Research Article

A Systematic Approach of Machine Learning for Healthcare: A Review

P. Sripalreddy[1*], Rashi Agrawal[2]

Abstract:

This review paper systematically examines the application of machine learning (ML) in the healthcare domain, aiming to provide a comprehensive overview of the current state, challenges, and future prospects. As the healthcare industry undergoes a transformative shift towards data-driven decision-making, ML techniques have emerged as powerful tools for extracting valuable insights from complex medical datasets.
The review adopts a systematic approach, categorizing the diverse applications of ML in healthcare into key themes.

Keywords: EHR data, Machine Learning Models, trade-off of embedding a model in a healthcare workflow; Challenges in Machine Learnings.

1.INTRODUCTION

Studying patient populations to identify causes, risk factors, effective treat- ments, and subtypes of disease has long been the purview of epidemiology (Dicker et al., 2006). Epidemiological methods such as randomized con- trolled trials and case-control studies are the cornerstones of evidence-based medicine. However, such methods are time-consuming and expensive, maynot be free of the biases they are designed to combat (Delgado-Rodr´ıguezand Llorca, 2004), and their results may not be applicable to real-worldpatient populations (Weng et al., 2014). Such studies are also difficult to exe- cute over large patient populations because of restrictive inclusion/exclusion criteria and the operational challenges of running large prospective studiesthat follow patients over long periods of time. Internationally, the adoption of electronic health records (EHRs) is increasing due to strategies and agen- cies that incentivize their use such as the Health Information for Economic and Clinical Health Act in the United States (Maxson et al., 2010), the National Agency for Health IT in Denmark, and the National eHealth Authority in India (Mossialos et al., 2016). As a result, methods that leverage EHRs to answer questions tackled by epidemiologists (Casey et al., 2016) and to increase precision in healthcare delivery are now commonplace (Parikh et al., 2016).
Data analysis approaches broadly fall into the following categories: descriptive, exploratory, inferential, predictive, and causal (Leek and Peng, 2015). A descriptive analysis reports summaries of data without interpreta- tion and an exploratory analysis identifies associations between variables ina dataset. An inferential study quantifies the degree to which an observed association in a population will hold outside the dataset from which it was derived, and a predictive study attempts to quantify the probability of an out- come at the level of an individual. Finally, a causal analysis determines how changes in one variable affect another. It is crucial to define the type of question being asked in a given study to determine the type of data analysis that is appropriate to use in answering the question.
Inferential analyses are widely used in clinical research. The primary goal of such studies is to

---

[1]Research Scholar, Department of Computer Science Engineering, Chhatrapati Shahu Ji Maharaj University, Kanpur. Email: sripal.patlola@gmail.com
[2]Research guide, Department of Computer Science Engineering, Chhatrapati Shahu Ji Maharaj University, Kanpur. Email: dr.rashiagrawal@gmail.com
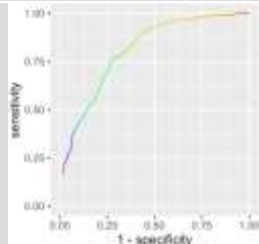
quantify how a change in one or more variables affects an observed outcome at a population level (Leek and Peng, 2015).Such analysis (referred to as the culture of data modeling by Breiman, 2001) assumes the existence of a model that produces observed outcomes frominput variables and evaluates a model's quality by measuring its goodness of fit to existing data. A model that has a good fit to existing data is considered adequately to define the process by which data are generated.

Predictive analyses, on the other hand, aim to predict outcomes for indi- viduals (Leek and Peng, 2015) by constructing a statistical model from observed data and using this model to generate a prediction for an individual based on their unique features. Predictive modeling is a type of algorithmic modeling (Breiman, 2001), which considers the process by which data are generated to be unknown (and perhaps unknowable). Such modelingapproaches measure performance by metrics such as precision, recall, and calibration (Table 19.1), which quantify different notions of the frequency with which a model's predictions are correct.

