Learning About Machine Learning: The Promise and Pitfalls of Big Data and the Electronic Health Record

Rahul C. Deo, MD, PhD; Brahmajee K. Nallamothu, MD, MPH

In medicine, we are often interested in understanding differences between those with and without a specific disease. Such differences may point toward modifiable risk factors or, when combined into a quantitative model, allow us to predict who is at high risk of disease development to direct treatment. Along these lines, predictive models of heart failure (HF) have been a popular target: in the past decade, over 28 models of incident HF have been published.[1] Most have good discriminatory properties (C statistics ranging from 0.70 to 0.89) and were developed using a modest number of predictors (typically <15), combined in linear models. Study populations varied from traditional epidemiological cohorts with regularly scheduled visits and physician-adjudicated outcomes to large patient databases that rely on diagnostic codes for defining HF cases and supplying predictive features.

Article, see p 649

In recent years, there has been increasing excitement about the application of machine-learning strategies toward these same problems.[2] Although such approaches are based on the same goal of patient classification, the machine-learning community, with its roots in computer science, tends to embrace far larger numbers of predictors, sometimes transformed or grouped or even derived empirically through feature engineering. These new predictors are then incorporated into a more flexible range of models to improve performance. Beyond attempting to achieve superior models, practitioners of machine learning also look to implement clinical decision support tools that use model predictions to guide physician behavior.

In this issue of *Circulation: Cardiovascular Quality and Outcomes*, Ng et al[3] describes an application of machine-learning approaches toward the problem of predicting incident HF within the electronic health record (EHR) with its potentially vast trove of data. A unique aspect of this study is the authors' interest in not just reporting their model performance but its sensitivity to critical aspects of the model building process, including quantity and diversity of input data. Their lessons have important and practical implications for others also trying to apply machine-learning approaches to their own EHRs.

The work by Ng et al[3] represents an example of supervised learning, in that patient samples are labeled as either having HF or not. To tackle this problem, the authors had to work over several critical steps. First, they had to decide on an appropriate response variable. Rather than use a time-to-event analysis, which was used in most prior studies, the authors have focused on a binary classification of patients into HF cases and controls, although they do look at different time horizons. In contrast with a 2-group classification problem, time-to-event analyses are hampered by availability of a limited set of machine-learning algorithms.

Once classification labels were defined, their next step was to focus on how to handle predictive features. In complex learning examples such as this one, features rarely can be taken raw and input into the model. The primary reasons are problems of sparsity and the curse of dimensionality—in which model performance degrades rapidly when the numbers of features greatly exceeds the number of training samples. Individual features are likely to be informative only if they are relevant to a large percentage of cases and controls. If few cases are positive for any given feature, it is unlikely to be discriminative; in particular, a negative result for that feature will be shared among the vast majority of cases and controls alike. To circumvent this, the authors grouped together features into superfeatures, such as grouping drugs into parent classes or individual diagnostic codes into hierarchical categories. This reduced the feature space from the tens of thousands to a workable level in the several hundreds. Feature grouping, although a necessity, can also obscure both the performance and the interpretability of the model.

After preprocessing features, the next step was to choose a model for prediction and an associated algorithm to fit its parameters. Ng et al[3] chose 2 different but well-established algorithms to model the data: the lasso variant of logistic regression and random forests.[3] Although different, both exemplify the principle of regularization, a key attribute of high-quality learning models in the face of large numbers of features. To understand the need for regularization, one must first remember that the goal of model developing is not to optimally predict labels within one's own data (the training data set), but to generalize to as of yet unseen data. The surest way to fit well to your training data is to have an incredibly flexible model with large number of features. However, such a model

The opinions expressed in this article are not necessarily those of the editors or of the American Heart Association.

From the Division of Cardiology, Department of Medicine; Cardiovascular Research Institute; Institute for Human Genetics; and Institute for Computational Health Sciences, University of California San Francisco, and California Institute for Quantitative Biosciences (R.C.D.); and VA Health Services Research and Development Center for Clinical Management Research, VA Ann Arbor Healthcare System, MI; Michigan Center for Health Analytics and Medical Prediction (M-CHAMP), Department of Internal Medicine, University of Michigan Medical School, Ann Arbor (B.K.N.).

Correspondence to Rahul C. Deo, MD, PhD, University of California, San Francisco, 555 Mission Bay Blvd S, Room 452S, San Francisco, CA 94158. E-mail rahul.deo@ucsf.edu

(*Circ Cardiovasc Qual Outcomes.* 2016;9:618-620. DOI: 10.1161/CIRCOUTCOMES.116.003308.)

***Circ Cardiovasc Qual Outcomes* is available at http://circoutcomes.ahajournals.org**

DOI: 10.1161/CIRCOUTCOMES.116.003308

will generally overfit to the training data set and do abysmally when you approach new data. Regularization imposes a penalty for the flexibility or complexity of the model to improve performance on unseen data. Both the lasso variant of logistic regression and random forest takes this issue into account. Random forest appeared to just edge out lasso logistic regression in this work, albeit at greater computational cost.

What conclusions can we derive from this work? The first is that their model performance (area under the receiver operating characteristic curve ≈0.79) was on par with or somewhat inferior to prior published models of incident HF (eg, Framingham Heart study,[4] area under the receiver operating characteristic curve ≈0.86; Kaiser Permanente model,[5] area under the receiver operating characteristic curve ≈0.88–0.89). Although more studies such as this one are needed, it is unlikely that machine-learning approaches applied to current EHR data will markedly improve on such existing predictive models. Their advantage must come from the ability to generate these models using data obtained at vastly lower costs and to incorporate their results more directly into clinical decision-making.

