Published in final edited form as:
J Biomed Inform. 2018 February ; 78: 54–59. doi:10.1016/j.jbi.2017.12.017.

Yield and Bias In Defining a Cohort Study Baseline from Electronic Health Record Data

Jason L. Vassy, MD, MPH, SM[a,b,d], **Yuk-Lam Ho, MPH**[a], **Jacqueline Honerlaw, RN, BSN, MPH**[a], **Kelly Cho, PhD, MPH**[a,c,d], **J. Michael Gaziano, MD, MPH**[a,c,d], **Peter W. F. Wilson, MD**[e,f], and **David R. Gagnon, MD, MPH, PhD**[a,g]

[a]VA Boston Healthcare System, Boston, MA, USA

[b]Division of General Internal Medicine, Brigham and Women's Hospital, Boston, MA, USA

[c]Division of Aging, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA

[d]Department of Medicine, Harvard Medical School, Boston, MA, USA

[e]Atlanta VA Medical Center, Atlanta, GA, USA

[f]Emory University Schools of Medicine and Public Health, Atlanta, GA, USA

[g]Boston University School of Public Health, Boston, MA, USA

Abstract

Aims—Despite growing interest in using electronic health records (EHR) to create longitudinal cohort studies, the distribution and missingness of EHR data might introduce selection bias and information bias to such analyses. We aimed to examine the yield and potential for these healthcare process biases in defining a study baseline using EHR data, using the example of cholesterol and blood pressure (BP) measurements.

Methods—We created a virtual cohort study of cardiovascular disease (CVD) from patients with eligible cholesterol profiles in the New England (NE) and Southeast (SE) networks of the Veterans Health Administration in the United States. Using clinical data from the EHR, we plotted the yield of patients with BP measurements within an expanding timeframe around an index date of cholesterol testing. We compared three groups: 1) patients with BP from the exact index date; 2) patients with BP not on the index date but within the network-specific 90th percentile around the index date; and 3) patients with no BP within the network-specific 90th percentile.

Results—Among 589,361 total patients in the two networks, 146,636 (61.0%) of 240,479 patients from NE and 289,906 (83.1%) of 348,882 patients from SE had BP measurements on the index date. Ninety percent had BP measured within 11 days of the index date in NE and within 5 days of the index date in SE. Group 3 in both networks had fewer available race data, fewer comorbidities and CVD medications, and fewer health system encounters.

Corresponding Author: Jason L. Vassy, MD, MPH, SM, VA Boston Healthcare System, 150 South Huntington Avenue, 152-G, Boston, MA 02130, Telephone: 857-364-2561; Fax 857-364-6990, jvassy@partners.org.

7. CONFLICT OF INTEREST

None declared.

Conclusions—Requiring same-day risk factor measurement in the creation of a virtual CVD cohort study from EHR data might exclude 40% of eligible patients, but including patients with infrequent visits might introduce bias. Data visualization can inform study-specific strategies to address these challenges for the research use of EHR data.

1. INTRODUCTION

Classic prospective cohort studies, such as the Framingham Heart Study and Nurses' Health Study,[1, 2] begin with a single baseline visit, when disease risk factors are measured with surveys, physical examinations, laboratory analyses, and other data collection methods. Participants are then observed at regular intervals over time for outcomes such as cardiovascular disease (CVD) events or cancer incidence. Such studies have made immeasurable contributions to understanding health and disease. However, as research budgets tighten, the creation of new cohort studies may no longer be sustainable for epidemiologic discovery.[3]

Electronic health records (EHR) have potential to fill this gap.[4–7] Concurrent with the decline in research funding is an increasing emphasis on the implementation and meaningful use of EHR in patient care.[8, 9] The wealth of data in EHR systems can be a cost-effective alternative to large cohort studies and surpass their scale by orders of magnitude. Moreover, EHR contain data unavailable from any other source, since they document the processes and outcomes of invasive data collection methods and other procedures. However, the creation of cohort studies from EHR data represents a use that is secondary to their primary purposes of clinical care and healthcare administration.[10] As a result, the content and frequency of data collection that is appropriate for these primary uses can introduce bias when the data are used for research.[11–14] Bias exists when the association observed between an exposure and outcome is not entirely due to causality[15] and can threaten the internal validity of a study's findings. Although dozens of biases have been described, most can be assigned to one of three categories: selection bias, information bias, and confounding.[16] Several features of research in EHR data make it susceptible to bias.[14, 17]