Machine learning is the process of learning a good enough statisticalmodel using observed data to predict outcomes or categorize observations in future data. Specifically, supervised machine learning methods train a model using observations on samples where the categories or predicted value of the outcome of interest are already known (a gold standard). The resulting model—which is often a penalized regression of some form—is typically applied to new samples to categorize or predict values of the outcome for previously unseen observations, and its performance evaluated by comparing predicted values to actual values for a set of test samples. Therefore, machine learning "lives" in the world of algorithmic modeling and should be evaluated as such. Regression models developed using machine learning methods cannot and should not be evaluated using criteria from the world of data modeling (Breiman, 2001). To do so would produce inaccurate assess- ments of a model's performance for its intended task, potentially misleading users into incorrect interpretation of the model's output.

EHRs provide access to a large number and variety of variables thatenable high-quality classification and prediction (Kennedy et al., 2013), while machine learning offers the methods to handle the large volumes of high-dimensional data that are typical in a healthcare setting. As a result, the application of machine learning to EHR data analysis is at the forefront of modern clinical informatics (Goldstein et al., 2016), fueling advances in both the science and practice of medicine. In the following sections, we presentan overview of how machine learning has been applied in clinical settings and summarize the advantages it offers over traditional analysis methods and caveats when using machine learning in real-world settings. We describe the

TABLE 1 Common Performance Measures Used to Evaluate MachineLearning Models

Performance Measure	Definition	Formula or Plot
Accuracy	The number of correct classifications made by a model (true positives and true negatives) divided bythe total number of predictions made	$A 5 \dfrac{TP\ 1\ TN\ TP}{1\ FP\ 1\ TN\ 1\ FN}$
Calibration	A measure of how closely predicted probabilities for anoutcome match the observedoutcome in test data, e.g., the Brier score	Brier score $5 \dfrac{1}{N} \sum_{i51}^{N} ðp_i\ 2\ o_iÞ$
Discrimination	A measure of how well a model discriminates between randomly selected true positive cases and true negative cases, usually measured as the area under the receiver operator curve (AUC)	

Performance Measure	Definition	Formula or Plot
Negative predictive value	The total number of correct negative classifications made (true negatives) divided by the total number of negative classifications made (true negatives and false negatives)	$NPV = \dfrac{TN}{TN + FN}$
Precision (also called positive predictive value)	The total number of correct positive classifications made (true positives) divided by the total number of positive classifications made (true positives and false positives)	$P \text{ or } PPV = \dfrac{TP}{TP + FP}$
Recall (also called sensitivity or the true positive rate)	The total number of correct positive classifications made (true positives) divided by the number of positive class members in the data (true positives and false negatives)	$R = \dfrac{TP}{TP + FN}$

(Continued)

TABLE .1 (Continued)

Performance Measure	Definition	Formula or Plot
Specificity (also called the true negative rate)	The total number of correct negative classifications made (true negatives) divided by the number of negative class members in the data (true negatives and false positives)	$S = \dfrac{TN}{TN + FP}$

For a binary classifier that classifies data points to be a member of either a positive class or a negative class, the following definitions are used in the measure formulas: True positive (TP)— a correct classification that a data point is a member of the positive class. True negative (TN)—a correct classification that a data point is a member of the negative class. False positive (FP)—an incorrect classification that a data point is a member of the positive class. False negative (FN)—an incorrect classification that a data point is a member of the negative class. For a classifier that predicts the probability of an outcome, p_i is the predicted probability of the outcome for data point i, and o_i is the whether the outcome was observed for that data point or not.

Methodological and operational challenges of using machine learning in research and practice. Lastly, we offer our perspective on opportunities for machine learning in medicine and applications that have the highest potential for impacting health and healthcare delivery.

2. FROM SCORING SYSTEMS TO STATISTICAL MODELS: LITERATURE REVIEW

In medicine, the use of scoring systems to categorize patients into different risk strata is quite common. For example, the Charlson Comorbidity Index (CCI) (Charlson et al., 1987) quantifies an individual's burden of disease and corresponding 1-year mortality risk. The widely used APACHE score (Haddad et al., 2008) quantifies the severity of disease for patients admitted to intensive care units. The Apgar score summarizes physiological measures as an indicator of infant morbidity (Casey et al., 2001). A recent entrant to risk scoring systems is the Rothman Index (Finlay et al., 2013), which is a score that quantifies risk of death based on vital signs, nursing assessments, and laboratory results. Disease-specific scoring systems have also been developed, such as the Dutch Lipid Clinic Network (DLCN) criteria for diag- nosing "unlikely," "possible," "probable," or "definite" familial hypercholes- terolemia (FH) (European Association for Cardiovascular Prevention and Rehabilitation et al., 2011).

