

---

## Review

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

Benjamin A Goldstein,<sup>1,2</sup> Ann Marie Navar,<sup>2,3</sup> Michael J Pencina,<sup>1,2</sup> and John P A Ioannidis<sup>4,5</sup>

<sup>1</sup>Department of Biostatistics & Bioinformatics, Duke University, Durham, NC 27710, USA, <sup>2</sup>Center for Predictive Medicine, Duke Clinical Research Institute, Duke University, Durham, NC 27710, USA, <sup>3</sup>Division of Cardiology at Duke University Medical Center, Durham, NC 27710, USA, <sup>4</sup>Department of Medicine, Stanford University, Palo Alto, CA 94305, USA, and <sup>5</sup>Department 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.<sup>1</sup> 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.<sup>2</sup>

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.<sup>3</sup> 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.<sup>4</sup> 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.<sup>5</sup>

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 unlike traditional algorithms that need to be translated to a clinical environment.

The goal of this systematic review is to assess how authors have utilized EHR data as the primary source to build and validate risk prediction models. Multiple reviews of prognostic models have been performed for a variety of disease areas.<sup>6–9</sup> In this review, we focus specifically on the unique components of EHR data and how they have been utilized. We evaluated the studies on aspects of the reported design, analysis, results, and evaluation that may be important for the validity and utility of these models. We highlight areas where researchers can leverage the unique components of EHRs and pitfalls that need to be considered.

## METHODS

We followed the PRISMA<sup>10</sup> guidelines 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.<sup>11</sup> suggested search terms for clinical prediction papers that were later updated by Geersing et al.<sup>12</sup> 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<sup>13</sup>) 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 TRIPOD<sup>14,15</sup> guidelines for reporting of risk prediction studies and the review by D’Agostino et al.<sup>16</sup> 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.

## RESULTS

Our initial search resulted in 8127 papers. Reviewer 1 identified 90 potentially acceptable papers while reviewer 2 identified 57. Upon comparison, 81<sup>17–97</sup> (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,<sup>98–107</sup> 17 of which were based on the QResearch database,<sup>108–124</sup> 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

Author	Year	Location	Sample size (no. of events)	Outcome	Time horizon
Himes	2009	Partners Healthcare Research Patient Data Registry	10 341 (843)	Chronic obstructive pulmonary disease	5 years
Hipsley-Cox	2009	QResearch, UK	3 773 585 (115 616)	Diabetes	10 Years
Hipsley-Cox	2009	QResearch, UK	3 633 812 (50 755)	Fracture	10 Years
Sairamesh	2009	Not reported	30 095 (NR)	Non-adherence	Not reported
Smits	2009	Five Health Centers Amsterdam, NL	3045 (470)	Persistent attenders	3 years
Amarasingham	2010	Parkland Hospital, TX	1372 (331)	Death or readmission	30 days
Crane	2010	Mayo Clinic, MN	13 457 (NA)	Number hospital visits	2 years
Hipsley-Cox	2010	QResearch, UK	3 610 918 (205 134 app)	Cardiovascular disease	10 years
Hipsley-Cox	2010	QResearch, UK	3 953 092 (110 250 app)	Kidney disease	10 years
Johnson	2010	Kaiser HMO	5171 (145)	Hyperkalemia	1–90 days
Lipsky	2010	CareFusion	8747 (1021)	Bacteremia	Variable
Liu	2010	Kaiser HMO	155 474 (NA)	Length of stay	Variable
Matheny	2010	Vanderbilt University Medical Center	26 107 (1352)	Acute Kidney Injury Risk and Injury	1–30 days
Robbins	2010	Massachusetts General and Brigham Women's Hospital	1074 (120)	Virological failure	1 year +
Skevofilakas	2010	Athens Hippokraton Hospital, Greece	55 (17)	Retinopathy	5 years
Tabak	2010	CareFusion	2 450 224 (105 579)	Mortality	Variable
Wu	2010	Geisenger Health System, PA	4489 (536)	Heart failure	6 months
Barrett	2011	Vanderbilt University Medical Center	832 (216)	Adverse events	30 days
Chang	2011	Taipei Medical University, Taiwan	4682 (505)	Hospital Acquired Infection	Not reported
Cheng	2011	Veterans Administration	12 589 (3329)	Diabetes	8 years
Hipsley-Cox	2011	QResearch, UK	3 594 690 (3970)	Gastric cancer	2 years
Hipsley-Cox	2011	QResearch, UK	3 673 278 (6071)	Lung cancer	2 years
Hipsley-Cox	2011	QResearch, UK	3 555 303 (21 699)	Thromboembolism	1 and 5 years
Kor	2011	Mayo Clinic, MN	4366 (113)	Acute Lung Injury	5 days
Lipsky	2011	CareFusion	3018 (646)	Amputation	Variable
Meyfroidt	2011	Leuven University Hospital, Belgium	961 (NA)	ICU length of stay and discharge	Variable
Puopolo	2011	Kaiser, HMO	1413 (350)	Infection	3 days
Robiscsek	2011	Northshore University Health System, Chicago, IL	52 574 (1804)	Methicillin-resistant <i>Staphylococcus aureus</i>	24 h
Saltzman	2011	CareFusion	61 726 (1813)	Mortality	Variable
Smith	2011	Kaiser HMO	4696 (1859)	Hospitalization or mortality	5 years
Westra	2011	15 Home Health Agencies, United States	1793 (684)	Improved urinary incontinence	Variable
Woller	2011	Intermountain Healthcare	143 975 (5288)	Venos Thromboembolism	90 days
Zhao	2011	Not reported	15 069 (98)	Pancreatic cancer	Not reported
Busch	2012	19 US Clinics	3168 (NA)	Survey score	1 year
Casserat	2012	3 US Hospices	21 074 (5562)	Mortality	7 days
Cholleti	2012	Emory University	232 705 (NR)	Readmission	30 days
Churpeck	2012	Northwestern Medical Center, IL	47 427 (2908)	Cardiac arrest and ICU transfer	30 min
Escobar	2012	Kaiser HMO	43 818 (4036)	Deterioration after admission	12 h
Hipsley-Cox	2012	QResearch, UK	3 587 653 (7401)	Colorectal cancer	2 years
Hipsley-Cox	2012	QResearch, UK	4 726 046 (88 457)	Fracture	10 years
Hipsley-Cox	2012	QResearch, UK	1 767 585 (1514)	Ovarian cancer	2 years
Hipsley-Cox	2012	QResearch, UK	3 608 311 (2196)	Pancreatic cancer	2 years
Hipsley-Cox	2012	QResearch, UK	3 599 890 (4500)	Renal cancer	2 years
Hivert	2012	Center Hospitalier Universitaire de Sherbrooke, Canada	31 823 (2997)	Diabetes of Coronary Heart Disease (CHD)	5 years
Hong	2012	National System, Korea	4 713 462 (57 706)	Hospital mortality	Variable
Karnick	2012	Marshfield Clinic, Wisconsin	1093 (547)	Atrial Fibrillation	1, 3, 5 years
Kawaler	2012	Marshfield Clinic, Wisconsin	720 (144)	Venous thromboembolism	90 days
Khan	2012	13 US Hospitals	227 (34)	Readmission	30 days
Mani	2012	Vanderbilt University Medical Center	2280 (228)	Diabetes	1 year, 180, 1 day
Monsen	2012	14 US Home Health Care Organizations	1643 (496)	Hospitalization	Not reported
Navarro	2012	Community Medical Center, CA	8298 (897)	Readmission	30 days
Nijhawan	2012	Parkland Hospital, TX	2476 (623)	Death 30 day admission	30 days
Tescher	2012	Mayo Clinic, MN	12 566 (416)	Pressure Ulcer	Time to event
Wang	2012	Veterans Administration	198 460 (14 124)	Hospitalization or mortality	30 days and 1 year
Alvarez	2013	Parkland Hospital, TX	46 974 (585)	Resuscitation or mortality	Variable
Baille	2013	University Pennsylvania Health System, PA	120 396 (17 337)	Readmission	30 days
Billings	2013	Five Primary Care Trusts, UK	1 836 099 (94 692)	Admission	1 year
Choudhry	2013	Eight Hospitals Chicago, IL	126 479 (9151)	Readmission	30 days
Churpeck	2013	Northwestern Medical Center, IL	59 643 (3038)	Cardiac arrest and ICU transfer	24 h
Eapen	2013	US National Linkage	30 828 (8374)	Death or readmission	30 days
Escobar	2013	Kaiser, HMO	391 584 (19 931)	Mortality	30 days
Hebert	2013	Northshore University Health System, Chicago, IL	829 (198)	Relapse	15–56 days
Hipsley-Cox	2013	QResearch, UK	6 673 458 (4 190 003)	Admissions	2 years
Hipsley-Cox	2013	QResearch, UK	1 908 467 (35 508)	Cancer	2 years
Hipsley-Cox	2013	QResearch, UK	1 942 245 (34 434)	Cancer	2 years
Jin	2013	Gangnam Severance Hospital, Seoul, Korea	19 303 (1328)	Bacteremia	24 h
Kennedy	2013	Veterans Administration	113 973 (4995)	Cardiovascular or cerebrovascular death	5 years
Mathias	2013	Northwestern Medical Faculty Foundation, IL	7463 (838)	Mortality	5 years
Perkins	2013	Geisenger Health System, PA	607 (116)	Readmission	30 days
Ramchandran	2013	Northwestern Medical Center, IL	3062 (294)	Mortality	30 days
Rothman	2013	Four US hospitals	171 250 (NA)	Patient condition defined through mortality	24 h mort, 30 day readmission
Singal	2013	Parkland Hospital, TX	1291 (347)	Readmission and death	30 days
Tabak	2013	CareFusion	512 484 (2083)	Mortality	Variable
Tabak	2013	CareFusion	102 626 (3197)	Mortality	Variable
Wang	2013	Veterans Administration	4 598 408 (378 863)	Mortality or hospitalization	90 days and 1 year