In this paper, we focus on one choice researchers make when creating a cohort study from EHR data: how to define each individual's baseline. Patients in a health system access care at varying frequencies and might therefore have incomplete ascertainment of a full set of risk factors at any one timepoint. This variable schedule of contact with the health system might introduce bias to any research using an EHR-defined cohort. Limiting the analyses only to those patients with complete baseline data at one timepoint might decrease the yield of eligible patients and introduce selection bias, error in cohort selection that favors patients who access care most often.[14, 15, 18] However, expanding the definition of the baseline timeframe might introduce information bias, error in cohort characterization,[16, 18] if patients with less frequent contact are more likely to have inaccurate measurements of their risk factors at any one instance. Together, these errors that result from the way data are collected in EHR have been termed healthcare process bias.[19, 20]

Examining the distribution of relevant EHR data can help investigators devise strategies to minimize this risk of bias.[21–23] We are using existing EHR data to define a "virtual"

cohort study of CVD risk. That is, we are seeking to define a cohort of patients from EHR records, identify a baseline "visit" at which relevant risk factors were measured, and then determine the association between those risk factors and subsequent CVD events. In this manuscript, we describe our approach to examining the impact of widening the baseline timeframe on the yield of eligible patient data, using the example of a laboratory measurement (cholesterol) and a vital sign (blood pressure, BP). We used data visualization to determine the impact of the variable rates of risk factor measurement on the yield and potential for bias when using EHR data for clinical epidemiology, examining whether patients with greater time between cholesterol and BP measurements are systematically different than those with narrower baselines. Specifically, we examined whether the time between cholesterol and BP measurements was associated with other important patient characteristics, including demographics, comorbidities, and laboratory values.

2. MATERIALS AND METHODS

2.1 Patient population

The Veterans Health Administration (VA) is an integrated health system caring for more than 8 million military veterans across the United States.[24] Veterans may receive VA healthcare if they meet certain requirements related to military service, disability, and income, and more than 90% of VA users are men. The health system is organized into of 23 regional Veterans Integrated Service Networks (VISN), each comprised of large medical centers, ambulatory centers, and community-based clinics with a shared electronic health record. We accessed patient data from the New England Network (VISN 1, NE), encompassing six states and caring for 981,000 patients, and the Southeast Network (VISN 7, SE), covering South Carolina and parts of Georgia and Alabama and caring for 1,458,000 patients. Because we are creating 10-year CVD models, we sought to identify baseline visits between 2000–2007. All NE and SE patients with at least one outpatient lipid profile (see below) were eligible, due to the centrality of lipids in CVD risk estimation.[25]

2.2 Data sources

All data came from the electronic VA corporate data warehouse (CDW),[26] which includes administrative data and clinical data from the single EHR (the Computerized Patient Record System, CPRS) used at all VA locations nationwide. Demographic data such as sex and race in the CDW are derived from administrative data sources. Clinical data such as International Classification of Disease (ICD) codes and vital signs including blood pressure (BP) are entered by care team members into non-redundant structured fields in the EHR at the point of care in both outpatient and inpatient settings. Medication data in the CDW are derived from VA pharmacy records, which contain structured data about outpatient and inpatient prescriptions ordered by VA providers and dispensed by VA pharmacies. The CDW also contains structured data on medications patients receive from outside the VA, entered by VA providers at structured data in the EHR. Laboratory results in the CDW originate from VA the laboratory information management system.