Scoring systems have the advantage of being easily interpretable and are usually straightforward to administer, but may not perform well at accurately reflecting the risk for individuals. For example, the DLCN criteria for FH have specificity ranging from 5.9% to 89.4% and sensitivity ranging from 41.5% to 99.3% for predicting the results of a genetic test for the disease (Damgaard et al., 2005). However, due to the association between age and several features of "definite" FH, the scores are more likely to misclassify younger individuals (Besseling et al., 2016). Similarly, the APACHE scores were found to have excellent discrimination, but poor calibration (Haddad et al., 2008) meaning that while the APACHE score could separate the high- risk group from the low-risk group, the outcome probabilities for specific individuals were inaccurate. This has implications for deciding how to use such scores in practice. For example, the excellent discrimination of the APACHE score makes it highly useful to decide whether to transfer a patient out of an ICU or not. However, it is inappropriate to use the score to quan-tify that an individual patient will need to be resuscitated with a 90% probability.

Given such limitations of traditional scoring systems, approaches that use machine learning to build well-calibrated statistical models that correctly classify individuals and provide accurate estimates of risk are growing in popularity. We have, for example, recently developed a statistical model—a penalized logistic regression model—to classify patients as "unlikely," "pos- sible," "probable," or "definite" FH based on the same criteria as the DLCN, but learning the variable weights from EHR data. Our model augments the DLCN criteria with variables that quantify health system utilization and sum- mary statistics of cholesterol levels and achieves an AUC of 95% with 82% precision and 86% recall. Other approaches have yielded comparable results for predicting FH (Weng et al., 2015; Besseling et al., 2016) and have signif- icant potential cost savings by reducing unnecessary genetic testing and simultaneously ensuring that true cases are not missed.

Using a similar approach, researchers in Canada updated and reevaluated the CCI with more recent administrative data and larger more diverse patient populations, and without relying on chart review (Quan et al., 2011). Rather than replacing the CCI altogether with a statistical model, they used statisti-cal models to assign new weights to comorbidities used in the CCI. Doing so resulted in a decrease in the number of conditions included in the Index from 17 to 12. This simpler index achieved comparable performance to the CCI, based on evaluation with data from six economically developed countries. Other risk models learned using EHR data include PhysiScore (Saria et al., 2010), a system that predicts preterm infant morbidity using similar physio- logical features as the Apgar score, and TrewScore (Henry et al., 2015), an early warning system for sepsis. Both achieved improved performance com- pared to traditional point-based scoring systems. Numerous other examples exist for models that predict sepsis (Gultepe et al., 2014), delayed wound healing (Jung et al., 2016), risk of being depressed (Huang et al., 2014), and cardiovascular adverse events (Ross et al., 2016).

These studies demonstrate that using machine learning approaches to do the "heavy lifting" of deriving appropriate variable weights (and in many cases, performing automated variable selection) can produce a model that outperforms the corresponding expert-derived scoring system and can update existing scoring models as data change over time (Jung and Shah, 2015). Proper use of machine learning methods that can handle large numbers of sparsely measured variables and that can identify variables with the most sig- nificant effect on model performance are crucial to the success of models built from heterogeneous and often messy EHR data.

3. CHALLENGES IN BUILDING AND DEPLOYINGEHR-BASED MODELS

The challenges inherent in retrospectively analyzing EHR data collected for administrative and clinical purposes to do research have been well character- ized (Hersh et al., 2013). They include biases due to using data from indivi- duals generated only when they visit their healthcare provider (and are therefore likely ill), incomplete records (and the related issue of missing data violating the missing-at-random assumption), loss to follow-up, the modifica- tion of records for billing purposes,

and potential errors introduced to a patient's record as they move through a healthcare system such as inaccurate coding or data entry (Table 19.2). Statistical methods commonly employed to address these difficulties include imputation to fill in missing data and propensity score based matching, weighting or stratification to select patient samples as controls (Schneeweiss et al., 2009; Brookhart et al., 2010; Toh et al., 2011; Patorno et al., 2013). These challenges and biases that affect retrospective studies also impact the use of machine learning to build models based on EHR data. The effectiveness of methods that minimize biases in retrospective studies when applied to high-dimensional EHR data is an area of active research.