(Continued)

Table 1. Continued

Author	Year	Location	Sample size (no. of events)	Outcome	Time horizon
Wells	2013	Cleveland Clinic, OH	33 067 (3661)	Stroke, Heart failure (HF), CHD, mortality	Variable
Atchinson	2014	All Children's Hospital, FL	400 (50)	Venous thromboembolism	Variable
Ayyagari	2014	Geisinger Health System, PA	34 797 (4272)	Stroke	3 years
Carter	2014	Nuffield Orthopedic Center, England	2130 (NA)	Length of stay	Variable
Churpeck	2014	University of Chicago Medical Center, IL	59 301 (2543)	Cardiac arrest and ICU transfer	Variable
Churpeck	2014	Five Chicago Hospitals, IL	269 999 (16 452)	Cardiac arrest and ICU transfer	8 h
Dai	2014	Boston Medical Center, MA	45 579 (NR)	Hospitalization for heart procedure	1, 3, 6, and 12 months
Goldstein	2014	US Dialysis Clinics	3394 (1697)	Sudden Cardiac Death	48 h, 30 day, 90, 180, 365
Gultepe	2014	UC Davis, CA	741 (151)	Mortality	Variable
Gupta	2014	Barwon Health, Australia	736 (NR)	Mortality	6, 12, 24 month
Hao	2014	Linkage Maine Hospitals	487 572 (96 703)	Readmission	30 days – time to event
Hebert	2014	Ohio State University Health Center, OH	3968 (888)	Readmission	30 days
Hipsley-Cox	2014	QRResearch, UK	5 446 646 (175 982)	Stroke	10 years
Hipsley-Cox	2014	QRResearch, UK	6 097 082 (632 500)	Bleeds	5 years
Huang	2014	Palo Alto Medical Foundation and Group Health Research	35 000 (5000)	Depression	6 months and 1 year
Huang	2014	Pediatric Clinics, MI	104 799 (11 737)	No Show	Variable
Konito	2014	University Hospital, Finland	23 528 (NA)	Patient Acuity	Next Day
Kotz	2014	GP Practices, Scotland	27 088 (728 658)	Chronic Obstructive Pulmonary Disease	10 years
Mani	2014	Vanderbilt University Medical Center	299 (209)	Sepsis	12 h
O'Leary	2014	Medstar Washington Hospital Center, Washington DC	219 (77)	Surgical Intervention	Variable
Puttkammer	2014	2 Hospitals, Haiti	923 (196)	ART Failure	Months 7–12
Rana	2014	Barwon Health, Australia	1660 (105)	Readmission	30 days and 12 months
Rapsomaniki	2014	Caliber Database, UK	102 023 (28107)	Mortality or MI	Variable
Roubinian	2014	Kaiser, HMO	444 969 (61 988)	Blood Transfusion	24 h
Sho	2014	Children's Hospital Pittsburgh, PA	157 (35)	<i>necrotizing enterocolitis totalis</i> (NEC) – Totalis	Variable
Smolin	2014	Rambam Medical Center, Israel	23 792 (4985)	Mortality	6 months and 1 year
Still	2014	Geisinger Health System, PA	690 (436)	Diabetes Remission	2 months, Variable
Sun	2014	Vanderbilt University Medical Center	1294 (615)	Hypertension Control	Variable
Tabak	2014	CareFusion	2 199 347 (52 603)	Mortality	Variable
Tran	2014	Barwon Health, Australia	7399 (13)	Suicide Attempt	30, 60, 90, 180 days
Uyar	2014	German Hospital, Turkey	2453 (270)	In-vitro Fertilization (IVF) Implantation	1 day
Zhai	2014	Cincinnati Children's Hospital, OH	7298 (526)	Transfer to PICU	24 h

