Published in final edited form as:

Clin Pharmacol Ther. 2018 May ; 103(5): 899–905. doi:10.1002/cpt.861.

Identifying patients with high data completeness to improve validity of comparative effectiveness research in electronic health records data

Kueiyu Joshua Lin, MD, ScD, MPH^{1,2,3}, Daniel E. Singer, MD^{1,2}, Robert J. Glynn, PhD, ScD^{1,3}, Shawn N. Murphy, MD, PhD⁴, Joyce Lii, MS, MA¹, and Sebastian Schneeweiss, MD, ScD^{1,3}

¹Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School

²Department of Medicine, Massachusetts General Hospital, Harvard Medical School

³Department of Epidemiology, Harvard School of Public Health

⁴Department of Neurology, Massachusetts General Hospital, Harvard Medical School

Abstract

Electronic health records (EHR)-discontinuity, i.e., having medical information recorded outside of the study EHR system, is associated with substantial information bias in EHR-based comparative effectiveness research (CER). We aimed to develop and validate a prediction model identifying patients with high EHR-continuity to reduce this bias. Based on 183,739 patients aged 65 in EHRs from two US provider networks linked with Medicare claims data from 2007–2014, we quantified EHR-continuity by Mean Proportion of Encounters Captured (MPEC) by the EHR system. We built a prediction model for MPEC using one EHR systems as training and the other as validation set. Patients with top 20% predicted EHR-continuity had 3.5–5.8 fold smaller misclassification of 40 CER-relevant variables, compared to the remaining study population. The comorbidity profiles did not differ substantially by predicted EHR-continuity. These findings suggest restriction of CER to patients with high predicted EHR-continuity may confer a favorable validity to generalizability trade-off.

Keywords

electronic medical records; data linkage; comparative effectiveness research; information bias; continuity

Corresponding author: Kueiyu Joshua Lin, Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, 1620 Tremont St. Suite 3030, Boston, MA 02120, Phone: (617) 278-0930; Fax: (617) 232-8602, jklin@mgh.harvard.edu.

Conflict of interest: none declared

Author Contributions

K.J.L., D.E.S., R.J.G., S.M., and S.S. wrote the manuscript; D.E.S., R.J.G., S.M., and S.S. designed the research; K.J.L., D.E.S., R.J.G., J.L., and S.S. performed the research; K.J.L. and J.L. analyzed the data.

Introduction

Because electronic health records (EHR) data contain rich clinical information essential for comparative effectiveness research (CER), the number of CER studies using EHR databases as the primary data source has grown exponentially in the last decade and there are currently more than 50 EHR-based research networks in the US.^{1,2} However, most EHR systems in the US, with the exception of highly integrated health plans, do not comprehensively capture medical encounters across all care settings and facilities, thus may miss substantial amount of information that characterized the health state of its patient population.

We defined “*EHR-discontinuity*” or lack of “*EHR-continuity*” as “*having medical information recorded outside of the study EHR system*”. Health conditions recorded at a clinic or hospital outside of a given EHR system are “invisible” to the investigators and therefore often assumed to be absent in the study. A false assumption about EHR-continuity will likely bias CER studies through substantial misclassification of the exposure, outcome, and confounding variables. Based on EHRs from two academic provider systems in the US linked with Medicare insurance claims data, we previously found the mean capture rates of all records by a single EHR system were only 18–24%, which translated into 11–17 fold larger misclassification of key variables in those with low than high EHR capture rates.³ We also found small to modest differences in comorbidity profiles between EHR continuity and non-continuity cohorts. These findings suggest that, when using EHR as the only data source for CER, restriction to patients with high EHR-continuity is a reasonable strategy to improve study validity.

However, direct measurement of EHR data completeness requires linkage to a secondary data source, which is often not feasible for privacy and compliance concerns (e.g., sensitive identifiers are often required for reliable linkage). Therefore, it is critical to develop strategies to best remedy data incompleteness due to EHR-discontinuity in CER studies using EHR when additional data are not available. We aimed to develop and validate algorithms to identify patients with high EHR-continuity with the predictors available in a typical EHR system. We also evaluated whether high EHR-continuity patients had representative comorbidity profiles when compared to the remaining population in the EHR.

