Review

Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review

Benjamin A Goldstein,1,2 Ann Marie Navar,2,3 Michael J Pencina,1,2 and John P A Ioannidis4,5

1Department of Biostatistics & Bioinformatics, Duke University, Durham, NC 27710, USA, 2Center for Predictive Medicine, Duke Clinical Research Institute, Duke University, Durham, NC 27710, USA, 3Division of Cardiology at Duke University Medical Center, Durham, NC 27710, USA, 4Department of Medicine, Stanford University, Palo Alto, CA 94305, USA, and 5Department of Health Research and Policy, and Statistics and Meta-Research Innovation Center at Stanford, Stanford University, Palo Alto, CA 94305, USA

Correspondence to Benjamin Goldstein, Department of Biostatistics & Bioinformatics, Duke University, 2424 Erwin Road, Suite 1104, Durham, NC 27705, USA; ben.goldstein@duke.edu.

Received 27 October 2015; Revised 25 January 2016; Accepted 20 February 2016

ABSTRACT

Objective: Electronic health records (EHRs) are an increasingly common data source for clinical risk prediction, presenting both unique analytic opportunities and challenges. We sought to evaluate the current state of EHR based risk prediction modeling through a systematic review of clinical prediction studies using EHR data.

Methods: We searched PubMed for articles that reported on the use of an EHR to develop a risk prediction model from 2009 to 2014. Articles were extracted by two reviewers, and we abstracted information on study design, use of EHR data, model building, and performance from each publication and supplementary documentation.

Results: We identified 107 articles from 15 different countries. Studies were generally very large (median sample size = 26 100) and utilized a diverse array of predictors. Most used validation techniques (n = 94 of 107) and reported model coefficients for reproducibility (n = 83). However, studies did not fully leverage the breadth of EHR data, as they uncommonly used longitudinal information (n = 37) and employed relatively few predictor variables (median = 27 variables). Less than half of the studies were multicenter (n = 50) and only 26 performed validation across sites. Many studies did not fully address biases of EHR data such as missing data or loss to follow-up. Average c-statistics for different outcomes were: mortality (0.84), clinical prediction (0.83), hospitalization (0.71), and service utilization (0.71).

Conclusions: EHR data present both opportunities and challenges for clinical risk prediction. There is room for improvement in designing such studies.

Key words: Electronic Medical Record; Review; Risk Assessment

INTRODUCTION

The use of electronic health records (EHRs) has increased dramatically in the past 5 years. In 2009, 12.2% of US hospitals had a basic EHR system, increasing to 75.5% by 2014.2 Beyond facilitating billing and patient care, the dynamic clinical patient information captured in structured EHRs provides opportunities for research, including developing and refining risk prediction algorithms.2
EHR-based risk prediction studies depart from traditional risk prediction studies in several significant ways. Traditionally, risk prediction algorithms have been developed from large cohort studies such as the Framingham Heart Study. These studies were designed to follow people for years or even decades. As such, they have predefined inclusion criterion, regular follow-up of participants, specified metrics to collect, and protocols for adjudicating outcomes. Unlike cohort data collected for research purposes, EHR data are collected de-facto, more frequently, and may lack the same standardization as cohort studies. As others have noted, EHR data come with many challenges. EHRs include all patients that touch a medical system, primarily capture data only when patients are ill, and collect metrics that clinicians deem to be necessary at each clinic visit. The data tend to be very “messy” leading to many potential analytic challenges and biases. Further, EHR-based outcomes and diagnoses vary based on how they are defined and from what data (ie, billing codes, medical problem lists, etc.) they are derived. However, there are multiple advantages to EHR-based risk prediction. Such de-facto data collection allows one to observe more metrics, on more individuals, at more time points, and at a fraction of the cost of prospective cohort studies. One can use the same set of data to predict a wide range of clinical outcomes – something not possible in most cohort studies. As data are sometimes observed with greater frequency (as opposed to yearly visits), it is also easier to predict near-term risk of events. Furthermore, patient populations derived from the EHR may be more reflective of the real-world than cohort studies that rely on volunteer participation. Finally, prediction models based on EHR data can often be readily implemented to predict near-term risk of events. Furthermore, patient populations tend to be very “messy” leading to many potential analytic challenges and biases. Further, EHR-based outcomes and diagnoses vary based on how they are defined and from what data (ie, billing codes, medical problem lists, etc.) they are derived.