A recent systematic review of published research on developing predic- tive models using EHR data (Goldstein et al., 2016) found that most predic- tive models developed for healthcare applications did not take advantage of the wealth of EHR data available. For example, only about half of the 60 studies that predicted a clinical outcome used information beyond diagnosis codes to define the outcome, and only eight studies used longitudinal mea- surements. The task of accurately ascertaining whether a specific outcome occurred (or not) is called electronic phenotyping (Richesson et al., 2016) and is the most important step in any statistical modeling exercise done with EHR data. A related study (Wei et al., 2016) found that in 4 out of 10 medi- cal conditions examined, the medical records of less than 10% of sampled patients known to have the condition contained the International Classification of Diseases diagnosis codes necessary to identify the condi- tion. For the remainder, data derived from clinical notes or medication records (just two of many possible EHR-derived sources) were required to determine the diagnosis. In light of the importance of accurate phenotype determination—both for the outcome as well as for the presence of comorbidities—several research efforts are focused on employing machine learning methods for the purpose of phenotyping itself (Agarwal et al., 2016; Halpern et al., 2016).

TABLE 2 Types of Errors in EHR and Claims Data, and Possible Causes

Type of Error	Definition	Possible Causes
Missing data	A patient's record is incomplete and has gaps in time when they received healthcare but it was not recorded	– Patient does not seek care because they lack healthcare insurance coverage – Loss to follow-up—a patient leaves their insurer or stops seeing their healthcare provider – Lag time or mistake in insurance claim processing – Error when populating insurance claims database – Incomplete record linkage—records belonging to a single patient are not linked
Missing data not missing at random	The reason data is missing from a patient's record is related to unobserved environmental or patient-specific factors	– Patient does not seek healthcare because they are very sick or healthy – Healthcare provider misdiagnoses a patient, makes an error when writing a note for a patient visit, or does not include all information in their note – Hospital/clinic staff miscodes diagnoses and/or medical procedures, or does not file insurance claim
Outcome or exposure misclassification	Incorrect labeling of patients who have a disease or do not, or who had an exposure (e.g., a medical treatment, a risk factor for a disease) or did not	

Another significant issue faced by researchers using machine learning to derive insights from EHR

data is that of external validity or generalizability (Iezzoni, 1999)—the performance of learned models at sites other than thethat which sourced the data used for training. One prevailing opinion is that models not validated on data from an external site (or sites) are not useful,and that models lacking external validity have failed in their task. Such thinking affects model design decisions and evaluation; for example, a model with fewer variables is often considered more generalizable, and therefore more useable, than one with more variables. We posit that one of the benefits of machine learning is that a model can be trained with data from any localsite and evaluated using data from that site itself. Therefore, if a modelachieves sufficient performance at a local site then it has achieved its pur- pose. So the discussion should be about sharing the model building workflow to retrain a model at a new site, rather than about "external validity." Indeed, learning a model using site-specific data and updating its parameters as a dataset changes over time will produce a model that is best suited for predic- tion at that site and with that data (Lee and Maslove, 2015). The success of such an approach across multiple sites (Peck et al., 2013) is evidence that the model building approach is valid.

The success of building and deploying a statistical model at a new site rests on the availability and consistent representation of data across sites. The Observational Health Data Sciences and Informatics (OHDSI) networkis an international collaboration of observational health researchers that has developed a Common Data Model (CDM) for EHR data currently used to structure hundreds of millions of patient records from more than 50 sites in12 countries. A query to a CDM-compliant database to construct a cohort at one site can be deployed at any other OHDSI site, without modification. Such portability has already facilitated large-scale research into treatmentpathways (Hripcsak et al., 2016) and presents the opportunity to develop sta- tistical models that are portable across sites. Researchers can share the work- flow used to construct and train a model, such that it can be retrained using data from a new site and deployed locally.