NA = Continuous outcome (eg, length of stay) so number of events is not applicable; NR = Not reported; appr = Exact number not reported. Estimated from the paper.

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 ( $n = 67$ ), informatics ( $n = 27$ ), and health services ( $n = 13$ ) 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, Geisinger 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 QRResearch ( $n = 17$ ).

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 like 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 QRResearch) 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

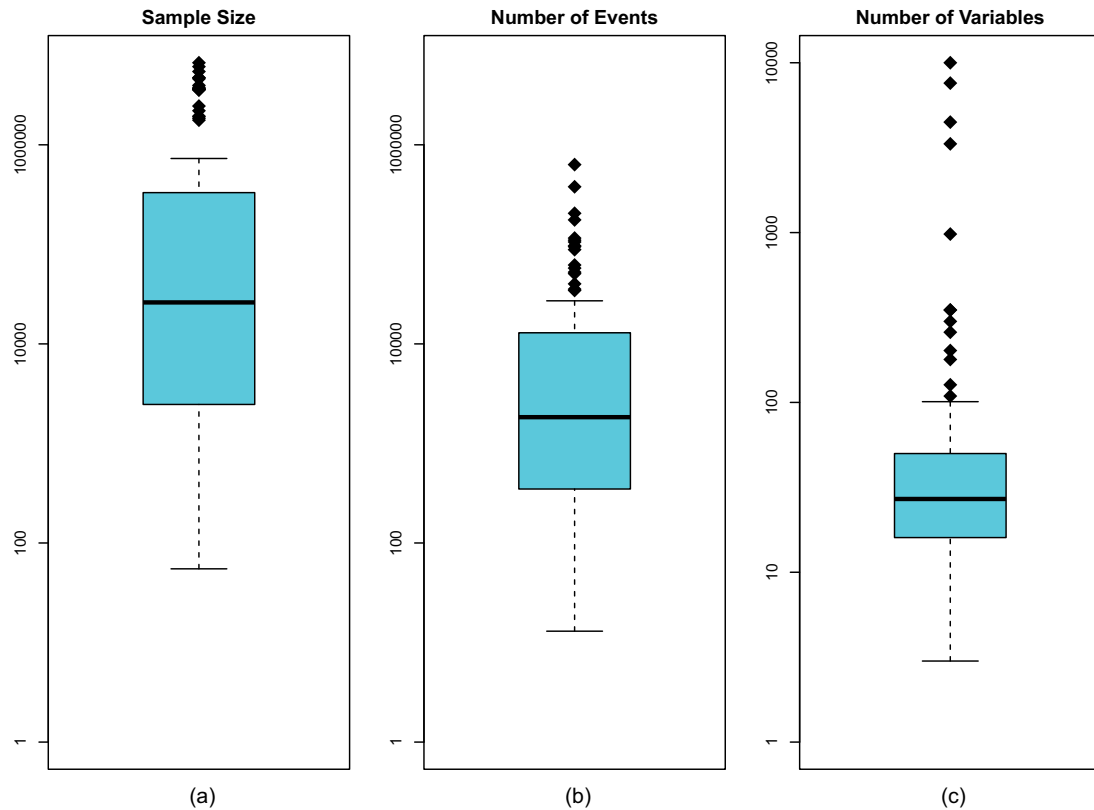


Figure 1. Leveraging of available data across different studies.

Table 2. Challenges of EHR Data

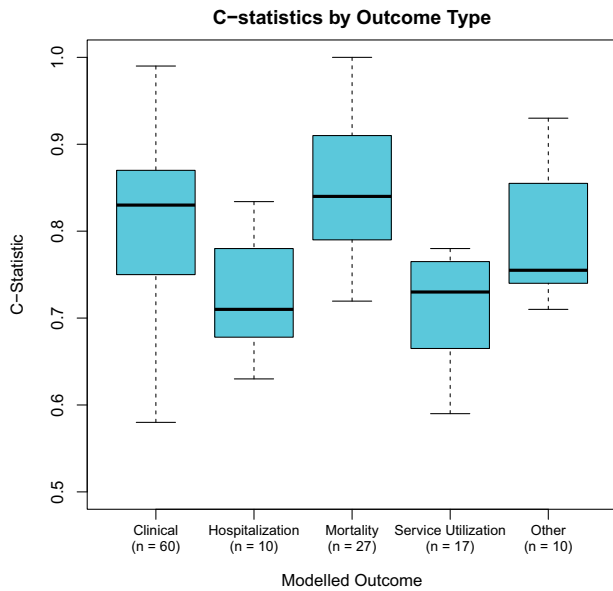
Longitudinal data	Missing data	Loss to follow-up/ Competing Event
Not considered ( $n = 71$ )	Not considered ( $n = 49$ )	Not considered ( $n = 40$ )
Peak value ( $n = 14$ )	Single imputation ( $n = 16$ )	In hospital event ( $n = 33$ )
Mean/Median ( $n = 7$ )	Multiple imputation ( $n = 21$ )	Time to event model ( $n = 22$ )
Count ( $n = 11$ )	Complete case ( $n = 10$ )	Linked to registry ( $n = 7$ )
Variability ( $n = 2$ )	Missing category/ indicator ( $n = 12$ )	Remove those lost ( $n = 5$ )
Trend ( $n = 3$ )	Drop variable ( $n = 3$ )	
Time Varying ( $n = 9$ )		

