edges gains while being clear-eyed about populations who don't benefit from those gains. Adopting methods that account for all factors that influence risk, and

An audio interview with Dr. Shrank is available at NEJM.org

for the interdependence of those factors, could be an important step in creating a more equitable health care payment system that better serves patients, including the most disadvantaged members of society.

1. Walker RJ, Smalls BL, Campbell JA, Strom Williams JL, Egede LE. Impact of social determinants of health on outcomes for type 2 diabetes: a systematic review. Endocrine 2014;47:29-48.
2. Joynt KE, Zuckerman R, Epstein AM. Social risk factors and performance under Medicare's value-based purchasing programs. Circ Cardiovasc Qual Outcomes 2017;10(5).
3. Nguyen CA, Gilstrap LG, Chernew ME, McWilliams JM, Landon BE, Landrum MB.
Social risk adjustment of quality measures for diabetes and cardiovascular disease in a commercially insured US population. JAMA Netw Open 2019;2(3):e190838.
4. Risk adjustment for socioeconomic status or other sociodemographic factors: technical report. Washington, DC: National Quality Forum, August 15, 2014 (https://www.qualityforum.org/Publications/2014/08/Risk_Adjustment_for_Socioeconomic_Status_or_Other_Sociodemographic_Factors.aspx).
5. Ash AS, Mick EO, Ellis RP, Kiefe CI, Allison JJ, Clark MA. Social determinants of health in managed care payment formulas. JAMA Intern Med 2017;177:1424-30.

Prediction Models — Development, Evaluation, and Clinical Application

Michael J. Pencina, Ph.D., Benjamin A. Goldstein, Ph.D., and Ralph B. D'Agostino, Ph.D.

When national lipid guidelines first incorporated a model based on data from the Framingham Heart Study — a turning point for the role of risk prediction in health care — that Massachusetts city was an anomaly: a community with extensive, available, longitudinal health data. Today, U.S. health care systems have amassed large, local data sets through adoption of electronic health records (EHRs) and the standardization associated with provider consolidation. More recently, payers have moved toward capitation and other value-based models. This shift places a higher premium on avoiding costly conditions altogether. These trends create greater demand for prediction models, since prevention is difficult without accurate identification of who specifically is at risk.

Prediction models' newfound importance and the emergence of model development based on machine learning raise questions about how to ensure their safety and efficacy, given their growing role in risk stratification, care pathways, and clinical outcomes. A systematic review comparing clinical prediction models based on regression with those based on machine learning revealed troubling weaknesses in model evaluation.[1] Given the number of emerging prediction models and their diverse applications, no single regulatory agency can review them all. This limitation, however, does not absolve models' developers and users from applying the utmost scrutiny in demonstrating effectiveness and safety. It also highlights the need for accepted standards for development, evaluation, and application of prediction models.

Fortunately, foundational principles for model creation and use have emerged.[2,3] These principles will have to be adapted and augmented for current conditions, which include new sources of data. We offer eight key considerations for the introduction and use of prediction models (see table for illustrative examples).[4,5]

1. Population at risk: Correct identification of persons at risk and the time when the model will be applied to inform intervention strategies is critical. Such identification requires a focus on the demographic characteristics, health status, and location of the patient population as well as on the clinical context in which a model will be used. The pooled cohort equations that drive current cholesterol guidelines, for example, are based on persons without atherosclerotic cardiovascular disease (ASCVD) who are 40 to 79 years of age and not currently receiving lipid-lowering treatment. Consequently, these equations should not be applied to people with a history of ASCVD, people currently taking lipid-lowering treatment, or people under 40 or over 79. Including broad swaths of the population who are at very low risk (e.g., women in obstetrics units) in a model for determining sepsis risk can make it harder to identify people with genuinely high risk. The

