition of coronary calcification, as evidenced by 2 measures. There was an increase in the model's ability to distinguish between those who do vs those who do not develop events (discrimination *C* statistic—defined as the probability that given one person who experienced an event and one did not, the model assigns a higher risk to the former,—increased from 0.76 to 0.81). There also was improvement in classification of individuals to clinically meaningful risk categories (net reclassification index, defined as the sum of differences in proportions of persons with and without events reclassified into more vs less appropriate risk categories, equaled 0.25).

This extensive reclassification establishes that individual risk within the original groups varied substantially. Still, there is no certainty that significant variation in risk does not persist among the individuals in the reclassified groups. The improved model remains imperfect because the *C* statistic is still far from 1. Another way to illustrate the imperfect nature of risk prediction involves using a different scale for evaluation of performance based on explained variation. Using discrimination slope (difference in average risks between those with events vs those without events) or other metrics of explained variation, such as different R^2-type

measures,[5] suggests that less than 20%, and often even less than 10%, of the variability in future cardiovascular event status can be captured by the model.[6] This further demonstrates that the components of the present algorithms used to estimate risk incompletely capture the determinants of risk.

The process of improved performance and reclassification of individuals could continue as long as new important risk factors are identified. Until all the major factors that determine risk have been identified and until all the necessary information is available for all the individuals who will make up the expanded cohorts from which the algorithms will be generated, individual risk prediction will remain imperfect. Expanding the pool of information and increasing the number of predictors requires samples with large numbers of participants. Current risk prediction models provide reasonable precision for a group, but uncertainty about any individual patient's true risk as estimated by these models remains.

How the Physician Can Mitigate Uncertainty in the Care of the Individual

Current risk prediction models are useful because they facilitate application of the overall results of clinical trials to patient cohorts that should resemble the patients who were in the original trial. Although many patients will resemble most of the patients in a treated group and consequently will have risk predictions that accurately estimate their outcomes, this will not be the case for all patients. Physicians should be aware of this uncertainty and remain alert to other features that were not principal characteristics of the participants in the study but may affect outcome in the individual patient. Moreover, when providing treatment recommendations, consideration should include not only who is at risk but also who is most likely to benefit from therapy. Considering "who will benefit?" rather than solely "who is at risk?" might be used as an additional criterion on which to base treatment decisions.[7,8]

Conclusions

Predictive analytics can improve clinical care by providing general recommendations for populations that can be incorporated into clinical guidelines. Predictive algorithms are an essential component of guideline recommendations. However, because predictive models imperfectly explain clinical outcomes, they do not estimate individual risk very well even when they accurately explain group risk. Consequently, these models cannot replace the physician in the process of care. Physicians are responsible for assessing an individual's constellation of problems and risk factors and then selecting with the patient the appropriate treatment for that individual. Physicians must understand the limitations as well as the strengths of the evidence as it applies to individual patients and must recognize additional factors not included in prediction models may alter risks or benefits.

ARTICLE INFORMATION

Conflict of Interest Disclosures: All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

Additional Contributions: We thank George Thanassoulis, MD, Research Institute of the McGill University Health Centre, and Curt Furberg, MD, PhD, professor emeritus, Wake Forest School of Medicine, Winston-Salem, NC, for their contribution to the development of concepts presented in this article, for which they were not compensated.

REFERENCES

1. D'Agostino RB Sr, Pencina MJ, Massaro JM, Coady S. Cardiovascular disease risk assessment: insights from Framingham. *Glob Heart*. 2013;8(1):11-23.

2. Stone NJ, Robinson JG, Lichtenstein AH, et al. 2013 ACC/AHA guidelines on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults. *J Am Coll Cardiol*. 2014;63(25 pt B):2889-2934.

3. Cohen LJ. *An Introduction to the Philosophy of Induction and Probability*. New York, NY: Oxford University Press; 1989.

4. 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.

5. Tjur T. Coefficients of determination in logistic regression models—a new proposal: the coefficient of discrimination. *Am Stat*. 2009;63(4):366-372.

6. 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*. 2008;27(2):157-172.

7. Sussman J, Vijan S, Hayward R. Using benefit-based tailored treatment to improve the use of antihypertensive medications. *Circulation*. 2013;128(21):2309-2317.

8. Sniderman AD, Toth PP, Thanassoulis G, Pencina MJ, Furberg CD. Taking a longer term view of cardiovascular risk: the causal exposure paradigm. *BMJ*. 2014;348:g3047.