Results

Performance of the prediction model

We identified 104,403 eligible patients in EHR system 1 as the training set and 79,336 in EHR system 2 as the validation set. The mean follow-up time for the study cohort was 3.2 and 3.0 years for the training and validation sets, respectively. Based on Lasso regression with 5-fold cross-validation using data in the training set, we built a prediction model with 20 selected variables (Table 2, adjusted R squared = 0.48). The AUC for predicting the measured EHR-continuity of ≥ 60% was 0.86 in the training and 0.88 in the testing set. The predicted EHR-continuity value was highly correlated with the measured EHR-continuity both in the training and testing set (Spearman coefficient = 0.78 and 0.82, respectively).

Misclassification by predicted EHR-continuity

- a. **Discrepancy in CCS calculation:** Based on data in the first year, Figure 2 demonstrates a clear trend of decreasing discrepancy in the calculated CCS ($CCS_{\text{difference}}$) associated with increasing predicted MPEC. In the training set, the mean $CCS_{\text{difference}}$ in the lowest decile of predicted EHR-continuity (1.60, 95% CI: 1.51–1.62) was 3.8 fold greater than that for the highest decile of predicted EHR-continuity (0.41, 95% CI: 0.39–0.44). A similar pattern was observed in the validation set (Figure 2).
- b. **Sensitivity of coding 40 selected variables:** Based on data in the first year, the mean sensitivity of EHR capturing the codes for 40 selected variables when compared to the linked claims-EHR data in the highest decile of predicted EHR-continuity (0.85, 95% CI: 0.83–0.87) was 31.5 fold greater than that for the lowest decile (0.03, 95% CI: 0.02–0.03) in the training set. A similar trend was observed in the validation set and when the analysis was done for 25 co-morbidity and 15 medication use variables separately (Figure 3A).
- c. **Standardized differences in classifying 40 selected variables:** Based on data in the first year, the Mean Standardized Differences between the proportions of the 40 selected variables ($MSD_{40_variables}$) based on EHR alone vs. the linked claims-EHR data in the lowest decile of predicted EHR-continuity ($MSD_{40_variables}=0.59$, 95% CI, 0.56–0.62) was 7.9 fold greater than that in the highest predicted continuity decile ($MSD_{40_variables}=0.08$, 95% CI 0.06–0.09) in the training set. A similar trend was observed in the validation set and when the analysis was done for 25 co-morbidity and 15 medication use variables separately (Figure 3B). When the top two deciles were combined, they had $MSD_{40_variables}$ of 0.10 (0.09–0.11) in the training set and 0.10 (95% CI, 0.09–0.11) in the validation set.

Similar patterns in all the years after cohort entry

When the top two deciles were combined, they consistently had 95 CIs of $MSD_{40_variables}$ include the *a priori* cut-off of 0.1 in all subsequent years (Table 3). $Sensitivity_{40_variables}$ ranges between 0.74 and 0.86 for those in the top two deciles of predicted continuity, which were 2.4–4.8 fold greater than the corresponding estimates in the remaining population (Table S1). Therefore we defined EHR continuity cohort as “*patients with top 20% predicted MPEC*”.

Assessing representativeness of those with high EHR-continuity

We observed small to modest differences between the distribution of CCS in the top two deciles of predicted MPEC and in the rest of the population, with a MSD for all CCS categories of 0.02 in the training set (Figure 4). Similar findings were found in the validation set and in all the years following the cohort entry (Table S2).

Sensitivity analysis

When assessing EHR continuity status every 180 days rather than every 365 days, the prediction model included the same set of predictors with the resulting prediction score highly correlated with the one generated by the primary analysis (Spearman coefficient =0.99). Very similar results were observed when assessing EHR continuity status every 545 and 730 days, respectively (Table S3). After Box-Cox transformation, MPEC was less skewed (Fisher-Pearson skewness coefficient²⁰ was reduced from 1.05 to 0.14). Using the transformed MPEC, our analysis yielded a prediction score highly correlated with the original model (Spearman coefficient =0.99). Similarly, the prediction score excluding those with zero MPEC was highly correlated with the original score (Spearman coefficient =0.99). We found the MPECs with various weighting schemes for inpatient vs outpatient encounters were all highly correlated with the MPEC used in the main analysis (Figure S1).