Considerations for Development of Prediction Models and Their Application in Cholesterol Guidelines and Sepsis Prevention.*			
Consideration	Comments	Pooled Cohorts Equation for Cholesterol Guidelines	Model for Improving Sepsis Prevention in Hospitals
1. Population at risk	Identify persons at risk to whom the model will be applied on the basis of demographic characteristics, health status, location, and clinical context.	Adults 40–79 yr of age free of CVD (model based on 24,626 NHLBI cohort participants)	Adult patients presenting to emergency department and admitted (model based on >42,000 encounters over 14 mo)
2. Outcome of interest	Use well-curated data, with outcomes that reflect the primary focus of care.	Atherosclerotic CVD (myocardial infarction, stroke, cardiovascular death)	Sepsis: ≥2 abnormal vital signs, suspected infection; ≥1 abnormal sign of end-organ failure
3. Time horizon	Starting point and duration of follow-up should align with goals of interventions.	10 yr after baseline measurement	4 hr; predictions made every hour of encounter
4. Predictors	Decisions about choices and number of predictors should take into account ease and time of collection, possible bias, model stability, and interpretation (e.g., understanding what outputs the model produces and identifying key predictors and their association with outcome).	Age, sex, race, smoking status, presence or absence of diabetes, systolic blood pressure, antihypertensive treatment, total and HDL cholesterol levels	34 time-varying physiological variables, 35 baseline covariates, 10 medication classes (more than 32 million data points)
5. Mathematical model	Balance performance with ability to understand, implement, and maintain the model. Burden of proof is placed on the more complex models. Be transparent and avoid black-box solutions.	Cox proportional-hazards regression	Multitask Gaussian Process Recurrent Neural Network Classifier
6. Model evaluation	Rigorously evaluate the model using data different from those used for development and collected in a setting that mirrors clinical application.	Validated in development and several external cohorts	Validation on 32,000 new encounters; model compared with National Early Warning Score for 48-hr sepsis detection
7. Translation to CDS	To translate model into CDS, determine intended use and what should be displayed. Separate evaluation of the CDS tool is necessary, including comparison with current practice (ideally, randomized).	Guidelines recommend lipid-lowering treatment when risk >0.075	Sepsis risk categorized as present, high, medium, or low; trial planned with sepsis bundle compliance as primary outcome
8. Clinical implementation	Incorporation into clinical workflows with training, performance engineering, monitoring, and updating, when necessary, is required.	Calculators on AHA/ACC websites, model programmed into EHR systems	CDS piloted in academic health system workflow

* AHA/ACC denotes American Heart Association/American College of Cardiology, CDS clinical decision support, CVD cardiovascular disease, EHR electronic health record, HDL high-density lipoprotein, and NHLBI National Heart, Lung, and Blood Institute.

more dissimilar the population used in developing the model is from the target population, the less accurate predictions will be.

2. Outcome of interest: Well-curated outcome data reflecting the primary focus of care are needed for model development. Using poorly defined or surrogate outcomes can lead to unintended adverse consequences. For example, data for in-hospital events, including death, may be the easiest to obtain, but prediction models that don't incorporate events occurring soon after

discharge might be misleading. Other decisions include appropriate handling of practice-driven choices (in the sepsis example in the table, for instance, administration of medications may be indicative of disease progression) and competing outcomes (which should be handled using statistical methods, rather than excluded or recorded as nonevents).

3. Time horizon: The starting point and duration of follow-up should align with the goals of interventions that will be informed by prediction models. Al-

though ranking patients according to risk may be less affected by the time horizon, absolute risks — which are often used in clinical decision making — will be more so. In a sepsis-prediction model, for example, the choice of starting point is closely connected with the level of care — emergency department, inpatient, or intensive care — and depends on whether our interest lies in prediction during the patient's entire stay or just the first few hours. A too-short window might improve model performance but reduce

actionability for outcomes that develop more slowly. The availability of data for updating predictions will also influence these choices.

4. Predictors: Historically, models relied on few informative predictors, which often lay along the causal pathway for the disease — in the pooled cohort equations, for example, lipid levels, blood pressure, smoking status, and presence or absence of diabetes. Drawing data from EHRs can substantially lengthen the list of potential predictors, but principles of parsimony still apply (so that models will be sufficiently stable in application). Ease of data collection and potential bias in model inputs should be scrutinized. Only predictors that can be measured at or before the point of model application should be included.

Model developers should identify the most important predictors and clarify the direction of their association with the outcome. In ASCVD models, for instance, antihypertensive treatment appears to increase risk, though we know from clinical trials that the opposite is true.[4] This apparent effect is explained by indication bias: people with higher risk are more likely to be treated.

5. Mathematical model: The majority of models in common use rely on regression techniques, but there is increasing interest in using advanced machine learning. It's important to strike a balance between the model's performance and users' ability to understand, implement, and maintain it. The burden of proof falls on developers of more complex models to demonstrate that they provide additional value — users and patients should not accept a true "black box." Reproducibility should be expected, and transparency of

reporting is essential, permitting independent evaluation and critical appraisal of the model and the methods used to derive it.

6. Model evaluation: Once a target model is identified, it needs to be rigorously evaluated using data that are different from those used to develop it and that have been collected in clinical settings as similar as possible to those where it will be applied. The lack of rigor and transparency in model validation is a pressing problem.[1] The availability of EHR data permits examination of heterogeneity by means of multiple evaluations involving different centers and hospitals. Developers should apply standard statistical measures of discrimination and calibration and should compare their model to models in current use.

7. Translation to clinical decision support: If a model performs well using independent, relevant data, it can be translated into clinical decision support (CDS). Users will have to make choices about how the model will be applied (for instance, allocation of limited resources to highest-risk persons or determination of risk levels that warrant various responses) and about what information should be displayed (such as estimated absolute risk, risk relative to peers, percentiles, categories and reference ranges, risk thresholds, and key drivers of risk). For example, the lipid guidelines recommend that the 10-year risk of ASCVD be calculated and that lipid-lowering treatment be considered for persons with a risk above 0.075.[4] If the prediction model and resulting CDS tool are to be incorporated into clinical information technology systems, they should initially run silently to allow for user feedback and data collection and evaluation.