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



**Figure 2.** Distribution of *c*-statistics across different outcomes. Thirteen modeled more than one outcome type.

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

**Table 3.** Areas of improvement

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

data from the Geisinger Health System, modeling the probability of heart failure,<sup>95</sup> 30-day readmission,<sup>66</sup> stroke,<sup>20</sup> and diabetes remission.<sup>82</sup>

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.<sup>125</sup> A key area for future improvement would be the use of multicenter studies.<sup>126</sup> 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.<sup>127</sup> 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<sup>128</sup> 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

patients is a key strength of EHRs. While the integration of such repeated observations is challenging from a statistical perspective, methods do exist.<sup>129</sup>

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,<sup>130</sup> 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 has been recognized that EHRs contain sicker people on average<sup>131</sup> and others have noted that this can lead to biased associations.<sup>132</sup> 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,<sup>133</sup> 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.<sup>2</sup> 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,<sup>134</sup> 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,<sup>135</sup> genetic data will likely become a regular field in EHRs and used in prediction models.<sup>136</sup> While this will create many foreseeable and unforeseeable technical and analytic challenges,<sup>137-140</sup> 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

1. Charles D, Gabriel M, Searcy T. Adoption of electronic health record systems among U.S. non-federal acute care hospitals: 2008-2014. 2015 <https://www.healthit.gov/sites/default/files/data-brief/2014HospitalAdoptionDataBrief.pdf>.
2. Rothman B, Leonard JC, Vigoda MM. Future of electronic health records: implications for decision support. *Mt Sinai J Med NY*. 2012;79(6):757-768.
3. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation*. 1998;97(18):1837-1847.
4. 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.
5. Richesson RL, Rusincovitch SA, Wixted D, *et al*. A comparison of phenotype definitions for diabetes mellitus. *J Am Med Inform Assoc*. 2013;20(e2):e319-e326.
6. Tangri N, Inker L, Levey AS. A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. *J Clin Epidemiol*. 2013;66(6):697.
7. Mallett S, Royston P, Waters R, Dutton S, Altman DG. Reporting performance of prognostic models in cancer: a review. *BMC Med*. 2010;8:21.
8. Counsell C, Dennis M. Systematic review of prognostic models in patients with acute stroke. *Cerebrovasc Dis Basel Switz*. 2001;12(3):159-170.
9. Collins GS, Mallett S, Omar O, Yu L-M. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Med*. 2011;9:103.
10. Liberati A, Altman DG, Tetzlaff J, *et al*. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *J Clin Epidemiol*. 2009;62(10):e1-e34.
11. Ingui BJ, Rogers MA. Searching for clinical prediction rules in MEDLINE. *J Am Med Inform Assoc*. 2001;8(4):391-397.