However, there are multiple advantages to EHR-based risk prediction models. Multiple reviews of prognostic models have been based on EHR data as the primary source to build and validate risk prediction studies using EHR data. Ingui et al. suggested search terms for clinical prediction papers that were later updated by Geersing et al. reaching a sensitivity of 0.97. To limit the search to EHR-based studies we added the search terms: “(‘Electronic Health Record’) OR (‘Electronic Medical Record’) OR EHR OR EHRs OR EMR OR EMRs).” We limited our search to recent papers published between January 1, 2009 and December 31, 2014 (final search March 1, 2015). We anticipated that few relevant papers, if any, would have been published before 2009, since use of EHR for predictive purposes is a recent evolution. Upon review of the search results, we noticed that some studies were not identified (particularly those from the QResearch database) since they did not explicitly mention EHR (or its permutation) in the abstract. Therefore, we searched specifically for these studies as well as additional studies by research groups for which we had already identified publications in this area.

Study Selection
We included all papers (including conference proceedings) published in English that used an EHR system as a primary data source to develop a prediction algorithm to predict a clinical event or outcome. This excluded certain common types of papers such as those evaluating predictor associations, developing a “computable phenotype” definition (ie, algorithmic definition for the presence of a clinical event), applying or validating an already developed prediction algorithm, using only simulated data, or proposing a methodological approach. Two reviewers (B.A.G., A.M.N.) independently reviewed all studies, reaching a consensus on all “approved” studies.

Data Extraction, Quality Assessment, Synthesis, and Analysis
First, for each paper, we categorized the journal as being a medical (eg, Diabetes Care), informatics (eg, Journal of the American Medical Informatics Association), or health services (eg, Medical Care) journal. To determine how to evaluate papers, we consulted the recently published TRIPOD14,15 guidelines for reporting of risk prediction studies and the review by D’Agostino et al. and then focused on the study aspects that were particular or unique to EHR-based studies. We evaluated papers on three general domains: study design, leveraging of EHR data, and model development, evaluation and reporting. For study design we abstracted information on: the nature of the clinical sample, whether it was a single or multicenter study, the outcome predicted and how it was abstracted from the EHR, the study design (cohort vs case-control), and the time horizon for the outcome. For leveraging of the data we abstracted: the sample size, number of events, the number of variables considered and whether they were predefined, use of longitudinal (repeated) data, consideration of informed presence, handling of missing data, and loss to follow-up. Finally, we abstracted information on: the type of prediction model used, whether validation was performed (both internal and external), which evaluation metrics were used, how the models performed, whether individual variables were assessed, and whether a final model was reported. For studies that evaluated more than one model type (eg, 30 day vs 1 year mortality, logistic regression vs machine learning derived model) we noted which models performed best.

Data were abstracted from the abstract, main text, and any supplement material (when available). Study characteristics were counted and cross-tabulated. No specific summary measures or syntheses were calculated across studies.

METHODS
We followed the PRISMA guidelines10 for reporting our systematic review.

Data Sources and Searches
Using PubMed we performed a systematic review of clinical prediction studies using EHR data. Ingui et al. suggested search terms for clinical prediction papers that were later updated by Geersing et al. reaching a sensitivity of 0.97. To limit the search to EHR-based studies we added the search terms: “((‘Electronic Health Record’) OR (‘Electronic Medical Record’) OR EHR OR EHRs OR EMR OR EMRs).” We limited our search to recent papers published between January 1, 2009 and December 31, 2014 (final search March 1, 2015). We anticipated that few relevant papers, if any, would have been published before 2009, since use of EHR for predictive purposes is a recent evolution. Upon review of the search results, we noticed that some studies were not identified (particularly those from the QResearch database) since they did not explicitly mention EHR (or its permutation) in the abstract. Therefore, we searched specifically for these studies as well as additional studies by research groups for which we had already identified publications in this area.