In addition to technical challenges around data models, external validity, and the need to use multiple data types, operational challenges also exist for the construction and deployment of a statistical model in a clinical setting. These challenges include data access, availability of personnel with the nec- essary skills to develop a model and interpret its output, and, perhaps most essential, a system for integrating the model into the healthcare practitioner's workflow it is designed to inform. For effective integration into the clinical workflow it is essential that the model's output be actionable, that is, that there are concrete interventions that can be executed in response to the model output. For example, if a model predicts a high risk of delayed wound healing, a physician can opt to begin hyperbaric oxygen therapy early (Junget al., 2016). The availability of an effective intervention is a key consider- ation in assessing the utility of a predictive model built using EHR data.

The success of models deployed for practice management therefore depends on more than just model performance—additional factors such as the cost of deploying the model and the effectiveness of any action triggered by the model are also necessary to consider. For example, factors to consider include the size of the patient population where intervention is possible, the prevalence of the outcome being predicted, the expected number of patients with the predicted outcome, the cost savings per patient exposed to the intervention, the estimated cost of implementing the intervention, the total target savings, and the number of patients subjected to the intervention required to meet that target. If deploying a model is expensive, or the affected population too small to result in cost savings for the practice in question, the model will not be "useful" even if its performance as measuredin terms of precision, recall and calibration is exceptional.

4.OPPORTUNITIES FOR MACHINE LEARNING IN HEALTHCARE

Despite the methodological challenges of working with EHR data and that researchers have yet to take full advantage of the universe of EHR-derived variables available for predictive modeling, there are many exciting opportunities for machine learning to improve health and healthcare delivery. Models that stratify patients into different risk categories to inform practice management have enormous potential impact on healthcare value (see Chapter 8: Health Information Technology and Value) (Parikh et al., 2016), and methods thatcan predict outcomes for individual patients bring

clinical practice one step closer to precision medicine (see Chapter 11: Bioinformatics and Precision Medicine) (Pencina and Peterson, 2016). Identifying high-cost and high-risk patients (Bates et al., 2014) in time to attempt targeted intervention will become increasingly necessary as healthcare providers take on the financial risk of treating their patients.

Machine learning approaches have already been used to characterize and predict a variety of health risks. Recent work in our group using penalized logistic regression to identify patients with undiagnosed peripheral artery dis- ease and predict their mortality risk found that such an approach outperforms a simpler stepwise logistic regression in terms of accuracy, calibration, and net reclassification (Ross et al., 2016). Such predictive models have been implemented in medical practice, resulting in more efficient and better qual- ity care. For example, a predictive model to stratify newborns' risk of sepsis decreased antibiotic prescribing by 33% 60% when deployed by Kaiser Permanente (Escobar et al., 2014). Recent work to learn reference ranges for pediatric heart and respiratory rates from EHR data resulted in a decrease in heart rate alarms in a pediatric acute care unit following implementation of the learned ranges in unit monitors (Goel et al., 2015), potentially reducing alarm fatigue.

Machine learning has also been applied to hospital and practice manage- ment, to streamline operations and improve patient outcomes. For example, models have been developed to predict demand for emergency department beds (Peck et al., 2012) and elective surgery case volume (Tiwari et al., 2014), to inform hospital staffing decisions. The Veterans Health Administration's Clinical Data Warehouse, which houses clinical data for more than 20 million patients, was used to develop models for risk of hospitalization and death of individual patients with AUC values ranging from 0.81 to 0.87 (Fihn et al., 2014). The risk scores calculated by these models are presented to Patient-Aligned Care Teams and are accessed by more than 1200 healthcare providers per month as part of their routine prac- tice. In another successful example, a 30-day readmission risk score for heart failure patients at the Parkland Health and Hospital System, called e-Score, was implemented in a prospective study to assess the effectiveness of a tar- geted intervention (involving intensive follow-up care) in patients with a high predicted risk of readmission (Amarasingham et al., 2013). The study found that the intervention, triggered by the predicted risk score, resulted in significantly reduced 30-day readmission compared to a control group. These examples demonstrate that predictive models can be deployed in healthcare settings and used effectively to improve resource allocation and patient care.

