VIEWPOINT

Validation and Utility Testing of Clinical Prediction Models
Time to Change the Approach

Amin Adibi, MSc
Respiratory Evaluation Sciences Program, Collaboration for Outcomes Research and Evaluation, Faculty of Pharmaceutical Sciences, University of British Columbia, Vancouver, British Columbia, Canada.

Mohsen Sadatsafavi, PhD, MD
Respiratory Evaluation Sciences Program, Collaboration for Outcomes Research and Evaluation, Faculty of Pharmaceutical Sciences, University of British Columbia, Vancouver, British Columbia, Canada.

John P. A. Ioannidis, MD, DSc
Meta-Research Innovation Center at Stanford (METRICS), Stanford Prevention Research Center, Department of Medicine, Stanford University School of Medicine, Stanford, California.

Corresponding Author: John P. A. Ioannidis, MD, DSc, Stanford Prevention Research Center, 1265 Welch Rd, Stanford, CA 94305 (jioannid@stanford.edu).

The growth in the publication of clinical prediction models (CPMs) has been exponential, largely as a result of an ever-increasing availability of clinical data, inexpensive computational power, and an expanding tool kit for constructing predictive algorithms. Such an abundance of CPMs has led to an overcrowded, confusing landscape in which it is difficult to identify and select the best, most useful models.[1] Few models are externally validated by the same researchers who developed them, and even fewer by independent investigators. Only 592 (43.3%) of 1366 cardiovascular CPMs in the Tufts PACE Clinical Predictive Model Registry reported at least 1 validation.[2] The proportions of models in the Tufts registry that reported at least 2, 3, and 10 validations were 20.1%, 12.8%, and 2.9%, respectively.[2] A few select CPMs, such as the Framingham Risk Score and EuroSCORE, have had numerous validations.

However, even these models are subject to modifications (eg, adding or removing a predictor variable), with the resulting modified model not revalidated externally. Fragmented efforts that assess only one model at a time do not allow for reliable ranking of the comparative performance of the many CPMs available for the same clinical application. A small number of specific models eventually dominate clinical practice based on tradition and herding behavior, rather than high-quality evidence that they outperform competitors.

> Such an abundance of clinical prediction models has led to an overcrowded landscape in which it is difficult to identify and select the most useful models.

The overabundance of de novo model development is wasteful and may exaggerate the loss of some good CPMs in the noise. For example, a systematic review found 363 CPMs for the risk of cardiovascular disease in general population.[3] However, usefulness of most of these models was deemed unclear because much-needed external validation studies and investigations of model utility were lacking. It is thus imaginable that some excellent CPMs may remain largely unused, while poor ones may remain popular by inertia. Moreover, besides gaps in assessing predictive accuracy (discrimination and calibration), clinical utility of CPMs is rarely assessed. In particular, randomized trials of CPMs remain uncommon.

Clinical utility would require evaluating whether using models results in better outcomes that matter to patients and physicians. However, assessing scores of CPMs proposed for each condition in any sort of clinical trial is utopian. A recent systematic review identified 408 clinical prediction models for chronic obstructive pulmonary disease outcomes,[4] which is more than the entire number of completed phase 3 trials in chronic obstructive pulmonary disease registered in ClinicalTrials.gov at the time of writing this Viewpoint. Moreover, if some trials are conducted, they may fail to show favorable clinical outcomes, not because predictive modeling is ineffective in the setting or condition under study, but because subpar models are evaluated. Overall, ways to identify the best models with the highest potential among the large space of competing candidates are needed.

It may be possible to meet this challenge by taking inspiration from the principles underlying natural selection and selective breeding. For millennia, humans have developed their desired phenotypic traits in plants and animals by giving a survival advantage to those who possessed them. Perhaps the same strategy can help to refine predictive analytics. What if a "survival advantage" could be given to the models that perform better, while waiting for the best models to emerge?

Clinical prediction models could be made responsive by hosting them in the cloud, implementing a standardized language to communicate inquiries to the models, and getting results from them in real time as part of routine patient-physician encounters and provision of care. With proper consideration for privacy and security, patient data could be streamed from administrative databases or electronic medical/health records (EMRs/EHRs) to continuously validate models. Through this process, better-performing models could obtain a survival advantage by ranking models based on their real-time external validation performance, in terms of both discrimination and calibration. Physicians and patients may prefer the models that have the best performance at any point in time in the settings that are most relevant to them.

An exciting possibility would be to create models that are constantly being updated and recalibrated,[1,5] similar to "living" meta-analyses, either automatically or manually, to adapt to the incoming data, new settings, and new clinical practices. This not only transforms models into dynamic entities beyond the snapshot version captured in a publication but also could facilitate clinical translation by allowing clinicians and health care centers to adapt models to their local settings.

Many of the components of such a system are already in place. Versatile standards have been developed for communicating health data, such as Fast Healthcare Interoperability Resources, which ensures semantic consistency as patient data are exchanged among different systems.[6] Standardized application programming interfaces that enable predictive models among other pieces of software to communicate queries and results with other software and medical devices are standard and commonplace in the software industry. Commercial cloud services that allow data to be shared securely and responsibly and in compliance with the Health Insurance Portability and Accountability Act (HIPAA) in the US (or equivalent requirements in other countries) are already available.