RESULTS
Our initial search resulted in 8127 papers. Reviewer 1 identified 90 potentially acceptable papers while reviewer 2 identified 57. Upon comparison, 81 (7.97% (1.0%)) were eventually included, 43 of which had been identified upfront by both reviewers, in line with the positive predictive value of 1% reported by Ingui et al. on their search filter. Upon secondary search we identified 26 additional papers, 17 of which were based on the QResearch database for a total of 107 papers. Table 1 lists the 107 abstracted papers and a supplemental table has full abstraction details. While papers were rejected for a variety of reasons, the most similar class of papers were those that were using EHR data to
Table 1. Characteristics of Studies Included in the Review

<table>
<thead>
<tr>
<th>Author</th>
<th>Year</th>
<th>Location</th>
<th>Sample size (no. of events)</th>
<th>Outcome</th>
<th>Time horizon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Floder</td>
<td>2010</td>
<td>Partners Healthcare Research Patient Data Registry</td>
<td>30 341 (843)</td>
<td>Chronic obstructive pulmonary disease</td>
<td>5 years</td>
</tr>
<tr>
<td>Hipsley-Cox</td>
<td>2009</td>
<td>QResearch, UK</td>
<td>3 773 585 (115 616)</td>
<td>Diabetes</td>
<td>10 Years</td>
</tr>
<tr>
<td>Hipsley-Cox</td>
<td>2009</td>
<td>QResearch, UK</td>
<td>3 633 812 (50 755)</td>
<td>Fracture</td>
<td>10 Years</td>
</tr>
<tr>
<td>Saramesh</td>
<td>2009</td>
<td>Not reported</td>
<td>30 095 (NR)</td>
<td>Non-adherence</td>
<td>Not reported</td>
</tr>
<tr>
<td>Smith</td>
<td>2010</td>
<td>Five Health Centers, Amstered, NL</td>
<td>3045 (1859)</td>
<td>Persistent attenders</td>
<td>3–7 years</td>
</tr>
<tr>
<td>Amarasingham</td>
<td>2010</td>
<td>Mayo Clinic, MN</td>
<td>13 457 (NA)</td>
<td>Number hospital visits</td>
<td>2 years</td>
</tr>
<tr>
<td>Hipsley-Cox</td>
<td>2010</td>
<td>QResearch, UK</td>
<td>3 610 918 (205 134 app)</td>
<td>Cardiovascular disease</td>
<td>10 years</td>
</tr>
<tr>
<td>Hipsley-Cox</td>
<td>2010</td>
<td>QResearch, UK</td>
<td>3 933 092 (210 250 app)</td>
<td>Kidney disease</td>
<td>10 years</td>
</tr>
<tr>
<td>Johnson</td>
<td>2010</td>
<td>Kaiser HMO</td>
<td>5171 (145)</td>
<td>Hyperkalemia</td>
<td>1–90 days</td>
</tr>
<tr>
<td>Lipsky</td>
<td>2010</td>
<td>CareFusion</td>
<td>8747 (1021)</td>
<td>Bacteremia</td>
<td>Variable</td>
</tr>
<tr>
<td>Liu</td>
<td>2010</td>
<td>Kaiser HMO</td>
<td>155 474 (NA)</td>
<td>Length of stay</td>
<td>Variable</td>
</tr>
<tr>
<td>Hipsley-Cox</td>
<td>2010</td>
<td>CareFusion</td>
<td>26 105 (1536)</td>
<td>Acute Kidney Injury Risk and Injury</td>
<td>1–30 days</td>
</tr>
<tr>
<td>Robbins</td>
<td>2010</td>
<td>Massachusetts General and Brigham Women's Hospital</td>
<td>1074 (120)</td>
<td>Virological failure</td>
<td>1 year +</td>
</tr>
<tr>
<td>Hipsley-Cox</td>
<td>2010</td>
<td>QResearch, UK</td>
<td>3 610 918 (205 134 app)</td>
<td>Acute Kidney Injury Risk and Injury</td>
<td>Variable</td>
</tr>
<tr>
<td>Matheny</td>
<td>2010</td>
