EDITORIAL

Risk Prediction With Electronic Health Records
The Importance of Model Validation and Clinical Context

Benjamin A. Goldstein, PhD; Ann Marie Navar, MD, PhD; Michael J. Pencina, PhD

Related article page 1007

Historically, risk models have been derived using data from large epidemiologic cohorts or clinical trials. Although these data sources are often high quality, their external generalizability may be limited for at least 2 reasons. First, the populations included in the cohort or trial are often narrowly defined and not representative of all adults.[1] Recent efforts to combine data from multiple cohorts have led to risk prediction models with broader external generalizability. The pooled cohort equations used in the 2013 American College of Cardiology/American Heart Association cholesterol guideline[2] and the Cohorts for Heart and Aging Research in Genomic Epidemiology-Atrial Fibrillation (CHARGE-AF)[3] score were both based on data pooled from multiple observational cohort studies. Second, how data are collected in prospective trials and cohort studies may not match how data are collected in clinical settings.

The rise of comprehensive electronic health records (EHRs) offers promising opportunities for application of existing risk scores and for development and validation of new models. The EHR data include large numbers of individuals, usually greatly exceeding sample sizes available in individual trials or registries; a recent review by Goldstein et al[4] found that risk scores developed from EHRs used data on a median of 26 100 people, with more than one-third of studies using more than 100 000 patient records. Moreover, collected data represent what is actually available in clinical practice. Implementation of risk models in the EHR greatly enhances the potential usability of those models by making them immediately available at the point of care, which obviates the need for clinicians to calculate the risk manually. On the other hand, EHR-based data also have limitations in terms of data quality and completeness of data capture.[5]

Thus, the question arises how well risk models developed in more traditional data sources perform when applied in the EHR. Because of the differences in the populations and how data are derived, the performance of risk scores derived in trials or cohorts may vary when implemented in the EHR. In this issue of *JAMA Cardiology*, Kolek et al[8] evaluated the cohort-derived CHARGE-AF score in their patient population using the Vanderbilt University Medical Center EHR. They found that the discrimination of the CHARGE-AF score in the EHR (C index, 0.708) was remarkably similar to what was observed in the original study's 2 validation cohorts (C indices, 0.66 and 0.71).[3] Although this is encouraging, similar results may not be achieved in other health systems. For example, the algorithm for identification of atrial fibrillation (AF) cases adopted by Kolek et al incorporated billing codes, echocardiographic analyses, and natural language processing, a sophisticated approach that is unlikely to be easily replicated across all EHRs and for all conditions to be studied.

Even using their advanced EHR-derived phenotype, the authors noted high degrees of miscalibration among high- and low-risk individuals, which underpredicted risk in the lowest-risk groups and overpredicted risk in the highest-risk groups. The authors concluded that the performance of the CHARGE-AF model is limited in their institution's EHR. The implications of a risk model with poor calibration at either end of the predicted risk spectrum depend on how this model is implemented in clinical practice. If the model is meant to estimate the burden of AF in different subgroups or to convey to each individual in the health system his or her numeric risk, problems arise. However, if the main goal is to identify the highest-risk adults, then the degree of overestimation at the highest end of risk matters less than accurate identification of who belongs in the highest risk category. This highlights the importance of knowing the clinical context of how a risk model will be used in assessing its utility. Metrics of calibration assess the former set of questions, whereas metrics of discrimination (eg, C indices) assess the later.[9]

Kolek and colleagues[8] suggest that those at the highest risk may benefit from more aggressive prevention. However, although observational data suggest potential roles for β-blockers, angiotensin-converting enzyme inhibitors, angiotensin receptor blockers, or statins to reduce incident AF,[6] it is important to note that at present no large randomized trials have demonstrated the effectiveness of primary preventive strategies in AF. Another possible use of an AF risk score is to increase screening for AF. However, how to screen, who to screen, and how to treat adults identified with subclinical disease is actively debated.[7] The simplest and most immediate application of an AF risk score might be in the identification of the highest-risk candidates for participation in clinical trials for more aggressive AF screening to test primary preventive strategies.

The clinical context of a prediction model should also guide the broader question of the extent to which EHR data should be leveraged to augment or replace the cohort-derived or trial-based risk scores. The work of Kolek et al[8] shows how to carefully define a cohort, phenotype an outcome, and extract extensive predictors. Using this study data, the authors could easily develop an AF risk model specific to their institution. For internal quality improvement initiatives (eg, prediction of readmission, targeted follow-up) or clinical trial recruitment, this approach could be preferred. This type of risk modeling is more and more common: during the past 7 years, publications using EHRs to develop risk prediction models rose steadily.[4] Perhaps not surprisingly, few of these studies have been multicenter studies, with most consisting of a risk model optimized for the specific center. By using the characteristics of the specific patient population as well as the idiosyncrasies of the EHRs, a center-based risk model has the potential to perform better than a more general model. However, locally derived models may or may not perform well in other health systems, particularly in those that lack equally robust algorithms for identification of clinical phenotypes. Moreover, if risk scores are being used to guide care, an argument can be made for standardized algorithms being applied across the nation. In this scenario, an approach that used data pooled across multiple cohorts or multiple EHR systems might be preferred.

The emergence of the EHRs greatly increases the interest in and potential applications of predictive models in health systems. However, as the work of Kolek et al[8] illustrates, before models can be recommended for health system–wide applications, validation work needs to be conducted to better understand the strengths and weaknesses of the selected model in the context of the intended clinical use.

ARTICLE INFORMATION

Author Affiliations: Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, North Carolina (Goldstein, Pencina); Center for Predictive Medicine, Duke Clinical Research Institute, Durham, North Carolina (Goldstein, Navar, Pencina); Division of Cardiology, Duke University School of Medicine, Durham, North Carolina (Navar).

Corresponding Author: 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).

Published Online: October 12, 2016. doi:10.1001/jamacardio.2016.3826

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

REFERENCES

1. Hordijk-Trion M, Lenzen M, Wijns W, et al; EHS-CR Investigators. Patients enrolled in coronary intervention trials are not representative of patients in clinical practice: results from the Euro Heart Survey on Coronary Revascularization. *Eur Heart J*. 2006;27(6):671-678.

2. Goff DC Jr, Lloyd-Jones DM, Bennett G, et al; American College of Cardiology/American Heart Association Task Force on Practice Guidelines. 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(25)(suppl 2):S49-S73.

3. Alonso A, Krijthe BP, Aspelund T, et al. Simple risk model predicts incidence of atrial fibrillation in a racially and geographically diverse population: the CHARGE-AF consortium. *J Am Heart Assoc*. 2013;2(2):e000102.

4. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review [published online May 17, 2016]. *J Am Med Inform Assoc*. doi:10.1093/jamia/ocw042

5. Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. *Med Care*. 2013;51(8)(suppl 3):S30-S37.

6. Camm AJ, Lip GYH, De Caterina R, et al; ESC Committee for Practice Guidelines (CPG). 2012 focused update of the ESC Guidelines for the management of atrial fibrillation: an update of the 2010 ESC Guidelines for the management of atrial fibrillation: developed with the special contribution of the European Heart Rhythm Association. *Eur Heart J*. 2012;33(21):2719-2747.

7. Healey JS, Sandhu RK. Are we ready for mass screening to detect atrial fibrillation? *Circulation*. 2015;131(25):2167-2168.

8. Kolek MJ, Graves AJ, Xu M, et al. Evaluation of a prediction model for the development of atrial fibrillation in a repository of electronic medical records [published online October 12, 2016]. *JAMA Cardiol*. doi:10.1001/jamacardio.2016.3366

9. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21(1):128-138.