Benefits of developing such an integrated infrastructure go beyond high-throughput validation of CPMs. Cloud-based model repositories with direct model access also could improve the peer-review of CPMs. Direct model access could prove useful in exposing potential shortcomings of models by allowing reviewers to deeply examine input-output relationships. Similar to preprint servers, such repositories could be part of the scientific publication process to enable sharing the model with and collecting feedback from a wider audience. Additionally, direct access to CPMs also could make predictive analytics more equitable by empowering patients and advocacy groups to scrutinize prediction models and advocate their values and interests.

Such a survival-of-the-fittest strategy makes model validation as automated, as nondisruptive, and as low maintenance as possible. The research community has sought to improve the reproducibility of predictive analytics by introducing reporting guidelines for predictive models and promoting preregistration.[7] This proposal could complement these efforts by facilitating clinical translation once the proper infrastructure is in place.

The use of such an infrastructure would facilitate also selecting and testing the best candidate CPMs for their clinical utility. Trials may use point-of-care randomization whereby patients, physicians, or clinical units could be randomized to 2 or more strategies. The EMR/EHR environment could identify eligible patients and encounters and flag them for randomization. Depending on the types of questions asked, these trials could have simple (modified) or even entirely waived informed consent. The tested strategies may address questions such as when it is best to provide predictive information (early or later in the process of the clinical encounter and provision of care); how predictive information should exactly be presented (eg, assertive vs suggestive); and, in some cases, whether offering specific predictive information is better than not making it available at all.

There is extensive evidence from A/B testing (the equivalent of clinical trials in the technology industry) that running tens of thousands of simple randomizations on an online environment is feasible.[8] Seemingly subtle differences on how information is exactly presented can make a big difference. For example, what font is used or how fast the website is may translate into tens of millions of dollars in revenue gained or lost. Moreover, it is impossible to guess these effects before a randomized trial is done.

The same issues may apply to the availability of predictive analytics. Making the information available early vs later in the care process may affect what other tests are done, what management choices are adopted, and eventually, clinical outcomes that matter. All processes and outcomes could be documented in detail within the EMR/EHR system. The same applies to how predictive information is presented, such as with various types of measures that convey discrimination and calibration; with various ways to provide estimates of absolute risks (eg, visually, with numbers, with different schematics); and with different types of explanations or accompanying material.

In addition, in many clinical areas it may be entirely unclear whether predictive information is clinically useful or not. Then, randomization may involve making this information available or withholding it and allowing physicians and patients to make decisions without it, based on their appraisal of the evidence or based on simple heuristics. In several settings, fast and frugal heuristics[9,10] that are informed just by answering a couple of key questions may outperform complex models in practice. For example, some complex models may confuse clinicians and may waste time and resources in trying to collect the necessary factors to inform them (if they are not part of routine data collection). Moreover, complex models may be difficult to understand and to communicate despite the best efforts, and this may lead to suboptimal use.

Regardless of the questions asked in these studies, it is important to use the predictive models that are proven to be best, as they evolve from the survival of the fittest screening that may occur. As thousands of CPMs are being developed, high-throughput evaluation, validation, and utility testing of predictive analytics are long due.

ARTICLE INFORMATION

Published Online: March 5, 2020. doi:10.1001/jama.2020.1230

Conflict of Interest Disclosures: None reported.

Funding/Support: This project is supported by a Canada Foundation for Innovation John R. Evans Leaders Fund (project 38222). Dr Sadatsafavi receives salary support from the Canadian Institutes of Health Research and Michael Smith Foundation for Health Research. The work of Dr Ioannidis is supported by an unrestricted gift from Sue and Bob O'Donnell. METRICS is supported with funding from the Laura and John Arnold Foundation.

Role of the Funder/Sponsor: The funders had no role in the preparation, review, or approval of the manuscript or decision to submit the manuscript for publication.

REFERENCES

1. Moons KGM, Kengne AP, Grobbee DE, et al. Risk prediction models, II: external validation, model updating, and impact assessment. *Heart*. 2012;98(9):691-698. doi:10.1136/heartjnl-2011-301247

2. Tufts PACE Clinical Predictive Model Registry. Accessed January 27, 2020. http://pacecpmregistry.org/

3. Damen JAAG, Hooft L, Schuit E, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ*. 2016;353:i2416. doi:10.1136/bmj.i2416

4. Bellou V, Belbasis L, Konstantinidis AK, Tzoulaki I, Evangelou E. Prognostic models for outcome prediction in patients with chronic obstructive pulmonary disease: systematic review and critical appraisal. *BMJ*. 2019;367:l5358. doi:10.1136/bmj.l5358

5. Janssen KJM, Moons KGM, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol*. 2008;61(1):76-86. doi:10.1016/j.jclinepi.2007.04.018

6. Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. *J Am Med Inform Assoc*. 2016;23(5):899-908. doi:10.1093/jamia/ocv189

7. Moons KGM, 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*. 2015;162(1):W1-W73. doi:10.7326/M14-0698

8. Kohavi R, Tang D, Xu Y, Hemkens LG, Ioannidis JPA. Online randomized controlled experiments at scale: lessons and extensions to medicine. *Trials*. 2020;21(1):150. doi:10.1186/s13063-020-4084-y

9. Djulbegovic B, Hozo I, Dale W. Transforming clinical practice guidelines and clinical pathways into fast-and-frugal decision trees to improve clinical care strategies. *J Eval Clin Pract*. 2018;24(5):1247-1254. doi:10.1111/jep.12895

10. Goldstein DG, Gigerenzer G. Fast and frugal forecasting. *Int J Forecast*. 2009;25(4):760-772. doi:10.1016/j.ijforecast.2009.05.010