<td>Vanderbilt University Medical Center</td>
<td>5171 (145)</td>
<td>Hyperkalemia</td>
<td>Variable</td>
</tr>
<tr>
<td>Liu</td>
<td>2010</td>
<td>Kaiser HMO</td>
<td>155 474 (NA)</td>
<td>Length of stay</td>
<td>Variable</td>
</tr>
<tr>
<td>Hipsley-Cox</td>
<td>2010</td>
<td>QResearch, UK</td>
<td>3 610 918 (205 134 app)</td>
<td>Acute Kidney Injury Risk and Injury</td>
<td>Variable</td>
</tr>
<tr>
<td>Matheny</td>
<td>2010</td>
<td>Kaiser HMO</td>
<td>155 474 (NA)</td>
<td>Length of stay</td>
<td>Variable</td>
</tr>
<tr>
<td>Lipsky</td>
<td>2010</td>
<td>CareFusion</td>
<td>8747 (1021)</td>
<td>Bacteremia</td>
<td>Variable</td>
</tr>
<tr>
<td>Liu</td>
<td>2010</td>
<td>Kaiser HMO</td>
<td>155 474 (NA)</td>
<td>Length of stay</td>
<td>Variable</td>
</tr>
<tr>
<td>Hipsley-Cox</td>
<td>2010</td>
<td>QResearch, UK</td>
<td>3 610 918 (205 134 app)</td>
<td>Acute Kidney Injury Risk and Injury</td>
<td>Variable</td>
</tr>
<tr>
<td>Matheny</td>
<td>2010</td>
<td>Kaiser HMO</td>
<td>155 474 (NA)</td>
<td>Length of stay</td>
<td>Variable</td>
</tr>
<tr>
<td>Lipsky</td>
<td>2010</td>
<td>CareFusion</td>
<td>8747 (1021)</td>
<td>Bacteremia</td>
<td>Variable</td>
</tr>
<tr>
<td>Liu</td>
<td>2010</td>
<td>Kaiser HMO</td>
<td>155 474 (NA)</td>
<td>Length of stay</td>
<td>Variable</td>
</tr>
<tr>
<td>Hipsley-Cox</td>
<td>2010</td>
<td>QResearch, UK</td>
<td>3 610 918 (205 134 app)</td>
<td>Acute Kidney Injury Risk and Injury</td>
<td>Variable</td>
</tr>
<tr>
<td>Matheny</td>
<td>2010</td>
<td>Kaiser HMO</td>
<td>155 474 (NA)</td>
<td>Length of stay</td>
<td>Variable</td>
</tr>
<tr>
<td>Lipsky</td>
<td>2010</td>
<td>CareFusion</td>
<td>8747 (1021)</td>
<td>Bacteremia</td>
<td>Variable</td>
</tr>
<tr>
<td>Liu</td>
<td>2010</td>
<td>Kaiser HMO</td>
<td>155 474 (NA)</td>
<td>Length of stay</td>
<td>Variable</td>
</tr>
<tr>
<td>Hipsley-Cox</td>
<td>2010</td>
<td>QResearch, UK</td>
<td>3 610 918 (205 134 app)</td>
<td>Acute Kidney Injury Risk and Injury</td>
<td>Variable</td>
</tr>
<tr>
<td>Matheny</td>
<td>2010</td>
<td>Kaiser HMO</td>
<td>155 474 (NA)</td>
<td>Length of stay</td>
<td>Variable</td>
</tr>
<tr>
<td>Lipsky</td>
<td>2010</td>
<td>CareFusion</td>
<td>8747 (1021)</td>
<td>Bacteremia</td>
<td>Variable</td>
</tr>
<tr>
<td>Liu</td>
<td>2010</td>
<td>Kaiser HMO</td>
<td>155 474 (NA)</td>
<td>Length of stay</td>
<td>Variable</td>
</tr>
<tr>
<td>Hipsley-Cox</td>
<td>2010</td>
<td>QResearch, UK</td>
<td>3 610 918 (205 134 app)</td>
<td>Acute Kidney Injury Risk and Injury</td>
<td>Variable</td>
</tr>
<tr>
<td>Matheny</td>
<td>2010</td>
<td>Kaiser HMO</td>
<td>155 474 (NA)</td>
<td>Length of stay</td>
<td>Variable</td>
</tr>
<tr>
<td>Lipsky</td>
<td>2010</td>
<td>CareFusion</td>
<td>8747 (1021)</td>
<td>Bacteremia</td>
<td>Variable</td>
</tr>
<tr>
<td>Liu</td>
<td>2010</td>
<td>Kaiser HMO</td>
<td>155 474 (NA)</td>
<td>Length of stay</td>
<td>Variable</td>
</tr>
<tr>
<td>Hipsley-Cox</td>
<td>2010</td>
<td>QResearch, UK</td>
<td>3 610 918 (205 134 app)</td>