For each patient, a baseline index date was assigned as described below. We used CDW data from the baseline index date to identify patient characteristics, including gender, race, and

ethnicity. Insurance status was taken from the outpatient visit closest to the index date and was categorized as VA only, VA plus private insurance, or VA plus other government insurance (including Medicare, Medicaid, and Civilian Health and Medical Program of the Uniformed Services [CHAMPUS]). Age was calculated as the difference between the index date and date of birth. Estimated glomerular filtration rate was calculated from the serum creatinine using the CKD-EPI method.[27] Smoking status was determined with a validated method developed in VA data as described previously.[28] We used ICD-9 codes before the index date to identify baseline CVD: acute and chronic ischemic heart disease (410.0–414.9), cerebrovascular disease (430–438), and peripheral vascular disease (443.9, 440.20–440.4). We identified diabetes by ICD-9 codes 250.00–250.93). Hypertension treatment was identified by an active prescription on the index date for at least one of the following medication classes, taken as a single agent or in combination formulations: diuretics, angiotensin-converting enzyme (ACE) inhibitors, angiotensin II receptor inhibitors, beta-blockers, and calcium channel blockers. Similarly, cholesterol-lowering treatment was identified by an active prescription for any of the following medications, taken as single agents or in combination formulations: statins, niacin, ezetimibe, gemfibrozil, fenofibrates, cholestyramine, and colestipol. Proprotein convertase subtilisin kexin 9 (PCSK9) inhibitors were not clinically available during the baseline period. Outpatient encounters included any day with at least one documented health system contact.

2.3 Analysis

For each patient, we anchored a baseline index date to the date of the first outpatient lipid results between 2000 and 2007. We chose lipid testing as the index date because of the centrality of cholesterol, including low-density lipoprotein (LDL), high-density lipoprotein (HDL), and triglycerides, in CVD risk estimation[25] and because patient lipid levels are generally measured less frequently than BP, a vital sign routinely collected in clinical care. Because we aimed to create a prospective cohort study for incident CVD events, we used outpatient, rather than inpatient, lipid measurements to identify baseline risk factors before, not at the time of, a CVD event, a common cause of hospitalization. In this health system, BP is generally measured at clinical appointments with providers such as primary care practitioners or subspecialists and not at encounters such as laboratory testing or optometry appointments.

We examined whether CVD risk factors varied by the time between CVD risk factor measurements in two different ways. First, we determined the proportion of patients with BP measured on the same date as their index date of lipid testing. We then expanded the definition of the baseline timeframe by one-day intervals before or after this date, plotting the cumulative proportion of eligible patients with BP measured (yield) with successive widening of the timeframe. Informed by this visualization of the data distribution, we compared the demographic and clinical characteristics of three mutually exclusive groups of patients: 1) patients with BP recorded on the index date of lipid testing; 2) patients with no recorded BP on the index date but within a time window to either side of the index date that included 90% of the patients in each network; and 3) patients with no BP within the network-specific 90th percentile. Second, we used linear regression to model the association between CVD risk factors (BP and lipid levels) and the time between the index date of lipid

testing and the date of the nearest-in-time BP measurement, adjusted for age, gender, race, diabetes, and hypertension and cholesterol medications. Specifically, the dependent variables in these models were systolic blood pressure (SBP), diastolic blood pressure (DBP), and LDL, HDL, and total cholesterol. $P<0.005$ indicated statistical significance.

2.4 Ethical considerations

The VA Boston Institutional Review Board approved this study and granted a waiver of informed consent. The study conforms to the Declaration of Helsinki.

3. RESULTS

We identified 589,361 patients with an outpatient lipid measurement in the two networks (Table 1). Among these, 146,636 (61.0%) of 240,479 patients from NE and 289,906 (83.1%) of 348,882 patients from SE had BP measurements on the same date as the index date of lipid testing. These yields reached 90% when the baseline window was expanded to 154 days before the index date in NE and 14 days before the index date in SE. Alternatively, 90% of patients had a BP measurement when the baseline window was extended 91 and 7 days after the index date in NE and SE, respectively (Figure 1, left panel). When the baseline window was extended symmetrically before *and* after the index date of lipid testing, 90% of patients had a BP measurement within 11 days in NE and 5 days in SE (Figure 1, right panel). When patients were categorized based on the length of time between their index date and date of nearest BP, the three groups did not differ in age, BP, or lipid levels, but Group 3 in both networks had fewer available race data, lower recorded prevalence of comorbidities, and fewer CVD medications and outpatient contact with the health system (Table 1).