12. Geersing G-J, Bouwmeester W, Zuihthoff P, *et al.* Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews. *PLoS One.* 2012;7(2):e32844.
13. Hippisley-Cox J, Stables D, Pringle M. QRESEARCH: a new general practice database for research. *Inform Prim Care.* 2004;12(1):49–50.
14. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *J Clin Epidemiol.* 2015;68(2):134–143.
15. Moons KGM, Altman DG, Reitsma JB, *et al.* Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med.* 2015;162(1):W1–W73.
16. D'Agostino RB, Pencina MJ, Massaro JM, Coady S. Cardiovascular Disease Risk Assessment: Insights from Framingham. *Glob Heart.* 2013;8(1):11–23.
17. Alvarez CA, Clark CA, Zhang S, *et al.* Predicting out of intensive care unit cardiopulmonary arrest or death using electronic medical record data. *BMC Med Inform Decis Mak.* 2013;13:28.
18. Amarasingham R, Moore BJ, Tabak YP, *et al.* An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. *Med Care.* 2010;48(11):981–988.
19. Atchison CM, Arlikar S, Amankwah E, *et al.* Development of a new risk score for hospital-associated venous thromboembolism in noncritically ill children: findings from a large single-institutional case-control study. *J Pediatr.* 2014;165(4):793–798.
20. Ayyagari R, Vekeman F, Lefebvre P, *et al.* Pulse pressure and stroke risk: development and validation of a new stroke risk model. *Curr Med Res Opin.* 2014;30(12):2453–2460.
21. Baillie CA, VanZandbergen C, Tait G, *et al.* The readmission risk flag: using the electronic health record to automatically identify patients at risk for 30-day readmission. *J Hosp Med Off Publ Soc Hosp Med.* 2013;8(12):689–695.
22. Barrett TW, Martin AR, Storrow AB, *et al.* A clinical prediction model to estimate risk for 30-day adverse events in emergency department patients with symptomatic atrial fibrillation. *Ann Emerg Med.* 2011;57(1):1–12.
23. Billings J, Georgiou T, Blunt I, Bardsley M. Choosing a model to predict hospital admission: an observational study of new variants of predictive models for case finding. *BMJ Open.* 2013;3(8):e003352.
24. Busch AB, Neelon B, Zelevinsky K, He Y, Normand S-LT. Accurately predicting bipolar disorder mood outcomes: implications for the use of electronic databases. *Med Care.* 2012;50(4):311–319.
25. Carter EM, Potts HWW. Predicting length of stay from an electronic patient record system: a primary total knee replacement example. *BMC Med Inform Decis Mak.* 2014;14:26.
26. Casarett DJ, Farrington S, Craig T, *et al.* The art versus science of predicting prognosis: can a prognostic index predict short-term mortality better than experienced nurses do? *J Palliat Med.* 2012;15(6):703–708.
27. Chang Y-J, Yeh M-L, Li Y-C, *et al.* Predicting hospital-acquired infections by scoring system with simple parameters. *PLoS One.* 2011;6(8):e23137.
28. Cheng P, Neugaard B, Foulis P, Conlin PR. Hemoglobin A1c as a predictor of incident diabetes. *Diabetes Care.* 2011;34(3):610–615.
29. Cholleti S, Post A, Gao J, *et al.* Leveraging derived data elements in data analytic models for understanding and predicting hospital readmissions. *AMIA Annu Symp Proc AMIA Symp AMIA Symp.* 2012;2012:103–111.
30. Choudhry SA, Li J, Davis D, Erdmann C, Sikka R, Sutariya B. A public-private partnership develops and externally validates a 30-day hospital readmission risk prediction model. *Online J Public Health Inform.* 2013;5(2):219.
31. Churpek MM, Yuen TC, Park SY, Gibbons R, Edelson DP. Using electronic health record data to develop and validate a prediction model for adverse outcomes in the wards\*. *Crit Care Med.* 2014;42(4):841–848.
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.
33. Dai W, Brisimi TS, Adams WG, Mela T, Saligrama V, Paschalidis IC. Prediction of hospitalization due to heart diseases by supervised learning methods. *Int J Med Inf.* 2015;84(3):189–197.
34. Eapen ZJ, Liang L, Fonarow GC, *et al.* Validated, electronic health record deployable prediction models for assessing patient risk of 30-day rehospitalization and mortality in older heart failure patients. *JACC Heart Fail.* 2013;1(3):245–251.
35. Escobar GJ, Gardner MN, Greene JD, Draper D, Kipnis P. Risk-adjusting hospital mortality using a comprehensive electronic record in an integrated health care delivery system. *Med Care.* 2013;51(5):446–453.
36. Escobar GJ, LaGuardia JC, Turk BJ, Ragins A, Kipnis P, Draper D. Early detection of impending physiologic deterioration among patients who are not in intensive care: development of predictive models using data from an automated electronic medical record. *J Hosp Med Off Publ Soc Hosp Med.* 2012;7(5):388–395.
37. Goldstein BA, Chang TI, Mitani AA, Assimes TL, Winkelmayer WC. Near-term prediction of sudden cardiac death in older hemodialysis patients using electronic health records. *Clin J Am Soc Nephrol.* 2014;9(1):82–91.
38. Gultepe E, Green JP, Nguyen H, Adams J, Albertson T, Tagkopoulos I. 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.* 2014;21(2):315–325.
39. Gupta S, Tran T, Luo W, *et al.* Machine-learning prediction of cancer survival: a retrospective study using electronic administrative records and a cancer registry. *BMJ Open.* 2014;4(3):e004007.
40. Hao S, Jin B, Shin AY, *et al.* Risk prediction of emergency department revisit 30 days post discharge: a prospective study. *PLoS One.* 2014;9(11):e112944.
41. Hebert C, Du H, Peterson LR, Robicsek A. Electronic health record-based detection of risk factors for Clostridium difficile infection relapse. *Infect Control Hosp Epidemiol Off J Soc Hosp Epidemiol Am.* 2013;34(4):407–414.
42. Hebert C, Shivade C, Foraker R, *et al.* Diagnosis-specific readmission risk prediction using electronic health data: a retrospective cohort study. *BMC Med Inform Decis Mak.* 2014;14:65.
43. Himes BE, Dai Y, Kohane IS, Weiss ST, Ramoni MF. Prediction of chronic obstructive pulmonary disease (COPD) in asthma patients using electronic medical records. *J Am Med Inform Assoc.* 2009;16(3):371–379.
44. Hivert M-F, Dusseault-Bélanger F, Cohen A, Courteau J, Vanasse A. Modified metabolic syndrome criteria for identification of patients at risk of developing diabetes and coronary heart diseases: longitudinal assessment via electronic health records. *Can J Cardiol.* 2012;28(6):744–749.
45. Hong KJ, Shin SD, Ro YS, Song KJ, Singer AJ. Development and validation of the excess mortality ratio-based Emergency Severity Index. *Am J Emerg Med.* 2012;30(8):1491–1500.
46. Huang SH, LePendu P, Iyer SV, Tai-Seale M, Carrell D, Shah NH. Toward personalizing treatment for depression: predicting diagnosis and severity. *J Am Med Inform Assoc.* 2014;21(6):1069–1075.
47. Huang Y, Hanauer DA. Patient no-show predictive model development using multiple data sources for an effective overbooking approach. *Appl Clin Inform.* 2014;5(3):836–860.
48. Jin SJ, Kim M, Yoon JH, Song YG. A new statistical approach to predict bacteremia using electronic medical records. *Scand J Infect Dis.* 2013;45(9):672–680.
49. Johnson ES, Weinstein JR, Thorp ML, *et al.* Predicting the risk of hyperkalemia in patients with chronic kidney disease starting lisinopril. *Pharmacoevidenciol Drug Saf.* 2010;19(3):266–272.
50. Karnik S, Tan SL, Berg B, *et al.* Predicting atrial fibrillation and flutter using electronic health records. *Conf Proc Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Conf.* 2012;2012:5562–5565.
51. Kawaler E, Cobian A, Peissig P, Cross D, Yale S, Craven M. Learning to predict post-hospitalization VTE risk from EHR data. *AMIA Annu Symp Proc AMIA Symp AMIA Symp.* 2012;2012:436–445.