A high-performing model does

not guarantee high-performing CDS. The CDS needs to be evaluated separately, using measures addressing clinical implications. Though care delivery incorporating a prediction model may frequently be superior to the alternative, a formal, randomized comparison with current practice is preferable. The outcomes in such a trial may include improved risk stratification, decision making, or other process measures, as well as cost-effectiveness relative to usual care.

8. Clinical implementation: Once the model and resulting CDS have been shown to improve care delivery, they can be incorporated into clinical workflows. Training and performance engineering will be needed to facilitate understanding and appropriate use.[5] Regular monitoring by an automated platform permits gathering of information on impact, appropriateness of use, and needed changes. As clinical populations shift and clinicians adapt to model alerts, new risk groups may emerge, necessitating adjustments. Although implementation is the last step in the process, its design principles should influence the decisions made in previous steps.

Progress in technology and available data create unprecedented opportunities for prediction models to inform, personalize, and improve care. Those opportunities also place an onus on professionals who develop, implement, maintain, and use these tools to do this work responsibly.

From the Department of Biostatistics and Bioinformatics, Duke Clinical Research Institute, Duke University School of Medicine, Durham, NC (M.J.P., B.A.G.); and the Department of Mathematics and Statistics, Boston University, Boston (R.B.D.)

1. Christodoulou E, Ma J, Collins GS et al. A systematic review shows no performance

benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019;110, 12-22.

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

3. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 2015; 162:55-63.

4. Goff DC Jr, Lloyd-Jones DM, Bennett G, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Circulation 2014;129:Suppl 2: S49-S73.

5. Sendak MP, Ratliff W, Sarro D, et al. Sepsis watch: a real-world integration of deep learning into routine clinical care. JMIR Med Inform (in press).

The Invisible Hand — Medical Care during the Pandemic

Michelle M. Kittleson, M.D., Ph.D.

I met him on March 3, 2020, a 70-year-old man with a 6-month history of classic stable angina. He had left-arm achiness whenever he walked uphill, reliably triggered by the same level of exertion and always relieved with rest. A stress test showed a large, reversible inferolateral defect, prompting consultation with me. At the time of his visit, travel restrictions had been issued for China, Iran, Italy, and South Korea, and the first Covid-19–related death had been reported in the United States. But on that day, in my office on Wilshire Boulevard in Los Angeles, Covid-19 wasn't even a blip on our radar.

Rather than ordering a knee-jerk coronary angiogram, I explained to the patient the 2014 Focused Update of the Guideline for the Diagnosis and Management of Patients with Stable Ischemic Heart Disease,[1] which recommends coronary angiography only in patients with presumed stable ischemic heart disease with unacceptable ischemic symptoms despite medical therapy. He was the perfect candidate for medical therapy, because when we first met he was receiving none.

Even better, the results of the ISCHEMIA (International Study of Comparative Health Effectiveness with Medical and Invasive Approaches) trial had been presented a few months before his visit.[2] This landmark trial compared optimal medical therapy or optimal medical therapy plus routine cardiac catheterization with revascularization in patients with stable angina. Its conclusion was a validation of the 2014 guidelines: an invasive approach did not reduce the risk of myocardial infarction or cardiovascular death. The results were also satisfying from a pathophysiological standpoint: a focal stenosis is the marker of the systemic disease of atherosclerosis, so it made more medical sense to treat the disease with medications rather than just fix the stenosis with a stent.

On November 16, 2019, the presentation of the ISCHEMIA trial was met with great fanfare at the American Heart Association Scientific Sessions. Internet medical pundits debated the finer points: Did the secondary end points actually favor intervention? Were the end points adjudicated fairly? And why were the results not simultaneously published in a high-impact journal? In retrospect, those passionate discussions seem quaint; just a day later, the yet-to-be-named SARS-CoV-2 infected the first patient in Hubei Province, China.

But on March 3, 2020, when faced with a patient who perfectly fit the profile of an ISCHEMIA trial enrollee, I delighted in the opportunity to provide guideline- and evidence-based therapy that made pathophysiological sense, though the patient was suspicious, and his wife more so. Still, they listened politely as I explained the pathophysiology of atherosclerosis. They even smiled when I described the limitations of the "oculostenotic reflex," an interventional cardiologist's shorthand for the see-a-blockage–fix-a-blockage approach to coronary artery disease. The man agreed to start taking an aspirin, a beta-blocker, and a statin, though his wife made this plan contingent on an angiogram scheduled a few weeks later.

I acquiesced to the angiogram because I knew that risk–benefit calculations are not just for physicians; patients perform them, too. I worried about the complications of potentially unnecessary angiography that would not improve survival and might not be necessary to improve quality of life, if medical therapy worked its magic. The patient and his wife worried about a ticking time bomb in his chest. I knew it was