When analyzed as a continuous variable in multivariable regression models, greater time between lipid and BP measurements was generally associated with more favorable levels of these quantitative CVD risk factors themselves, although the magnitudes of these associations were too small to be clinically meaningful. Greater time between measurements was statistically significantly associated with lower SBP in NE (−0.21 mmHg per 100 days, 95% CI −0.24, −0.18) and SE (−0.20 mmHg per 100 days, 95% CI −0.23, −0.17) and lower DBP in NE (−0.12 mmHg per 100 days, 95% CI −0.14, −0.11) and SE (−0.10 mmHg per 100 days, −0.11, −0.08). Greater time between measurements was also significantly associated with lower LDL (−0.22 mg/dL per 100 days, 95% CI −0.27, −0.17) and total cholesterol (−0.37 mg/dL per 100 days, 95% CI −0.43, −0.31) in NE and with lower total cholesterol (−0.37 mg/dL per 100 days, 95% CI −0.43, −0.30) and higher HDL cholesterol (0.08 mg/dL per 100 days, 95% CI 0.05, 0.10) in SE.

4. DISCUSSION

4.1. Summary of findings

Using the example of cholesterol and BP measurements, we examined the impact of widening the timeframe that defines the virtual baseline “visit” on the yield of eligible patient data for a cohort study created from the EHR. We found that requiring same-day measurements of cholesterol and BP might exclude up to 40% of eligible patients. Widening the baseline timeframe increased the yield of eligible patients, albeit at the expense of

including patients who were systematically different than those with a narrower time between measurements. Specifically, time between cholesterol and BP measurements was associated with the risk factor levels themselves and, perhaps more problematically, with the missingness of key variables such as race and ethnicity. This observation raises the concern of misclassification, a type of information bias: if patients with less frequent risk factor measurement are more likely to have missing race and ethnicity data, their CVD and diabetes statuses might also be misclassified as negative. Examining these relationships can be a first step to addressing the resulting biases.

4.2. Contextualization with prior work

The risk for bias in using EHR for research has been well documented. [11–14, 17–19] It is known that invoking certain data sufficiency requirements selects for a sicker patient population, since these patients have more frequent data collection[14, 17]. In our cohort, shorter time between dates of lipid and BP measurements was associated with a greater number of comorbidities and with poorer BP and cholesterol values. While the magnitude of these associations is unlikely to be clinically significant, they further illustrate the non-random patterns of risk factor measurement and data missingess in our EHR-defined cohort. The relationship we observed between the measurements of a laboratory value and a vital sign are consistent with those of Pivovarov and colleagues, who found that, among 20 years of EHR data from 14,000 ambulatory internal medicine patients, the testing patterns of certain laboratory tests conveyed separate information from the test results' numerical values themselves.[29] Interestingly, however, the numerical value of LDL cholesterol testing was not associated with its frequency of testing in that study, a finding the authors attributed to healthcare processes such as guidelines for screening and monitoring. Our analyses of outpatient measurements are also consistent with results from 10,000 patients receiving anesthetic services at one medical center, where illness severity was associated with a greater number of days with clinical data[30]. The decision of how to define the baseline timeframe of an EHR-derived cohort study can be viewed as a missing data problem,[31] with competing biases at either end of the spectrum. Narrower timeframes will select a sicker population, while wider timeframes will include patients who are more likely to have misclassified exposures and outcomes.