52. Kennedy EH, Wiitala WL, Hayward RA, Sussman JB. Improved cardiovascular risk prediction using nonparametric regression and electronic health record data. *Med Care*. 2013;51(3):251–258.
53. Khan A, Malone ML, Pagel P, Vollbrecht M, Baumgardner DJ. An electronic medical record-derived real-time assessment scale for hospital readmission in the elderly. *WMJ Off Publ State Med Soc Wis*. 2012;111(3):119–123.
54. Kontio E, Airola A, Pahikkala T, et al. Predicting patient acuity from electronic patient records. *J Biomed Inform*. 2014;51:35–40.
55. Kor DJ, Warner DO, Alsara A, et al. Derivation and diagnostic accuracy of the surgical lung injury prediction model. *Anesthesiology*. 2011;115(1):117–128.
56. Liu V, Kipnis P, Gould MK, Escobar GJ. Length of stay predictions: improvements through the use of automated laboratory and comorbidity variables. *Med Care*. 2010;48(8):739–744.
57. Mani S, Chen Y, Elasy T, Clayton W, Denny J. Type 2 diabetes risk forecasting from EMR data using machine learning. *AMIA Annu Symp Proc AMIA Symp AMIA Symp*. 2012;2012:606–615.
58. Mani S, Ozdas A, Aliferis C, et al. Medical decision support using machine learning for early detection of late-onset neonatal sepsis. *J Am Med Inform Assoc*. 2014;21(2):326–336.
59. Matheny ME, Miller RA, Iklizer TA, et al. Development of inpatient risk stratification models of acute kidney injury for use in electronic health records. *Med Decis Mak Int J Soc Med Decis Mak*. 2010;30(6):639–650.
60. Mathias JS, Agrawal A, Feinglass J, Cooper AJ, Baker DW, Choudhary A. Development of a 5 year life expectancy index in older adults using predictive mining of electronic health record data. *J Am Med Inform Assoc*. 2013;20(e1):e118–e124.
61. Meyfroidt G, Güiza F, Cottam D, et al. Computerized prediction of intensive care unit discharge after cardiac surgery: development and validation of a Gaussian processes model. *BMC Med Inform Decis Mak*. 2011;11:64.
62. Monsen KA, Swanberg HL, Oancea SC, Westra BL. Exploring the value of clinical data standards to predict hospitalization of home care patients. *Appl Clin Inform*. 2012;3(4):419–436.
63. Navarro AE, Enguñanos S, Wilber KH. Identifying risk of hospital readmission among Medicare aged patients: an approach using routinely collected data. *Home Health Care Serv Q*. 2012;31(2):181–195.
64. Nijhawan AE, Clark C, Kaplan R, Moore B, Halm EA, Amarasingham R. An electronic medical record-based model to predict 30-day risk of readmission and death among HIV-infected inpatients. *J Acquir Immune Defic Syndr* 1999. 2012;61(3):349–358.
65. O'Leary EA, Desale SY, Yi WS, et al. Letting the sun set on small bowel obstruction: can a simple risk score tell us when nonoperative care is inappropriate? *Am Surg*. 2014;80(6):572–579.
66. Perkins RM, Rahman A, Bucaloiu ID, et al. Readmission after hospitalization for heart failure among patients with chronic kidney disease: a prediction model. *Clin Nephrol*. 2013;80(6):433–440.
67. Puttkammer N, Zeliadt S, Balan JG, et al. Development of an electronic medical record based alert for risk of HIV treatment failure in a low-resource setting. *PLoS One*. 2014;9(11):e112261.
68. Ramchandran KJ, Shega JW, Von Roenn J, et al. A predictive model to identify hospitalized cancer patients at risk for 30-day mortality based on admission criteria via the electronic medical record. *Cancer*. 2013;119(11):2074–2080.
69. Rana S, Tran T, Luo W, Phung D, Kennedy RL, Venkatesh S. Predicting unplanned readmission after myocardial infarction from routinely collected administrative hospital data. *Aust Health Rev Publ Aust Hosp Assoc*. 2014;38(4):377–382.
70. Rapsomaniki E, Shah A, Perel P, et al. Prognostic models for stable coronary artery disease based on electronic health record cohort of 102 023 patients. *Eur Heart J*. 2014;35(13):844–852.
71. Robbins GK, Johnson KL, Chang Y, et al. Predicting virologic failure in an HIV clinic. *Clin Infect Dis Off Publ Infect Dis Soc Am*. 2010;50(5):779–786.
72. Robicsek A, Beaumont JL, Wright M-O, Thomson RB, Kaul KL, Peterson LR. Electronic prediction rules for methicillin-resistant *Staphylococcus aureus* colonization. *Infect Control Hosp Epidemiol Off J Soc Hosp Epidemiol Am*. 2011;32(1):9–19.
73. Rothman MJ, Rothman SI, Beals J. Development and validation of a continuous measure of patient condition using the Electronic Medical Record. *J Biomed Inform*. 2013;46(5):837–848.
74. Roubinian NH, Murphy EL, Swain BE, et al.; NHLBI Recipient Epidemiology and Donor Evaluation Study-III (REDS-III), and Northern California Kaiser Permanente DOR Systems Research Initiative. Predicting red blood cell transfusion in hospitalized patients: role of hemoglobin level, comorbidities, and illness severity. *BMC Health Serv Res*. 2014;14:213.
75. Sairamesh J, Rajagopal R, Nemana R, Argenbright K. Early warning and risk estimation methods based on unstructured text in electronic medical records to improve patient adherence and care. *AMIA Annu Symp Proc AMIA Symp AMIA Symp*. 2009;2009:553–557.
76. Sho S, Neal MD, Sperry J, Hackam DJ. A novel scoring system to predict the development of necrotizing enterocolitis totalis in premature infants. *J Pediatr Surg*. 2014;49(7):1053–1056.
77. Singal AG, Rahimi RS, Clark C, et al. An automated model using electronic medical record data identifies patients with cirrhosis at high risk for readmission. *Clin Gastroenterol Hepatol Off Clin Pract J Am Gastroenterol Assoc*. 2013;11(10):1335–1341.e1.
78. Skevofilakas M, Zarkogianni K, Karamanos BG, Nikita KS. A hybrid Decision Support System for the risk assessment of retinopathy development as a long term complication of type 1 diabetes mellitus. *Conf Proc Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Annu Conf*. 2010;2010:6713–6716.
79. Smith DH, Johnson ES, Thorp ML, et al. Predicting poor outcomes in heart failure. *Perm J*. 2011;15(4):4–11.
80. Smits FT, Brouwer HJ, van Weert HCP, Schene AH, ter Riet G. Predictability of persistent frequent attendance: a historic 3-year cohort study. *Br J Gen Pract JR Coll Gen Pract*. 2009;59(559):e44–e50.
81. Smolin B, Levy Y, Sabbach-Cohen E, Levi L, Mashiah T. Predicting mortality of elderly patients acutely admitted to the Department of Internal Medicine. *Int J Clin Pract*. 2015;69(4):501–508.
82. Still CD, Wood GC, Benotti P, et al. Preoperative prediction of type 2 diabetes remission after Roux-en-Y gastric bypass surgery: a retrospective cohort study. *Lancet Diabetes Endocrinol*. 2014;2(1):38–45.
83. Sun J, McNaughton CD, Zhang P, et al. Predicting changes in hypertension control using electronic health records from a chronic disease management program. *J Am Med Inform Assoc*. 2014;21(2):337–344.
84. Tabak YP, Sun X, Hyde L, Yaitanes A, Derby K, Johannes RS. Using enriched observational data to develop and validate age-specific mortality risk adjustment models for hospitalized pediatric patients. *Med Care*. 2013;51(5):437–445.
85. Tabak YP, Sun X, Johannes RS, Hyde L, Shorr AF, Lindenauer PK. Development and validation of a mortality risk-adjustment model for patients hospitalized for exacerbations of chronic obstructive pulmonary disease. *Med Care*. 2013;51(7):597–605.
86. Tabak YP, Sun X, Nunez CM, Johannes RS. Using electronic health record data to develop inpatient mortality predictive model: Acute Laboratory Risk of Mortality Score (ALaRMS). *J Am Med Inform Assoc*. 2014;21(3):455–463.
87. Tescher AN, Branda ME, Byrne TJO, Naessens JM. All at-risk patients are not created equal: analysis of Braden pressure ulcer risk scores to identify specific risks. *J Wound Ostomy Cont Nurs Off Publ Wound Ostomy Cont Nurses Soc*. 2012;39(3):282–291.
88. Tran T, Luo W, Phung D, et al. Risk stratification using data from electronic medical records better predicts suicide risks than clinician assessments. *BMC Psychiatry*. 2014;14:76.
89. Uyar A, Bener A, Ciray HN. Predictive modeling of implantation outcome in an in vitro fertilization setting: an application of machine learning methods. *Med Decis Mak Int J Soc Med Decis Mak*. 2014.