Yet, the most promising aspects of this study are unrelated to this particular model's performance within this particular EHR. More important is their work on learning about machine-learning approaches and their admirable decision to openly share the source code used during the analysis with readers. In a series of analyses, for example, the authors evaluated model performance after systematically varying several data requirements, including the prediction window length, observation window length, number of different data domains (diagnosis, medications, hospitalization), data quantity of patient records, and density of patient encounters. Not surprisingly, the model of Ng et al[3] was most effective in predicting imminent HF diagnoses—those occurring within 1 year or less—and did more poorly for predictions that happened at an interval more removed from the observation window where predictive features were collected. As the authors note, this has obvious limitations in terms of clinical utility—but may, nonetheless, still benefit a subset of patients. Among predictive features, individual diagnoses and medication order data types were most valuable on their own for improving model performance. In the combined model, knowledge about patient hospitalizations also improved discrimination substantially.

Another valuable feature of this study was its insights into how much data were needed to improve the model. The authors found that beyond a duration of 2 years of observation, there was limited benefit to the model, perhaps again related to the fact that HF diagnoses were most likely to be made soon after the observation window elapsed, and so the immediate window preceding diagnosis may reflect an escalation of treatment of conditions that may predispose to the development of HF. The authors also found that a larger training set was of value, but the benefits leveled off around 3500 to 4000 patients. Finally, model performance was best with a high density of visits in the observation window, although by applying thresholds of minimum number of visits, one begins to curtail the number of patients eligible for prediction. By tackling issues around how model performance varies because of these inputs, the authors rightfully conclude that we can generate possible guidelines for training effective disease onset predictive models.

Despite this promise, there remain some important pitfalls. The validity of predictive models relies on accurately defining cases and controls. EHRs have many uses but few would say that harried physicians and diagnostic coders use *International Classification of Diseases*, 9th Revision, codes as a comprehensive representation of a patient's diagnoses. In the development of models of incident diagnoses, there is a greater concern of the impact of false negatives than false positives. Specifically, it is important that the patients did not already have HF or we run the risk of using physician actions based on an implicit (ie, noncoded) diagnosis of HF, such as prescribing HF medications or targeting specific features of the physical examination, as predictors of a future diagnosis of HF. Although the authors have done their best to guard against this possibility with careful inclusion criteria in their cohort construction, an examination of some of the predictive features highly associated with the outcome (eg, combined α-beta blocker use) suggests that some patients might already have had HF or perhaps asymptomatic reduced ejection fraction.

What are the next steps? There is already activity in a few directions. One avenue will incorporate novel features such as wearable devices with the hope of improving prediction of incident disease or hospitalization. One may also try additional groupings of existing features with the hope of improving the model. In addition, it will be important to replicate this work in other health systems because one of the major gaps in machine-learning approaches is generalizability. All of the unique features the authors discovered in this single health system may have little to no applicability in a hospital across the street. Finally, one hopes that these predictive models will transition into actionable insights through interventions, such as clinical decision support—with the end result of better outcomes for patients and populations. Demonstration of efficacy may require randomized messages within the EHR—akin to Google A/B testing—although there may be ethical hurdles to implementing these within routine care. The gap between demonstrating improvements in model performance and impact through actual implementation remains large. This is an area that the health services and outcomes research community must help navigate in the future.

In summary, Ng et al[3] have described a rigorous path to incorporating machine learning into EHR-based research and understanding how decisions regarding data amount and type influence model performance within a real-world EHR. It will not be an easy road to success, but it is rewarding to see these early steps toward learning more about machine learning because these will help pave the way toward fulfilling its ultimate promise for clinical practice.

Disclosures

None.

References

1. Echouffo-Tcheugui JB, Greene SJ, Papadimitriou L, Zannad F, Yancy CW, Gheorghiade M, Butler J. Population risk prediction models for incident heart failure: a systematic review. *Circ Heart Fail*. 2015;8:438–447. doi: 10.1161/CIRCHEARTFAILURE.114.001896.
2. Deo RC. Machine learning in medicine. *Circulation*. 2015;132:1920–1930. doi: 10.1161/CIRCULATIONAHA.115.001593.

3. Ng K, Steinhubl SR, deFilippi C, Dey S, Stewart WF. Early detection of heart failure using electronic health records: practical implications for time before diagnosis, data diversity, data quantity, and data density. *Circ Cardiovasc Qual Outcomes*. 2016;9:649–658. doi: 10.1161/CIRCOUTCOMES.116.002797.
4. Velagaleti RS, Gona P, Larson MG, Wang TJ, Levy D, Benjamin EJ, Selhub J, Jacques PF, Meigs JB, Tofler GH, Vasan RS. Multimarker approach for the prediction of heart failure incidence in the community. *Circulation*. 2010;122:1700–1706. doi: 10.1161/CIRCULATIONAHA.109.929661.
5. Goyal A, Norton CR, Thomas TN, Davis RL, Butler J, Ashok V, Zhao L, Vaccarino V, Wilson PW. Predictors of incident heart failure in a large insured population: a one million person-year follow-up study. *Circ Heart Fail*. 2010;3:698–705. doi: 10.1161/CIRCHEARTFAILURE.110.938175.

Key Words: Editorials ■ heart failure ■ linear models ■ machine learning ■ medicine ■ risk factors

Circulation
Cardiovascular Quality and Outcomes

Learning About Machine Learning: The Promise and Pitfalls of Big Data and the Electronic Health Record

Rahul C. Deo and Brahmajee K. Nallamothu

Circ Cardiovasc Qual Outcomes. 2016;9:618-620; originally published online November 8, 2016;
doi: 10.1161/CIRCOUTCOMES.116.003308
Circulation: Cardiovascular Quality and Outcomes is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231

Print ISSN: 1941-7705. Online ISSN: 1941-7713