4.3. Data visualization

Data visualization is one tool to help investigators examine and address these biases when using EHR data for research.[22, 23, 32, 33] Pivovarov and colleagues used histograms to examine laboratory testing dynamics; in doing so, they identified multimodality in testing patterns associated with, for example, inpatient versus outpatient status.[29] Baseline measurements of cholesterol and BP are central to CVD prediction in the virtual cohort study we are creating. Our V-shaped plot of the timeframe between these measurements will inform how we examine the impact of data distribution on the analytic validity of our CVD prediction models. Sensitivity analyses are classically used in epidemiology to determine the robustness of statistical associations to potential biases.[34] In our case, sensitivity analyses that include only Group 1, Groups 1 and 2, or Groups 1–3 in our cohort will enable us to determine any resulting bias in our model, indicated by any change in the magnitude of the association between each risk factor and the risk of a CVD event. Investigators creating

longitudinal cohorts from EHR data may wish to use similar plots for exposures and outcomes beyond CVD; the choice of data distribution to examine will depend on the primary hypotheses and natural history of the exposures and outcomes of interest. For example, Albers and Hripcsak have used EHR data to demonstrate diurnal variation in serum creatinine values,[17] suggesting that studies of acute changes in renal function might require methods to handle time-based signals in a narrow window. On the other hand, it may be acceptable to have greater time between baseline assessments of more stable risk factors, such as diabetes status and smoking habits in the prediction of a future CVD event. In addition, a longer follow-up time between the baseline and the incident event of interest, such as cancer diagnosis, is also likely to diminish the relative importance of the time between the baseline measurements of different risk factors. Ultimately, each team of investigators will need to decide the appropriate methods for examining and addressing potential sources of bias in their studies.

Our data plots revealed another potential source of healthcare process bias in our data: the variation in CVD risk factor data collection between two regional networks of the same national integrated health system. Ninety percent of patients had a BP measurement within 11 and 5 days of their baseline lipid measurement in NE and SE, respectively. However, when including BP measurements only taken before the index lipid test, 90% of patients in SE had an eligible BP within 14 days, while for NE, this timeframe had to be extended to 154 days to include 90% of patients. Reasons for such variation might include more frequent BP measurement in SE or more frequent utilization of non-VA healthcare in NE. Our illustration of data variability between two networks within one system suggests the likelihood of even greater variability in the EHR data of different health systems, as demonstrated elsewhere.[12] This is an important consideration for researchers combining EHR data across different health systems to power larger-scale analyses.

4.4 Addressing healthcare process bias

Recognizing the potential for healthcare process bias is the first step toward choosing strategies for mitigating that bias.[20] Strategies for handling the missing or infrequent data that result from healthcare processes can occur at the pre-analytic and analytic phases of EHR-based research. Before analysis, collateral or complementary data sources can be sought.[35] The Electronic Medical Records and Genomics (eMERGE) network, a multi-institution consortium using EHR data to define clinical phenotypes, has combined EHR and pharmacy insurance claims data to improve the detection of cases of resistant hypertension.[36] As another example, data from sources such as the National Death Index and the Centers for Medicare and Medicaid Services can supplement missing data such as medical diagnoses and race. Validating EHR-derived data against gold-standard research measurements in the same cohort might determine the confidence that can be placed in EHR data alone.[36] For example, the Million Veteran Program is using this approach to validate EHR-derived smoking status against participant survey data. In the analytic phase of research, investigators often limit analyses only to patients with a complete set of data.[14, 31] For example, the eMERGE network has defined completeness as having one biobanked sample, at least two clinical visits, and data from each of several data categories.[36] Our results and those of others suggest that such data sufficiency requirements might introduce

bias that should be examined before investigators choose this strategy over other methods to account for missing data,[14] including multiple imputation.[31] The pattern of data collection itself can be informative. This was recently illustrated in a machine learning study of serum ferritin values, in which patient demographics and other concurrent laboratory results had remarkable accuracy for predicting high ferritin levels and in some cases predicted iron-deficiency anemia more accurately than measured ferritin.[37] Additional approaches such as lagged regression models[21] and other temporal-informed methods[38–41] [42] [43] can be used to account for healthcare process bias and improve the specificity and sensitivity of time-based analyses in EHR data.