90. Wang L, Porter B, Maynard C, *et al*. Predicting risk of hospitalization or death among patients with heart failure in the veterans health administration. *Am J Cardiol*. 2012;110(9):1342–1349.
91. Wang L, Porter B, Maynard C, *et al*. Predicting risk of hospitalization or death among patients receiving primary care in the Veterans Health Administration. *Med Care*. 2013;51(4):368–373.
92. Wells BJ, Roth R, Nowacki AS, *et al*. Prediction of morbidity and mortality in patients with type 2 diabetes. *Peer J*. 2013;1:e87.
93. Westra BL, Savik K, Oancea C, Choromanski L, Holmes JH, Bliss D. Predicting improvement in urinary and bowel incontinence for home health patients using electronic health record data. *J Wound Ostomy Cont Nurs Off Publ Wound Ostomy Cont Nurses Soc*. 2011;38(1):77–87.
94. Woller SC, Stevens SM, Jones JP, *et al*. Derivation and validation of a simple model to identify venous thromboembolism risk in medical patients. *Am J Med*. 2011;124(10):947–954.e2.
95. Wu J, Roy J, Stewart WF. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. *Med Care*. 2010;48(6 Suppl):S106–S113.
96. Zhai H, Brady P, Li Q, *et al*. Developing and evaluating a machine learning based algorithm to predict the need of pediatric intensive care unit transfer for newly hospitalized children. *Resuscitation*. 2014;85(8):1065–1071.
97. Zhao D, Weng C. Combining PubMed knowledge and EHR data to develop a weighted bayesian network for pancreatic cancer prediction. *J Biomed. Inform*. 2011;44(5):859–868.
98. Churpek MM, Yuen TC, Edelson DP. Predicting clinical deterioration in the hospital: the impact of outcome selection. *Resuscitation*. 2013;84(5):564–568.
99. Churpek MM, Yuen TC, Park SY, Meltzer DO, Hall JB, Edelson DP. Derivation of a cardiac arrest prediction model using ward vital signs\*. *Crit Care Med*. 2012;40(7):2102–2108.
100. Churpek MM, Yuen TC, Winslow C, *et al*. Multicenter development and validation of a risk stratification tool for ward patients. *Am J Respir Crit Care Med*. 2014;190(6):649–655.
101. Kotz D, Simpson CR, Viechtbauer W, van Schayck OCP, Sheikh A. Development and validation of a model to predict the 10-year risk of general practitioner-recorded COPD. *NPJ Prim Care Respir Med*. 2014;24:14011.
102. Lipsky BA, Kollef MH, Miller LG, Sun X, Johannes RS, Tabak YP. Predicting bacteremia among patients hospitalized for skin and skin-structure infections: derivation and validation of a risk score. *Infect Control Hosp Epidemiol*. 2010;31(8):828–837.
103. Lipsky BA, Weigelt JA, Sun X, Johannes RS, Derby KG, Tabak YP. Developing and validating a risk score for lower-extremity amputation in patients hospitalized for a diabetic foot infection. *Diabetes Care*. 2011;34(8):1695–1700.
104. Liu V, Kipnis P, Gould MK, Escobar GJ. Length of stay predictions: improvements through the use of automated laboratory and comorbidity variables. *Med Care*. 2010;48(8):739–744.
105. Puopolo KM, Draper D, Wi S, *et al*. Estimating the probability of neonatal early-onset infection on the basis of maternal risk factors. *Pediatrics*. 2011;128(5):e1155–e1163.
106. Saltzman JR, Tabak YP, Hyett BH, Sun X, Travis AC, Johannes RS. A simple risk score accurately predicts in-hospital mortality, length of stay, and cost in acute upper GI bleeding. *Gastrointest Endosc*. 2011;74(6):1215–1224.
107. Tabak YP, Sun X, Derby KG, Kurtz SG, Johannes RS. Development and validation of a disease-specific risk adjustment system using automated clinical data. *Health Serv Res*. 2010;45(6 Pt 1):1815–1835.
108. Hippisley-Cox J, Coupland C, Brindle P. Derivation and validation of QStroke score for predicting risk of ischaemic stroke in primary care and comparison with other risk scores: a prospective open cohort study. *BMJ*. 2013;346:f2573.
109. Hippisley-Cox J, Coupland C, Robson J, Brindle P. Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: cohort study using QResearch database. *BMJ*. 2010;341:c6624.
110. Hippisley-Cox J, Coupland C, Robson J, Sheikh A, Brindle P. Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore. *BMJ*. 2009;338:b880.
111. Hippisley-Cox J, Coupland C. Predicting risk of osteoporotic fracture in men and women in England and Wales: prospective derivation and validation of QFractureScores. *BMJ*. 2009;339:b4229.
112. Hippisley-Cox J, Coupland C. Predicting the risk of chronic Kidney Disease in men and women in England and Wales: prospective derivation and external validation of the QKidney Scores. *BMC Fam Pract*. 2010;11:49.
113. Hippisley-Cox J, Coupland C. Development and validation of risk prediction algorithm (QThrombosis) to estimate future risk of venous thromboembolism: prospective cohort study. *BMJ*. 2011;343:d4656.
114. Hippisley-Cox J, Coupland C. Identifying patients with suspected gastro-oesophageal cancer in primary care: derivation and validation of an algorithm. *Br J Gen Pract J R Coll Gen Pract*. 2011;61(592):e707–e714.
115. Hippisley-Cox J, Coupland C. Identifying patients with suspected lung cancer in primary care: derivation and validation of an algorithm. *Br J Gen Pract J R Coll Gen Pract*. 2011;61(592):e715–e723.
116. Hippisley-Cox J, Coupland C. Derivation and validation of updated QFracture algorithm to predict risk of osteoporotic fracture in primary care in the United Kingdom: prospective open cohort study. *BMJ*. 2012;344:e3427.
117. Hippisley-Cox J, Coupland C. Identifying women with suspected ovarian cancer in primary care: derivation and validation of algorithm. *BMJ*. 2012;344:d8009.
118. Hippisley-Cox J, Coupland C. Identifying patients with suspected colorectal cancer in primary care: derivation and validation of an algorithm. *Br J Gen Pract J R Coll Gen Pract*. 2012;62(594):e29–e37.
119. Hippisley-Cox J, Coupland C. Identifying patients with suspected pancreatic cancer in primary care: derivation and validation of an algorithm. *Br J Gen Pract J R Coll Gen Pract*. 2012;62(594):e38–e45.
120. Hippisley-Cox J, Coupland C. Identifying patients with suspected renal tract cancer in primary care: derivation and validation of an algorithm. *Br J Gen Pract J R Coll Gen Pract*. 2012;62(597):e251–e260.
121. Hippisley-Cox J, Coupland C. Predicting risk of emergency admission to hospital using primary care data: derivation and validation of QAdmissions score. *BMJ Open*. 2013;3(8):e003482.
122. Hippisley-Cox J, Coupland C. Symptoms and risk factors to identify men with suspected cancer in primary care: derivation and validation of an algorithm. *Br J Gen Pract J R Coll Gen Pract*. 2013;63(606):e1–e10.
123. Hippisley-Cox J, Coupland C. Symptoms and risk factors to identify women with suspected cancer in primary care: derivation and validation of an algorithm. *Br J Gen Pract J R Coll Gen Pract*. 2013;63(606):e11–e21.
124. Hippisley-Cox J, Coupland C. Predicting risk of upper gastrointestinal bleed and intracranial bleed with anticoagulants: cohort study to derive and validate the QBleed scores. *BMJ*. 2014;349:g4606.
125. Siontis GCM, Tzoulaki I, Castaldi PJ, Ioannidis JPA. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol*. 2015;68(1):25–34.
126. Wiens J, Guttig J, Horvitz E. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. *J Am Med Inform Assoc*. 2014;21(4):699–706.
127. Liu J, Hong Y, D'Agostino RB, *et al*. Predictive value for the Chinese population of the Framingham CHD risk assessment tool compared with the Chinese Multi-Provincial Cohort Study. *JAMA*. 2004;291(21):2591–2599.
128. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. *J Am Med Inform Assoc*. 2014;21(4):578–582.