<td>Acute Kidney Injury Risk and Injury</td>
<td>Variable</td>
</tr>
<tr>
<td>Matheny</td>
<td>2010</td>
<td>Kaiser HMO</td>
<td>155 474 (NA)</td>
<td>Length of stay</td>
<td>Variable</td>
</tr>
<tr>
<td>Lipsky</td>
<td>2010</td>
<td>CareFusion</td>
<td>8747 (1021)</td>
<td>Bacteremia</td>
<td>Variable</td>
</tr>
<tr>
<td>Liu</td>
<td>2010</td>
<td>Kaiser HMO</td>
<td>155 474 (NA)</td>
<td>Length of stay</td>
<td>Variable</td>
</tr>
<tr>
<td>Hipsley-Cox</td>
<td>2010</td>
<td>QResearch, UK</td>
<td>3 610 918 (205 134 app)</td>
<td>Acute Kidney Injury Risk and Injury</td>
<td>Variable</td>
</tr>
<tr>
<td>Matheny</td>
<td>2010</td>
<td>Kaiser HMO</td>
<td>155 474 (NA)</td>
<td>Length of stay</td>
<td>Variable</td>
</tr>
<tr>
<td>Lipsky</td>
<td>2010</td>
<td>CareFusion</td>
<td>8747 (1021)</td>
<td>Bacteremia</td>
<td>Variable</td>
</tr>
<tr>
<td>Liu</td>
<td>2010</td>
<td>Kaiser HMO</td>
<td>155 474 (NA)</td>
<td>Length of stay</td>
<td>Variable</td>
</tr>
<tr>
<td>Hipsley-Cox</td>
<td>2010</td>
<td>QResearch, UK</td>
<td>3 610 918 (205 134 app)</td>
<td>Acute Kidney Injury Risk and Injury</td>
<td>Variable</td>
</tr>
<tr>
<td>Matheny</td>
<td>2010</td>
<td>Kaiser HMO</td>
<td>155 474 (NA)</td>
<td>Length of stay</td>
<td>Variable</td>
</tr>
<tr>
<td>Lipsky</td>
<td>2010</td>
<td>CareFusion</td>
<td>8747 (1021)</td>
<td>Bacteremia</td>
<td>Variable</td>
</tr>
<tr>
<td>Liu</td>
<td>2010</td>
<td>Kaiser HMO</td>
<td>155 474 (NA)</td>
<td>Length of stay</td>
<td>Variable</td>
</tr>
<tr>
<td>Hipsley-Cox</td>
<td>2010</td>
<td>QResearch, UK</td>
<td>3 610 918 (205 134 app)</td>
<td>Acute Kidney Injury Risk and Injury</td>
<td>Variable</td>
</tr>
<tr>
<td>Matheny</td>
<td>2010</td>
<td>Kaiser HMO</td>
<td>155 474 (NA)</td>
<td>Length of stay</td>
<td>Variable</td>
</tr>
<tr>
<td>Lipsky</td>
<td>2010</td>
<td>CareFusion</td>
<td>8747 (1021)</td>
<td>Bacteremia</td>
<td>Variable</td>
</tr>
<tr>
<td>Liu</td>
<td>2010</td>
<td>Kaiser HMO</td>
<td>155 474 (NA)</td>
<td>Length of stay</td>
<td>Variable</td>
</tr>
<tr>
<td>Hipsley-Cox</td>
<td>2010</td>
<td>QResearch, UK</td>
<td>3 610 918 (205 134 app)</td>
<td>Acute Kidney Injury Risk and Injury</td>
<td>Variable</td>
</tr>
<tr>
<td>Matheny</td>
<td>2010</td>
<td>Kaiser HMO</td>
<td>155 474 (NA)</td>
<td>Length of stay</td>
<td>Variable</td>
</tr>
<tr>
<td>Lipsky</td>
<td>2010</td>
<td>CareFusion</td>
<td>8747 (1021)</td>
<td>Bacteremia</td>
<td>Variable</td>
</tr>
<tr>
<td>Liu</td>
<td>2010</td>
<td>Kaiser HMO</td>
<td>155 474 (NA)</td>
<td>Length of stay</td>
<td>Variable</td>
</tr>
</tbody>
</table>
validate a phenotype or type of diagnosis. These papers report very similar metrics; however, they are not aiming to predict a future event but detect a current clinical status. Over the 6-year study period the number of papers increased from 5 in 2009 to 31 in 2014. Papers were published in what we defined as medical journals, informatics journals, and health services journals. Studies originated from EHRs in 15 different countries.