4.5 Limitations

Our analyses have a few limitations to note. First, they are limited to patients receiving care from the Veterans Health Administration, a patient population with fewer women and greater burden of physical and psychological comorbidities than many health systems. Although our specific quantitative results about CVD risk factor measurement may not generalize to EHR data from other health systems, we believe our approach to visualizing and addressing the potential for healthcare process biases might. Second, in this exercise, we chose to illustrate the temporal relationship between the measurements of only two important clinical CVD factors, BP and cholesterol. Of course, CVD risk depends on numerous other factors, including clinical variables such as body-mass index and renal function and lifestyle factors such as smoking, diet, and exercise. Third, these analyses are limited by the lack of CVD outcome data, but as we define our cohort study and collect CVD outcomes, sensitivity analyses as described above will inform the degree to which variable data distribution might bias our prediction models.

5. CONCLUSIONS

In conclusion, plotting the timeline of data distribution helps researchers assess and mitigate the inherent potential for healthcare process bias when using EHR for research or other secondary purposes.

Acknowledgments

This work was supported by VA Merit Award I01-CX001025. JLV is supported by NIH L30-DK089597, KL2-TR001100, and Career Development Award IK2-CX001262 from the United States Department of Veterans Affairs Clinical Sciences Research and Development Service.

References

1. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation. 1998; 97:1837–47. [PubMed: 9603539]
2. Fiebach NH, Hebert PR, Stampfer MJ, Colditz GA, Willett WC, Rosner B, et al. A prospective study of high blood pressure and cardiovascular disease in women. Am J Epidemiol. 1989; 130:646–54. [PubMed: 2773913]
3. FitzGerald GA. Evolution in translational science: Whither the CTSAs? Sci Transl Med. 2015; 7:284fs15.

4. Callard F, Broadbent M, Denis M, Hotopf M, Soncul M, Wykes T, et al. Developing a new model for patient recruitment in mental health services: a cohort study using Electronic Health Records. BMJ open. 2014; 4:e005654.
5. Dziadkowiec O, Callahan T, Ozkaynak M, Reeder B, Welton J. Using a Data Quality Framework to Clean Data Extracted from the Electronic Health Record: A Case Study. EGEMS (Washington, DC). 2016; 4:1201.
6. Kopcke F, Lubgan D, Fietkau R, Scholler A, Nau C, Sturzl M, et al. Evaluating predictive modeling algorithms to assess patient eligibility for clinical trials from routine data. BMC Med Inform Decis Mak. 2013; 13:134. [PubMed: 24321610]
7. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc. 2014; 21:221–30. [PubMed: 24201027]
8. Buntin MB, Jain SH, Blumenthal D. Health information technology: laying the infrastructure for national health reform. Health Aff (Millwood). 2010; 29:1214–9. [PubMed: 20530358]
9. Byrd JB, Vigen R, Plomondon ME, Rumsfeld JS, Box TL, Fihn SD, et al. Data quality of an electronic health record tool to support VA cardiac catheterization laboratory quality improvement: the VA Clinical Assessment, Reporting, and Tracking System for Cath Labs (CART) program. Am Heart J. 2013; 165:434–40. [PubMed: 23453115]
10. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PR, Bernstam EV, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care. 2013; 51:S30–7. [PubMed: 23774517]
11. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc. 2013; 20:117–21. [PubMed: 22955496]
12. Madigan D, Ryan PB, Schuemie M, Stang PE, Overhage JM, Hartzema AG, et al. Evaluating the impact of database heterogeneity on observational study results. Am J Epidemiol. 2013; 178:645–51. [PubMed: 23648805]
13. Albers DJ, Hripcsak G. Using time-delayed mutual information to discover and interpret temporal correlation structure in complex populations. Chaos (Woodbury, NY). 2012; 22:013111.
14. Weber GM, Adams WG, Bernstam EV, Bickel JP, Fox KP, Marsolo K, et al. Biases introduced by filtering electronic health records for patients with “complete data”. J Am Med Inform Assoc. 2017:ocx071.
15. Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004; 15:615–25. [PubMed: 15308962]
16. Grimes DA, Schulz KF. Bias and causal associations in observational research. Lancet. 2002; 359:248–52. [PubMed: 11812579]
17. Albers DJ, Hripcsak G. A statistical dynamics approach to the study of human health data: resolving population scale diurnal variation in laboratory data. Physics letters A. 2010; 374:1159–64. [PubMed: 20544004]
18. Hripcsak G, Knirsch C, Zhou L, Wilcox A, Melton G. Bias Associated with Mining Electronic Health Records. Journal of Biomedical Discovery and Collaboration. 2011; 6:5.
19. Hripcsak G, Albers DJ. Correlating electronic health record concepts with healthcare process events. J Am Med Inform Assoc. 2013; 20:e311–e8. [PubMed: 23975625]
20. Hripcsak G, Albers DJ. High-fidelity phenotyping: richness and freedom from bias. J Am Med Inform Assoc. 2017
21. Levine ME, Albers DJ, Hripcsak G. Comparing lagged linear correlation, lagged regression, Granger causality, and vector autoregression for uncovering associations in EHR data. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2016; 2016:779–88. [PubMed: 28269874]
22. Swayne DF, Buja A. Missing Data in Interactive High-Dimensional Data Visualization. Computational Statistics. 1998; 13:15–26.
23. Meyer RD, Cook D. Visualization of data. Curr Opin Biotechnol. 2000; 11:89–96. [PubMed: 10679340]
24. US Department of Veterans Affairs VHA. Health Benefits: Veterans Eligibility.