129. Henderson R, Diggle P, Dobson A. Joint modelling of longitudinal measurements and event time data. *Biostat Oxf Engl*. 2000;1(4):465–480.
130. Kharrazi H, Wang C, Scharfstein D. Prospective EHR-based clinical trials: the challenge of missing data. *J Gen Intern Med*. 2014;29(7):976–978.
131. Rusanov A, Weiskopf NG, Wang S, Weng C. Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research. *BMC Med Inform Decis Mak*. 2014;14:51.
132. Goldstein BA, Bhavsar NA, Phelan M, Pencina MJ. Controlling for informed presence bias due to the number of health encounters in an Electronic Health Record. *Am J Epidemiol*. (in press).
133. Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. *Clin Chem*. 2008;54(1):17–23.
134. Mandic S, Go C, Aggarwal I, Myers J, Froelicher VF. Relationship of predictive modeling to receiver operating characteristics. *J Cardiopulm Rehabil Prev*. 2008;28(6):415–419.
135. Gottesman O, Kuivaniemi H, Tromp G, *et al.*; eMERGE Network. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet Med Off J Am Coll Med Genet*. 2013;15(10):761–771.
136. Goldstein BA, Knowles JW, Salfati E, Ioannidis JPA, Assimes TL. Simple, standardized incorporation of genetic risk into non-genetic risk prediction tools for complex traits: coronary heart disease as an example. *Front Genet*. 2014;5:254.
137. Kho AN, Rasmussen LV, Connolly JJ, *et al.* Practical challenges in integrating genomic data into the electronic health record. *Genet Med Off J Am Coll Med Genet*. 2013;15(10):772–778.
138. Ury AG. Storing and interpreting genomic information in widely deployed electronic health record systems. *Genet Med Off J Am Coll Med Genet*. 2013;15(10):779–785.
139. Van Driest SL, Wells QS, Stallings S, *et al.* Association of arrhythmia-related genetic variants with phenotypes documented in electronic medical records. *JAMA*. 2016;315(1):47–57.
140. Feero WG. Establishing the clinical validity of arrhythmia-related genetic variations using the electronic medical record: a valid take on precision medicine? *JAMA*. 2016;315(1):33–35.