Designing EHR Prediction Studies

Most studies used a cohort design (n = 94), with the rest using case-control. Studies differed on whether single or multiple EHRs were used. Forty studies were single-center studies. Another 17 studies occurred across multiple hospitals but used the same EHR system. These included studies within the Veterans Administration (VA) system, Geisenger Health System, and Kaiser Permanente. The remaining 50 studies encompassed linkage across different EHR systems in multiple hospitals across a city or region in the same country. This included studies that aggregated data into single databases like CareFusion (n = 7) and QResearch (n = 17).

Many studies considered multiple end points (eg, readmission and mortality). Mortality, both in- and out-of-hospital was the most common outcome (n = 27). Of studies that looked at overall mortality, most relied on linkage with the Social Security Death Index and/or the National Death Index. Twenty papers were published looking at service utilization with a primary focus on 30-day readmission (n = 11). Although re-hospitalization can occur at different hospitals than the index hospital, only half of these studies were multicenter. The largest class of papers (n = 60) were those predicting a clinical endpoint (eg, venous thromboembolism or myocardial infarction).

There was heterogeneity in how outcomes definitions were defined and reported. Several of these studies (n = 6) did not report how outcomes were defined, 22 relied on only ICD-9/10 codes, while 32 used additional information such as laboratory measures (eg, glucose levels for diabetes) or medications (eg, anti-depressant use for depression).

Finally, we examined the time horizon of prediction. Of the 81 studies with defined end-points (ie, not having variable time in-hospital mortality), 34 predicted an outcome of <90 days, with a 30-day horizon being the most common (n = 22). Conversely, 32 (17 of which were from QResearch) studies looked at events at 1-year or further out.

Leveraging EHR Data

Figure 1 displays how different studies leveraged the amount of available data. Overall the sample sizes were large. There was a median size of 26 100 observations and 39 studies had sample sizes above 100 000. The median number of events was 2543. Among cohort studies, there was heterogeneity in the prevalence of the outcome, with a median event rate of 8.2% (Interquartile Range (IQR): 3%, 20%). The median number of predictors was 27, with only 29 studies using 50 or more predictors, and 46 studies using 20 or fewer.

Challenges inherent to EHR data include modeling repeated measurements, handling missing data, and considering loss to follow-up. Table 2 describes how different studies approached these issues. Most (n = 70) studies did not consider repeated...
measurements. Among those that did, most used summary metrics with only 8 studies modeling multiple measures over time. Another area of challenge for EHR data is the presence of missingness. Missing data can be present in the predictors or in the outcomes (via censoring). Only 58 studies assessed missingness, with the most common strategy being multiple imputation. A related complication is informative observations – where the presence of an observation (clinic visit) is itself meaningful. No study assessed the role informative observations may play. Finally, censoring and loss to follow-up were not always well investigated. We determined that censoring was not applicable to any study that looked at in-hospital events, leaving 73 applicable studies. Of these, only 12 assessed the role censoring may play.

Model Development, Evaluation and Reporting

Generalized linear models (ie, logistic regression, Cox regression, etc.) were the most common algorithms used (n = 84) to develop the prediction model. Other approaches included, Bayesian methods (n = 11), random forests (n = 10), and regularized regression (ie, LASSO and ridge regression) (n = 7). Most studies that used regression incorporated some form of variable selection (n = 67), most often via stepwise approaches. Conversely, studies that used machine learning methods were more likely to include all of the selected variables (n = 14). All but 13 studies used some form of validation. The most common form was split sample (n = 67), followed by cross-validation (n = 21) and then bootstrapping (n = 9), with some studies using multiple forms of validation. Of the 50 studies that involved multiple hospital and EHR systems, 26 performed their validation across the sites – ie, training in one (set of) site(s) and validating in another.