Discussion

Based on two large US metropolitan EHR systems, we developed and validated a prediction score that was highly correlated with the measured EHR-continuity in both training and validation sets. Those with top 20% predicted EHR-continuity were found to have acceptable classification of CER relevant variables. In the validation set, patients with top 20% predicted EHR-continuity had 2.7 fold greater sensitivity and 4.0 fold smaller MSD, compared to the rest of the population.

Our findings suggest that restriction to the patient with high predicted EHR-continuity could substantially reduce the misclassification due to EHR-discontinuity. Because large-scale linkage between EHR data and a secondary database is rarely done due to privacy concerns, researchers with access only to EHR can use our model based on variables available in a typical EHR database to identify the EHR continuity cohort to improve study validity. In addition, we demonstrated only small differences in the co-morbidity profiles in patients with high vs. low EHR-continuity. These results thus suggest restricting a CER analysis to those with high EHR-continuity likely confer a favorable benefit (reducing information bias) to risk (losing generalizability) ratio.

It is important to note that our prediction model was only intended to rank patients based on predicted EHR-continuity, not to predict the absolute values of EHR-continuity metric (i.e., EHR encounter capture rates). We assume that people can be ranked based on the likelihood of receiving most care in the EHR system, which is plausible because active patients in an EHR often share some common features, such as receiving routine vaccines and screening tests, etc. In contrast, predicting the absolute values of EHR capture rates could be much more challenging. Besides, we found the top-ranked 20% of study cohort to have acceptable variable classification, based on a cut-off suggested in the context of achieving adequate confounding adjustment. As tolerance of misclassification may differ by research contexts, researchers could use the ranks based on our prediction model along with the data on mean misclassification metrics by predicted ranks to customize a high EHR-continuity cohort to meet their study needs. For instance, if a study requires a mean sensitivity of EHR capturing CER-relevant codes to be above 0.8, the investigators should select the highest decile of predicted EHR-continuity (Figure 3A).

There are some limitations. First, we provided data regarding the information bias when classifying exposure, outcome, and confounding variables due to EHR-discontinuity, but the influence of this bias on comparative estimates (e.g. relative risks.) is research question and context specific. Therefore, future investigations in a wide range of research questions with vs. without applying our approach are needed to evaluate the ultimate impact of EHR-discontinuity on CER. Second, both our training and validation sets used patients from US metropolitan provider networks. Medical seeking behaviors may be different in a rural than in a metropolitan area, and further validation in other different EHR systems is needed to confirm generalizability of our results. Next, our study cohort consisted of only those aged 65 and older. The older adults are the most critical population to investigate the impact of EHR-continuity on study validity using EHR because they often need more complex care which may not be fulfilled in one system due to resource limitations. Moreover, US integrated health systems, in which the EHR data completeness is considered sufficient, do not have representative elderly populations. Hence, the issues our approach sought to remedy are most relevant in the older adults. Nonetheless, it is important to note that our findings may not be applicable to the younger populations. Lastly, limiting to patients with high EHR-continuity will inevitably reduce study sizes and statistical power.

Conclusions

Based on two large academic EHR systems in the US, we developed and validated a prediction score to identify patients with high EHR-continuity. Patients with top 20% predicted EHR-continuity were found to have much reduced variable misclassification based on EHR alone, but their comorbidity profiles did not substantially differ from those with lower EHR-continuity. Our findings support the strategy to restrict a CER study to patients with high EHR-continuity, as the risk of losing generalizability is relatively small in comparison to the benefit of substantial misclassification reduction. These results are relevant for the majority of healthcare systems in the US that are not integrated with a payor/insurer and where EHR-discontinuity in the EHR system is likely.