Finally, machine learning is not a panacea, and not all things that can be predicted will be actionable. For example, we may be able to accurately pre- dict progression from stage 3 to stage 4 chronic renal failure. In the absence of effective treatment options—besides dialysis and kidney transplant—the prediction does not do much to improve the management of the patient. As another extreme example, imagine a model that predicts that treatment of a particular cancer (say lung cancer) will be ineffective based on the genotypic profile of the tumor. Would we withhold treatment based on such a predic- tion? Such extreme situations aside, there is ample middle ground where the use of predictive modeling in healthcare will enable proactive treatment, more efficient use of resources, and deliver better care at lower cost. Multiple other industries—such as finance, retail, aviation, web commerce, and election campaigning—have made the transition to incorporate machine learning, and the resulting predictive models, into their respective workflows. It is time for healthcare to make that transition.

5. CONCLUSIONS

Machine learning techniques applied to EHR data can generate actionable insights, from improving upon patient risk score systems, to predicting the onset of disease, to streamlining hospital operations. Statistical models that leverage the variety and richness of EHR-derived data (as opposed to using a small set of expert-selected and/or traditionally used features) are still rela- tively rare and offer an exciting avenue for further research. New types of data, such as from wearables, bring their own opportunities and challenges. Challenges in effectively using machine learning methods include the avail- ability of personnel with the skills to build, evaluate, and apply learned mod- els, as well as the

assessing the real-world cost benefit trade-off of embedding a model in a healthcare workflow. As machine learning usingEHR data touches a large number of healthcare problems, it will motivate applied research in statistics and computer science. Best practices informed by the success (and failure) of machine learning models will move clinical informatics forward as a field and continue to transform medicine and the delivery of clinical care.

REFERENCES

1. Agarwal, V., Vibhu, A., Tanya, P., Banda, J.M., Veena, G., Leung, T.I., et al., 2016. Learning statistical models of phenotypes using noisy labeled training data. J. Am. Med. Inform. Assoc. 23, 1166—1173.
2. Amarasingham, R., Patel, P.C., Toto, K., Nelson, L.L., Swanson, T.S., Moore, B.J., et al., 2013. Allocating scarce resources in real-time to reduce heart failure readmissions: a prospective, controlled study. BMJ Qual. Saf. 22, 998—1005.
3. Bates, D.W., Saria, S., Ohno-Machado, L., Shah, A., Escobar, G., 2014. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff. 33, 1123—1131.
4. Besseling, J., Reitsma, J.B., Gaudet, D., Brisson, D., Kastelein, J.J.P., Hovingh, G.K., et al., 2016. Selection of individuals for genetic testing for familial hypercholesterolaemia: development and external validation of a prediction model for the presence of a mutation causing familial hypercholesterolaemia. Eur. Heart J. 38 (8), 565—573.
5. Breiman, L., 2001. Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat. Sci. 16, 199—231.
6. Brookhart, M.A., Stürmer, T., Glynn, R.J., Rassen, J., Schneeweiss, S., 2010. Confounding control in healthcare database research: challenges and potential approaches. Med. Care 48, S114—S120.
7. Casey, B.M., McIntire, D.D., Leveno, K.J., 2001. The continuing value of the Apgar score for the assessment of newborn infants. N. Engl. J. Med. 344, 467—471.
8. Casey, J.A., Schwartz, B.S., Stewart, W.F., Adler, N.E., 2016. Using electronic health records for population health research: a review of methods and applications. Annu. Rev. Public Health 37, 61—81.
9. Charlson, M.E., Pompei, P., Ales, K.L., MacKenzie, C.R., 1987. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J. Chronic Dis. 40, 373—383.
10. Damgaard, D., Larsen, M.L., Nissen, P.H., Jensen, J.M., Jensen, H.K., Soerensen, V.R., et al., 2005. The relationship of molecular genetic to clinical diagnosis of familial hypercholesterolemia in a Danish population. Atherosclerosis 180, 155—160.
11. Delgado-Rodr´ıguez, M., Llorca, J., 2004. Bias. J. Epidemiol. Community Health 58, 635—641.
12. Dicker, R., Coronado, F., Koo, D., Parrish, R.G., 2006. Principles of Epidemiology in Public Health Practice. US Department of Health and Human Services, Atlanta, GA.
13. Escobar, G.J., Puopolo, K.M., Wi, S., Turk, B.J., Kuzniewicz, M.W., Walsh, E.M., et al., 2014. Stratification of risk of early-onset sepsis in newborns $ 34 weeks' gestation. Pediatrics 133, 30—36.
14. European Association for Cardiovascular Prevention and Rehabilitation, Reiner, Z., Catapano, A. L., De Backer, G., Graham, I., Taskinen, M.-R., et al., 2011. ESC/EAS Guidelines for the management of dyslipidaemias: the Task Force for the management of dyslipidaemias of the European Society of Cardiology (ESC) and the European Atherosclerosis Society (EAS).Eur. Heart J. 32, 1769—1818.
15. Fihn, S.D., Francis, J., Clancy, C., Nielson, C., Nelson, K., Rumsfeld, J., et al., 2014. Insights from advanced analytics at the Veterans Health Administration. Health Aff. 33, 1203—1211. Finlay, G.D., Duncan Finlay, G., Rothman, M.J., Smith, R.A., 2013. Measuring the modified early warning score and the Rothman Index: advantages of utilizing the electronic medical
16. record in an early warning system. J. Hosp. Med. 9, 116—119.