Most of the papers (n = 90) assessed the model’s discrimination via the c-statistic. Figure 2 shows the c-statistics across outcome type. For studies that reported more than one c-statistic (eg, over different time horizons), the best fitting model is reported. The mortality and clinical prediction studies had higher median c-statistics (c = 0.84 and 0.83, respectively) than hospitalization and service utilization prediction studies (c = 0.71 and 0.73, respectively). The c-statistic was not related to publication year (Spearman r = 0.07), suggesting that studies published in recent years are not only the ones with better predictive discrimination ability. Of the 12 studies that assessed model performance over different time horizons, all but 3 reported that the models performed worse the further out they forecast. Furthermore, of 7 studies that reported internal as well as external validation results, 5 had stronger performance on the
internal sample. Finally, of the 8 models that reported performance for both logistic regression and machine learning models, 6 reported better performance among the machine learning models. Beyond c-statistics, other evaluation metrics were less common. Only a minority of papers assessed model calibration (n = 48) typically via the Hosmer-Lemeshow test or Calibration Slope. Another way models were evaluated was through comparison to existing risk scores, which 26 studies did. This included comparing to human judgment (n = 7). All but 2 studies reported that their algorithm performed as well or better than the comparative approach. Of these studies, four calculated the Net Reclassification Index. Few studies used metrics that assess the potential impact of the model on practical decision-making, with only 26 studies reporting the positive predictive value (PPV), with a median value of 1.7%. The PPV was uncorrelated with the reported AUC (r = −0.18, 95% CI, −0.45 to 0.12).

Almost all studies (n = 95) assessed the association of some or all of the predictor variables. Finally, over three-quarters of the studies (n = 83) reported the final model so that they could be reproduced and/or implemented. This was most common among those using linear models (logistic regression, LASSO etc.) where 73 of 85 studies reported the model coefficients.

**DISCUSSION**

Over the past 6 years, at least 107 studies have been published creating prediction models using EHR data, with the number of these studies increasing over time. Overall, we found room for improvement in maximizing the advantages of EHR-data for risk modeling and addressing inherent challenges (see Table 3).

The primary advantage of EHR data is its size. As expected most studies had large sample sizes, with 39 studies having a sample size of over 100,000 people. The latter group is larger than other common large epidemiological cohorts and is more comparable to large registries. However, unlike registries, EHRs are not disease-specific, allowing one to look at multiple outcomes with the same data source. For example, 4 different algorithms were published with data from the Geisenger Health System, modeling the probability of heart failure, 30-day readmission, stroke, and diabetes remission.

Another advantage of the large data size in EHRs is the opportunity to create validation sets. Almost all studies performed some validation either through cross-validation or sample splitting. However, external validation was uncommon and almost all studies validated performance within the same EHR. This limits generalizability and may reduce discrimination when these models are applied in other sites or in other EHR systems. A key area for future improvement would be the use of multicenter studies. The use of hospital networks in a single region can ensure fuller capture of patient encounters – improving internal data reliability. Studies performed in closed networks like Kaiser Permanente and the Veterans Affairs (VA) illustrate this design. More important is the validation across different systems. Risk scores such as the Framingham Risk Score were designed to be general scores that could be used in any population and have been adopted even across different countries. However, hospitals (and by extension EHRs) serve specific patient populations and a strong score should leverage the unique characteristics of that population. Of the 49 studies that used multiple sites, only 26 validated across those sites, 17 of which were performed by the QResearch team. This represents a lost opportunity to assess the external validity of a prediction algorithm. While the limited data suggest scores perform worse externally, an important open question is how well a prediction algorithm developed in one center will port to another. This then raises some important questions: should we expect that a model developed in one site should port over to another center? More importantly, should individual centers attempt to optimize their prediction model for their particular center or try to create generalizable scores? Efforts such as PCORNet that allow for creation of linked hospital networks may help facilitate multicenter analyses needed to create and assess more broadly generalizable EHR-based risk scores.

The advantages of the size of EHR data are not limited to numbers of patients – a key advantage is access to a large number of potential predictor variables. In defining clinical outcomes, most studies used a variety of data elements, moving beyond just billing codes. However, we found that many studies, did not fully utilize the depth of information on patients available in the medical record to identify predictor variables. Many studies instead opted for smaller predefined lists. Moreover, few studies used longitudinal measurements for patient. The opportunity to observe changes in

---

**Table 3. Areas of improvement**

<table>
<thead>
<tr>
<th>Challenge</th>
<th>Way to Improve</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Multicenter studies</td>
<td>Use EHRs from multiple sites Validate Across Sites</td>
<td>Assess portability of models</td>
</tr>
<tr>
<td>2 Predictor variables</td>
<td>Incorporate time-varying (longitudinal) factors Use larger variable sets</td>
<td>Better leverage from the available data</td>
</tr>
<tr>
<td>3 Consideration of biases</td>
<td>Missing data Loss to follow-up</td>
<td>Assess the robustness of the models</td>
</tr>
<tr>
<td>4 Evaluation metrics</td>
<td>External validation</td>
<td>Metrics of Clinical Utility (eg, PPV, Net Benefit)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Develop an understanding of how models will impact clinical decision making</td>
</tr>
</tbody>
</table>