Materials and Methods

Data sets

We linked longitudinal claims data from Medicare to EHR data for two medical care networks. The first network (EHR system1) consists of 1 tertiary hospital, 2 community hospitals, and 17 primary care centers. The second network (EHR system 2) includes 1 tertiary hospital, 1 community hospital, and 16 primary care centers. The EHR database contains information on patient demographics, medical diagnoses, procedures, medications, and various clinical data. The Medicare claims data contain information on demographics, enrollment start and end dates, dispensed medications and performed procedures, and medical diagnoses.⁴

Study population

Among patients aged 65 and older with at least 180 days of continuous enrollment in Medicare (including inpatient, outpatient, and prescription coverage) from 2007/1/1 to

2014/12/31, we identified those with at least one EHR encounter during their active Medicare enrollment period. The date when these criteria were met was assigned as the index (cohort entry) date after which we started the evaluation of their EHR data completeness and classification of key variables. Those with private commercial insurance and Medicare as secondary payor were excluded to ensure we have comprehensive claims data for the study population.

Study design

Whether an EHR system holds adequate data for a particular individual (so called “EHR continuity status”) may change over time because patients may seek medical care in different provider systems over time. Therefore, we allowed the EHR continuity status to change every 365 days (Figure 1). The assumption was that most patients aged 65 and older would present for a regular follow-up with records in the claims data at least annually. A short assessment period may lead to unstable estimates of the capture rates and a long period would make the continuity status less time flexible. We followed patients until the earliest of the following: 1) loss of Medicare coverage; 2) death; 3) 2014/12/31, the end of the study period.

Measurement of EHR-continuity in an EHR system

To assess EHR-continuity, we calculated **Mean Proportions of Encounters Captured (MPEC)** by the EHR data:

$$\text{MPEC} = \left(\frac{\text{Outpatient encounters recorded in EHR}}{\text{Outpatient encounters recorded in claims data}} + \frac{\text{Inpatient encounters recorded in EHR}}{\text{Inpatient encounters recorded in claims data}} \right) / 2$$

Patients generally have substantially more outpatient than inpatient visits. This formula purposefully gives higher weight to inpatient than to outpatient visits. This is consistent with usual data considerations in CER where recording of inpatient diagnosis is considered more complete and accurate than in outpatient settings.⁵ The incomplete terminal year during follow-up (with length less than 365 days) was not used to calculate MPEC to avoid unstable estimates.

Building a prediction model for having high EHR-continuity

Based on clinical knowledge, we came up with a list of candidate proxy indicators for having high EHR-continuity in the following groups: (a) International Classification of Diseases (ICD) codes for routine care; (b) Preventive interventions often associated with routine care visits; (c) Recording of diagnoses or medications in the EHR; (d) Having certain type and number of encounters in the EHR; (e) Seeing the same provider repeatedly in the system (see Table S4 for detailed definitions). Using data from EHR system 1 as the training set, we built a model predicting continuous MPEC by Lasso regression with 5-fold cross-validation.⁶ To avoid violating the independence assumption for Lasso regression, we only used the first complete calendar year for each person for building prediction model but the model performance was evaluated in the subsequent years and in the independent validation dataset.

Performance of the prediction models

- A. Discrimination and correlation with measured MPEC:** We previously found that 60% was the minimum MPEC needed to achieve acceptable classification of the selected variables according to one possible cut-off suggested in the context of confounding adjustment.^{3,7} In the training and validation set, we computed area under the receiver operating characteristic curve (AUC) using our model to predict MPEC = 60%. We also evaluated how the predicted score was correlated with the measured continuity metric, MPEC, by Spearman rank correlation coefficient.
- B. Key variables for misclassification evaluation:** (1) Combined comorbidity score (CCS), which outperformed two widely used comorbidity scores.⁸ (2) 40 selected variables commonly used as drug exposure (n=15), outcome (n=10), or confounders (n=15) in CER (Table 1 and Table S5). The 10 outcome variables were based on previously validated algorithms.^{9–11, 12,13, 14,15, 16} For each year following the index date, we evaluated the classification of all the listed variables, stratified by deciles of predicted MPEC.
- C. Metrics of misclassifications:** For each individual, we calculated CCS based on EHR alone (CCS_{EHR}) and that based on linked claims-EHR data (CCS_{full}). The difference ($CCS_{difference} = CCS_{full} - CCS_{EHR}$) represents how much a patient's CCS would be underestimated if relying on only EHR than with access to both claims and EHR data. We then computed group mean of $CCS_{difference}$ for patients with different levels of predicted MPEC. In its derivation, each 1 point increase (decrease) in CCS corresponded to a 35% increase (decrease) in the odds of dying in one year.⁸ For the 40 selected variables listed in Table 1, we quantified their misclassification by two methods: (a) Sensitivity of positive coding in EHR when compared to coding in the linked claims-EHR data:

$$Sensitivity_{variable(i)} = \frac{\# \text{ of patients with variable}(i) = 1 \text{ based on EHR alone}}{\# \text{ of patients with variable}(i) = 1 \text{ based on the linked claims plus EHR data}}$$