25. Stone NJ, Robinson J, Lichtenstein AH, Merz CN, Blum CB, Eckel RH, et al. 2013 ACC/AHA Guideline on the Treatment of Blood Cholesterol to Reduce Atherosclerotic Cardiovascular Risk in Adults: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Circulation. 2013
26. Price LE, Shea K, Gephart S. The Veterans Affairs's Corporate Data Warehouse: Uses and Implications for Nursing Research and Practice. Nurs Adm Q. 2015; 39:311–8. [PubMed: 26340242]
27. Levey AS, Stevens LA, Schmid CH, Zhang YL, Castro AF 3rd, Feldman HI, et al. A new equation to estimate glomerular filtration rate. Ann Intern Med. 2009; 150:604–12. [PubMed: 19414839]
28. McGinnis KA, Brandt CA, Skanderson M, Justice AC, Shahrir S, Butt AA, et al. Validating smoking data from the Veteran's Affairs Health Factors dataset, an electronic data source. Nicotine & tobacco research : official journal of the Society for Research on Nicotine and Tobacco. 2011; 13:1233–9. [PubMed: 21911825]
29. Pivovarov R, Albers DJ, Sepulveda JL, Elhadad N. Identifying and mitigating biases in EHR laboratory tests. Journal of biomedical informatics. 2014; 51:24–34. [PubMed: 24727481]
30. 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. [PubMed: 24916006]
31. Newgard CD, Lewis RJ. Missing data: How to best account for what Is not known. JAMA. 2015; 314:940–1. [PubMed: 26325562]
32. Estiri H, Chan Y-F, Baldwin L-M, Jung H, Cole A, Stephens KA. Visualizing Anomalies in Electronic Health Record Data: The Variability Explorer Tool. AMIA Summits on Translational Science Proceedings. 2015; 2015:56–60.
33. Huang C-W, Lu R, Iqbal U, Lin S-H, Nguyen PA, Yang H-C, et al. A richly interactive exploratory data analysis and visualization tool using electronic medical records. BMC Med Inform Decis Mak. 2015; 15:92. [PubMed: 26563282]
34. Szklo, M., Nieto, FJ. Epidemiology: Beyond the Basics. 2. Boston: Jones and Bartlett Publishers; 2007. p. 392-4.
35. Weber GM, Mandl KD, Kohane IS. Finding the missing link for big biomedical data. JAMA. 2014; 311:2479–80. [PubMed: 24854141]
36. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. J Am Med Inform Assoc. 2013; 20:e147–54. [PubMed: 23531748]
37. Luo Y, Szolovits P, Dighe AS, Baron JM. Using Machine Learning to Predict Laboratory Test Results. Am J Clin Pathol. 2016; 145:778–88. [PubMed: 27329638]
38. Dahlem D, Maniloff D, Ratti C. Predictability Bounds of Electronic Health Records. Sci Rep. 2015; 5:11865. [PubMed: 26148751]
39. Hripcsak G, Albers DJ, Perotte A. Parameterizing time in electronic health record studies. J Am Med Inform Assoc. 2015; 22:794–804. [PubMed: 25725004]
40. Lasko TA, Denny JC, Levy MA. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. PLoS One. 2013; 8:e66341. [PubMed: 23826094]
41. Albers DJ, Elhadad N, Tabak E, Perotte A, Hripcsak G. Dynamical phenotyping: using temporal analysis of clinically collected physiologic data to stratify populations. PLoS One. 2014; 9:e96443. [PubMed: 24933368]
42. Hagar Y, Albers D, Pivovarov R, Chase H, Dukic V, Elhadad N. Survival analysis with electronic health record data: Experiments with chronic kidney disease. Statistical Analysis and Data Mining: The ASA Data Science Journal. 2014; 7:385–403.
43. Perotte A, Ranganath R, Hirsch JS, Blei D, Elhadad N. Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis. J Am Med Inform Assoc. 2015; 22:872–80. [PubMed: 25896647]