---

*Figure 2. Distribution of c-statistics across different outcomes. Thirteen modeled more than one outcome type.*
patients is a key strength of EHRs. While the integration of such repeated observations is challenging from a statistical perspective, methods do exist.129

Two of the largest challenges in EHR-based studies are the presence of missing data and informative presence. While missing data has been an acknowledged challenge in EHR studies,130 little more than half of the studies commented on the presence of missing data, with studies using a variety of analytic approaches. More importantly, no study commented on the issue of informative presence. It is has been recognized that EHRs contain sicker people on average131 and others have noted that this can lead to biased associations.132 However, there has been minimal work in this area, and it is unclear how such biased observations impact prediction models. Another challenge is the potential for loss to follow-up, which few studies assessed. Outside of comprehensive medical systems (eg, Kaiser Permanente, VA) it is unclear what role such biases may play. This is particularly important for studies that assess hospital readmission as patients can get readmitted elsewhere, without knowledge to the researcher.

The final area that studies can show improvement is in the use of evaluation metrics. Almost all studies assessed the model’s discrimination via the c-statistic. Because the value of the c-statistic is independent of the prevalence of the outcome,133 it is useful for comparing models across different diseases. However, this makes the c-statistic poorly suited for assessing clinical utility as many outcomes in EHR based studies are relatively rare. The promise of EHR based prediction models is to improve clinical decision making. Therefore, to assess clinical utility, metrics that take prevalence into account such as PPV should also be assessed. Moreover, as our results illustrate, and others have noted,134 there is minimal relationship between the PPV and the c-statistic.

One area we have not considered is the (future) role of genetic data. As evidenced by efforts like the eMERGE Network,135 genetic data will likely become a regular field in EHRs and used in prediction models.136 While this will create many foreseeable and unforeseeable technical and analytic challenges,137–140 this also creates a problem for model application: any patient that is not a regular patient in the health system will likely not have these data on file. Our ability to develop risk predictors that use information available beyond the point-of-care, such as genetic and socio-economic factors, is dependent on our ability to resolve these issues.

There are some limitations in our analysis. As the search results suggest, this is a very dynamic field, and there are likely some studies published that we did not capture. In particular, any paper that did not specify the EHR as the data source would be missed unless captured by our research team name-specific searches. Another limitation is that since we aimed to understand the status quo of the literature in general, we did not do meta-analysis for any of the results for specific outcomes. Future studies could focus on particular outcomes (eg, 30-day readmission) to see how they perform and which algorithms are optimal. In addition, our review only focuses on the published results and reported information in publications. Therefore, it is possible that some studies did consider, eg, missing data but simply did not report it. Finally, this review only focused on the development of the algorithm and the reported metrics. As more of the models find their way into clinical practice, it will be important to assess how they perform prospectively.

We suggest a number of areas that researchers should consider when conducting EHR-based predictions studies. Firstly, it is clear that more work is needed in the implementation of multicenter studies that use multiple EHRs, and these studies should attempt to validate their results across the different centers. Moreover, it is important to assess in each case whether we should strive for generally deployable scores or center-specific ones. There is room to use more predictor variables, particularly longitudinal information. Future work needs to consider the impact of informed presence and how that influences prediction models. Similarly, more consideration needs to be taken for missing data as well as loss to follow-up. While model reporting was very good in the studies that used logistic regression, it is important to consider what the added gain would be of a more complex model. Finally, since risk models are often used for clinical decision support, evaluation metrics should assess how the algorithms impact clinical decision making.

FUNDING

This study was supported by the National Institute of Diabetes and Digestive and Kidney Diseases grant K25DK097279 to B.A.G. The funders played no role in the study design or analysis.

COMPETING INTERESTS

None.

CONTRIBUTORS

B.A.G. conceptualized, designed, performed, and wrote up the study. A.M.N. reviewed papers and edited the manuscript. M.J.P. guided analytic strategy and edited the manuscript. J.P.A.I designed the review, guided analytic strategy, and edited the manuscript.

REFERENCES


32. Crane SJ, Tung EE, Hanson GJ, Cha S, Chaudhry R, Takahashi PY. Use of an electronic administrative database to identify older community dwelling adults at high-risk for hospitalization or emergency department visits: the elders risk assessment index. BMC Health Serv Res. 2010;10:338.