$i = 1 - 40$

By design, because the gold standard was classification based on all available data, specificity was expected to be 100% for all variables but if the study EHR system did not capture medical information recorded in other systems, sensitivity could be low; (b) Standardized difference comparing the classification based on only EHR vs. that based on the linked claims-EHR data: Standardized difference is a measure of distance between two group means standardized by their standard deviations. This metric is often used to assess balance of covariates for exposure groups under comparison.¹⁷ Within levels of MPEC, we computed mean sensitivity and mean standardized difference (MSD) over the 40 variables. We used a formula derived by Becker¹⁸ to construct 95% confidence intervals (CIs) of the calculated standardized differences, accounting for the correlation between the repeated variable classifications in the same population. To have a reference point, it was suggested that a standardized difference of less than 0.1 indicates

satisfactory balance of covariates in the context of achieving adequate confounding adjustment.⁷ Because reducing misclassification of confounders is one of the major pathways to improve study validity, this cut-off is relevant for our study.

Evaluation of the representativeness of the cohort with high EHR-continuity

Those within top deciles of predicted MPEC needed to achieve satisfactory variable classification (MSD <0.1) would be defined as the “*EHR continuity cohort*”. We compared the proportions of all CCS categories based on claims data in those within vs. outside of EHR continuity cohort to see if those with high predicted EHR-continuity had similar comorbidity profiles when compared to the remaining population. We used claims data for the representativeness assessment, assuming similar completeness of claims data across different levels of EHR-continuity.

Sensitivity analyses

(1) The prediction model was developed based on data in the first year following the index date. We applied the model to data in the subsequent years to see if similar patterns would be observed. (2) We evaluated if our results were sensitive to the length of EHR continuity assessment period. We compared results when assessing EHR continuity status every 180, 545, and 730 days instead of every 365 days. (3) We evaluated if our results were sensitive to outliers or skewed distribution of the measured EHR-continuity metric, MPEC. We repeated our analyses after: a) Box-Cox transformation of MPEC (aiming to achieve a distribution closer to normal distributions)¹⁹; b) excluding from the analysis those with zero MPEC. (4) We assessed if our results were sensitive to different weighting schemes for inpatient vs outpatient encounters in the MPEC formula. The statistical analyses were conducted with SAS 9.4 (SAS Institute Inc., Cary, NC).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Randhawa GS. Building electronic data infrastructure for comparative effectiveness research: accomplishments, lessons learned and future steps. *J Comp Eff Res*. 2014; 3:567–572. [PubMed: 25494561]
2. Corley DA, Feigelson HS, Lieu TA, McGlynn EA. Building Data Infrastructure to Evaluate and Improve Quality: PCORnet. *J Oncol Pract*. 2015; 11:204–206. [PubMed: 25980016]
3. Lin, KJ., Glynn, RJ., Singer, DE., Schneeweiss, S. The impact of care-discontinuity on recording patient characteristics critical for comparative effectiveness research using electronic health records. 22nd Annual International Meeting of International Society For Pharmacoeconomics and Outcomes Research; Boston, USA. 2017;
4. Hennessy S. Use of health care databases in pharmacoepidemiology. *Basic & clinical pharmacology & toxicology*. 2006; 98:311–313. [PubMed: 16611207]
5. Fang, MC., et al. Validity of Using Inpatient and Outpatient Administrative Codes to Identify Acute Venous Thromboembolism: The CVRN VTE Study; Medical care; 2016.
6. Tibshirani R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B*. 1996; 58:267–288.

7. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med*. 2009; 28:3083–3107. [PubMed: 19757444]
8. Gagne JJ, Glynn RJ, Avorn J, Levin R, Schneeweiss S. A combined comorbidity score predicted mortality in elderly patients better than existing scores. *Journal of clinical epidemiology*. 2011; 64:749–759. [PubMed: 21208778]
9. Birman-Deych E, et al. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Med Care*. 2005; 43:480–485. [PubMed: 15838413]
10. Wahl PM, et al. Validation of claims-based diagnostic and procedure codes for cardiovascular and gastrointestinal serious adverse events in a commercially-insured population. *Pharmacoepidemiology and drug safety*. 2010; 19:596–603. [PubMed: 20140892]
11. Andrade SE, et al. A systematic review of validated methods for identifying cerebrovascular accident or transient ischemic attack using administrative data. *Pharmacoepidemiology and drug safety*. 2012; 21(Suppl 1):100–128. [PubMed: 22262598]
12. Tamariz L, Harkins T, Nair V. A systematic review of validated methods for identifying venous thromboembolism using administrative and claims data. *Pharmacoepidemiology and drug safety*. 2012; 21(Suppl 1):154–162. [PubMed: 22262602]
13. Waikar SS, et al. Validity of International Classification of Diseases, Ninth Revision, Clinical Modification Codes for Acute Renal Failure. *Journal of the American Society of Nephrology : JASN*. 2006; 17:1688–1694. [PubMed: 16641149]
14. Cushman M, et al. Deep vein thrombosis and pulmonary embolism in two cohorts: the longitudinal investigation of thromboembolism etiology. *The American journal of medicine*. 2004; 117:19–25. [PubMed: 15210384]
15. Cunningham A, et al. An automated database case definition for serious bleeding related to oral anticoagulant use. *Pharmacoepidemiology and drug safety*. 2011; 20:560–566. [PubMed: 21387461]
16. Myers RP, Leung Y, Shaheen AA, Li B. Validation of ICD-9-CM/ICD-10 coding algorithms for the identification of patients with acetaminophen overdose and hepatotoxicity using administrative data. *BMC Health Serv Res*. 2007; 7:159. [PubMed: 17910762]
17. Franklin JM, Rassen JA, Ackermann D, Bartels DB, Schneeweiss S. Metrics for covariate balance in cohort studies of causal effects. *Statistics in medicine*. 2014; 33:1685–1699. [PubMed: 24323618]
18. Becker BJ. Synthesizing standardized mean change measures. *British Journal of Mathematical and Statistical Psychology*. 1988; 41:257–278.
19. Box GEPCD. An analysis of transformations revisited. *Journal of the Royal Statistical Society*. 1964; 26:211–252.
20. Newell KM, Hancock PA. Forgotten moments: a note on skewness and kurtosis as influential factors in inferences extrapolated from response distributions. *Journal of motor behavior*. 1984; 16:320–335. [PubMed: 15151856]

Study Highlights

- **What is the current knowledge on the topic?** Electronic health record (EHR)-discontinuity, i.e., receiving care outside of the study EHR, is associated with substantial information bias on variables essential for comparative effectiveness research (CER).
- **What question did this study address?** How can we reduce information bias due to EHR-discontinuity in comparative effectiveness research using EHR as the primary data source?
- **What this study adds to our knowledge?** We developed and validate a prediction rule to identify patients with high EHR-continuity and demonstrated the information bias was much reduced in this sub-cohort in an external population. The high care-continuity patients were found to have similar comorbidity profiles compared to the remaining population.
- **How this might change clinical pharmacology or translational science?** Using our algorithm to restrict a CER analysis using EHR to patients with high EHR-continuity can substantially reduce information bias while likely preserving generalizability of the study findings.

*Assess the completeness of the records captured by the EHR and classification of key variables during the same period
EHR=electronic health record

Figure 1.
Assessment of electronic health record continuity and proxy indicators

EHR= electronic health record; Combined comorbidity score¹⁰ ranges between -2 to 26 with a higher score associated with higher mortality.

Figure 2.

Mean difference in combined comorbidity score based on linked claims-EHR data vs. EHR alone

A. Quantified by sensitivity **B. Quantified by mean standardized difference (MSD)****

EHR=electronic health record;

*See list of variables in Table 1; ** MSD=Mean Standardized Difference between classification based on EHR vs. the linked Claims-EHR data

Figure 3.