17. Goel, V., Poole, S., Kipps, A., Palma, J., Platchek, T., Pageler, N., et al., 2015. Implementation of data driven heart rate and respiratory rate parameters on a pediatric acute care unit. Stud. Health Technol. Inform. 216, 918.
18. Goldstein, B.A., Navar, A.M., Pencina, M.J., Ioannidis, J.P., 2016. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J. Am. Med. Inform. Assoc. 27 (1), 198—208.
19. Gultepe, E., Green, J.P., Nguyen, H., Adams, J., Albertson, T., Tagkopoulos, I., 2014. From vital signs to clinical outcomes for patients with sepsis: a machine learning basis for a clinical decision support system. J. Am. Med. Inform. Assoc. 21, 315—325.
20. Haddad, Z., Falissard, B.F., Chokri, K.C., Kamel, B.K., Nader, B.N., Nagi, S.N., et al., 2008. Disparity in outcome prediction between APACHE II, APACHE III and APACHE IV. Crit. Care 12, P501.
21. Halpern, Y., Yoni, H., Steven, H., Youngduck, C., David, S., 2016. Electronic medical record phenotyping using the anchor and learn framework. J. Am. Med. Inform. Assoc. 23, 731—740.
22. Henry, K.E., Hager, D.N., Pronovost, P.J., Saria, S., 2015. A targeted real-time early warning score (TREWScore) for septic shock. Sci. Transl. Med. 7, 299ra122.
23. Hersh, W.R., Weiner, M.G., Embi, P.J., Logan, J.R., Payne, P.R.O., Bernstam, E.V., et al., 2013. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med. Care 51, S30—S37.
24. Hripcsak, G., Ryan, P.B., Duke, J.D., Shah, N.H., Park, R.W., Huser, V., et al., 2016. Characterizing treatment pathways at scale using the OHDSI network. Proc. Natl. Acad. Sci. U. S. A. 113, 7329—7336.
25. Huang, S.H., Paea, L., Iyer, S.V., Ming, T.-S., David, C., Shah, N.H., 2014. Toward personalizing treatment for depression: predicting diagnosis and severity. J. Am. Med. Inform. Assoc. 21, 1069—1075.
26. Iezzoni, L.I., 1999. Statistically derived predictive models. Caveat emptor. J. Gen. Intern. Med.
27. 14, 388—389.
28. Jung, K., Shah, N.H., 2015. Implications of non-stationarity on predictive modeling using EHRs.
29. J. Biomed. Inform. 58, 168—174.
30. Jung, K., Covington, S., Sen, C.K., Januszyk, M., Kirsner, R.S., Gurtner, G.C., et al., 2016.
31. Rapid identification of slow healing wounds. Wound Repair Regen. 24, 181—188.
32. Kennedy, E.H., Wiitala, W.L., Hayward, R.A., Sussman, J.B., 2013. Improved cardiovascular risk prediction using nonparametric regression and electronic health record data. Med. Care 51, 251—258.
33. Lee, J., Maslove, D.M., 2015. Customization of a severity of illness score using local electronic medical record data. J. Intensive Care Med. 32 (1), 38—47.
34. Leek, J.T., Peng, R.D., 2015. Statistics. What is the question? Science 347, 1314—1315.
35. Maxson, E., Jain, S., Kendall, M., Mostashari, F., Blumenthal, D., 2010. The regional extension center program: helping physicians meaningfully use health information technology. Ann. Intern. Med. 153, 666—670.
36. Mossialos, E., Wenzl, M., Osborn, R., Sarnak, D., 2016. 2015 International Profiles of Health Care Systems. The Commonwealth Fund.
37. Parikh, R.B., Kakad, M., Bates, D.W., 2016. Integrating predictive analytics into high-value care: the dawn of precision delivery. JAMA 315, 651—652.
38. Patorno, E., Grotta, A., Bellocco, R., Schneeweiss, S., 2013. Propensity score methodology for confounding control in health care utilization databases. Epidemiol. Biostat. Public Health 10. Peck, J.S., Benneyan, J.C., Nightingale, D.J., Gaehde, S.A., 2012. Predicting emergency department inpatient admissions to improve same-day patient flow. Acad. Emerg. Med. 19,
39. E1045—E1054.
40. Peck, J.S., Gaehde, S.A., Nightingale, D.J., Gelman, D.Y., Huckins, D.S., Lemons, M.F., et al., 2013. Generalizability of a simple approach for predicting hospital admission from an emergency department. Acad. Emerg. Med. 20, 1156—1163.
41. Pencina, M.J., Peterson, E.D., 2016. Moving from clinical trials to precision medicine: the role

for predictive modeling. JAMA 315, 1713—1714.

42. Quan, H., Li, B., Couris, C.M., Fushimi, K., Graham, P., Hider, P., et al., 2011. Updating and validating the Charlson comorbidity index and score for risk adjustment in hospital dis- charge abstracts using data from 6 countries. Am. J. Epidemiol. 173, 676—682.

43. Richesson, R.L., Jimeng, S., Jyotishman, P., Abel, K., Denny, J.C., 2016. A survey of clinical phenotyping in selected national networks: demonstrating the need for high-throughput, portable, and computational methods. Artif. Intell. Med. 71, 57—61.

44. Ross, E.G., Shah, N.H., Dalman, R.L., Nead, K., Cooke, J., Leeper, N.J., 2016. The use of machine learning for the identification of peripheral artery disease and future mortality risk. J. Vasc. Surg. 64, 1515—1522.e3.

45. Saria, S., Rajani, A.K., Gould, J., Koller, D., Penn, A.A., 2010. Integration of early physiological responses predicts later illness severity in preterm infants. Sci. Transl. Med. 2, 48ra65.

46. Schneeweiss, S., Rassen, J.A., Glynn, R.J., Avorn, J., Mogun, H., Brookhart, M.A., 2009. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiology 20, 512—522.

47. Tiwari, V., Furman, W.R., Sandberg, W.S., 2014. Predicting case volume from the accumulating elective operating room schedule facilitates staffing improvements. Anesthesiology 121, 171—183.

48. Toh, S., García Rodríguez, L.A., Hernán, M.A., 2011. Confounding adjustment via a semi-
49. automated high-dimensional propensity score algorithm: an application to electronic medical records. Pharmacoepidemiol. Drug Saf. 20, 849—857.

50. Wei, W.-Q., Teixeira, P.L., Mo, H., Cronin, R.M., Warner, J.L., Denny, J.C., 2016. Combining billing codes, clinical notes, and medications from electronic health records provides super- ior phenotyping performance. J. Am. Med. Inform. Assoc. 23, e20—e27.

51. Weng, C., Li, Y., Ryan, P., Zhang, Y., Liu, F., Gao, J., et al., 2014. A distribution-based method for assessing the differences between clinical trial target populations and patient populations in electronic health records. Appl. Clin. Inform. 5, 463—479.

52. Weng, S.F., Kai, J., Andrew Neil, H., Humphries, S.E., Qureshi, N., 2015. Improving identification of familial hypercholesterolaemia in primary care: derivation and validation of the familial hypercholesterolaemia case ascertainment tool (FAMCAT). Atherosclerosis 238, 336—343.