Decreasing misclassification of key variables associated with increasing predicted EHR-continuity

Figure 4.

Representativeness: comparison of combined comorbidity score in those with predicted high vs. low EHR-continuity in EHR system 1

Table 1

The 40 patient characteristics to be assessed for accuracy of classification

25 co-morbidity variables	<p>a. 15 variables commonly used as covariates for confounding adjustment: Dementia, atrial fibrillation, chronic lung disease, chronic liver disease, chronic kidney disease, cancer, diabetes, hypertension, anemia, psychosis, depression, pneumonia, HIV, fracture, and RA</p> <p>b. 10 variables with validated algorithm commonly used as outcome variables: ischemic stroke, ICH, MI, AKI, hepatotoxicity, GI Bleeding, major bleeding, DVT, PE, and CHF^{9-11,12,13,14,15,16}</p>
15 medication use variables	antiplatelet agents, antidiabetics, antihypertensives, NSAIDs, opioids, antidepressants, antipsychotics, anticonvulsants, PPIs, antiarrhythmics, statins, dementia, hormone therapy, antibiotics, and oral anticoagulants

CHF= congestive heart failure, HIV= human immunodeficiency virus, RA=rheumatoid arthritis, AKI=acute kidney injury, ICH= intracranial hemorrhage, MI= myocardial infarction, PE= pulmonary embolism, DVT= deep vein thrombosis, NSAIDs= nonsteroidal anti-inflammatory drugs, PPI= Proton pump inhibitors

Table 2

Prediction model for EHR continuity cohort

Variable	Coefficient
Intercept	−0.010
Having seen the same provider twice	0.049
Having seen the same provider ≥ 3 times	0.087
Having general medical exam*	0.078
Mammography*	0.075
Pap smear*	0.009
PSA Test*	0.103
Colonoscopy*	0.064
Fecal occult blood test*	0.034
Influenza vaccine*	0.102
Pneumococcal vaccine*	0.031
Having BMI recorded*	0.017
Having 2 of the above routine care facts**	0.049
With any one medication use record	0.002
With at least 2 medication use records	0.074
Having A1C ordered or value recorded*	0.018
Having at least one inpatient or outpatient encounter	0.091
Having at least two outpatient encounters	0.050
With 1 diagnosis recorded in the EHR	−0.026
With at least 2 diagnoses recorded in the EHR	0.037
Having any ED visit in the EHR	0.078

** having 2 of the facts followed by *, EHR=electronic health records, PSA= prostate specific antigen

Table 3

Difference in measuring 40 selected variables in EHR vs. EHR plus claims data

Training set							
Year after cohort entry	1	2	3	4	5	6	7
	Mean standardized difference* (95% CI)						
Top two deciles of predicted EHR-continuity	0.10 (0.09-0.11)	0.10 (0.09-0.11)	0.10 (0.09-0.11)	0.10 (0.09-0.11)	0.09 (0.07-0.10)	0.07 (0.06-0.09)	0.06 (0.05-0.08)
The remaining population	0.36 (0.35-0.37)	0.41 (0.40-0.42)	0.41 (0.40-0.43)	0.42 (0.40-0.43)	0.40 (0.39-0.41)	0.38 (0.36-0.40)	0.35 (0.33-0.37)
Validation set							
Year after cohort entry	1	2	3	4	5	6	7
	Mean standardized difference* (95% CI)						
Top two deciles of predicted EHR-continuity	0.10 (0.09-0.11)	0.11 (0.10-0.12)	0.11 (0.09-0.12)	0.11 (0.09-0.13)	0.09 (0.07-0.11)	0.08 (0.06-0.10)	0.07 (0.05-0.09)
The remaining population	0.40 (0.39-0.41)	0.48 (0.47-0.49)	0.49 (0.48-0.51)	0.49 (0.47-0.51)	0.47 (0.45-0.49)	0.46 (0.43-0.48)	0.41 (0.38-0.43)

CI=confidence interval, EHR= electronic health records.

* a mean standardized difference (MSD) indicates acceptable discrepancy⁷; Estimates with 95 CIs included 0.1 were marked in boldface.