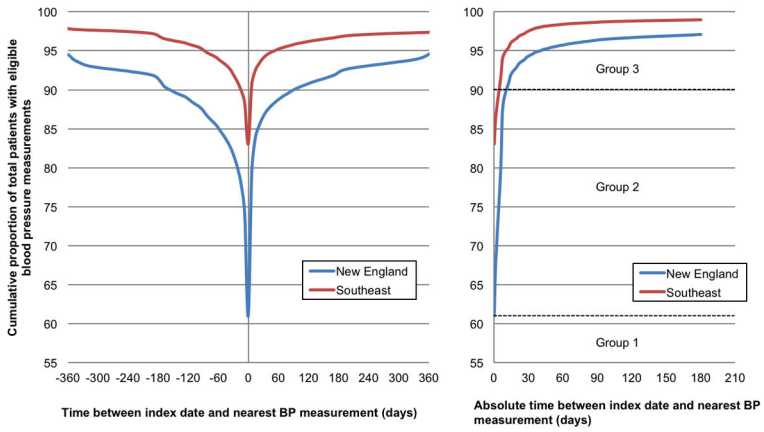

Figure 1. Yield of patients with eligible blood pressure (BP) measurement around an index date of lipid testing

Cumulative proportion of patients versus time (days) between index date and nearest BP measurement, analyzed before and after the index date separately (left panel) and by absolute time between the index date and BP measurement (right panel). Dotted lines (right panel) separate Groups 1, 2, and 3 from the New England network (see text).

Table 1

Patient characteristics by time between blood pressure and lipid measurements

	New England Network			Southeast Network		
	Group 1	**Group 2**	**Group 3**	**Group 1**	**Group 2**	**Group 3**
	BP at t=0	**BP between t=0 and t=90th percentile (+/−11d)**	**No BP at t≤90th percentile (+/−11d)**	**BP at t=0**	**BP between t=0 and t=90th percentile (+/−5d)**	**No BP at t≤90th percentile (+/−5d)**
n (%)	146,636 (61.0)	70,347 (29.3)	23,496 (9.8)	289,906 (83.1)	27,076 (7.8)	31,900 (9.1)
Age (SD), years	64.3 (14.2)	65.5 (14.1)	62.0 (17.4)	59.7 (14.3)	60.9 (14.7)	59.2 (15.8)
Gender (% male)	96.2	96.3	94.4	93.7	92.7	92.2
Race						
White (%)	76.3	71.8	61.1	45.3	44.0	38.9
Black or African American (%)	4.2	3.2	3.4	29.7	21.3	19.8
Other * (%)	0.9	0.8	0.8	0.9	1.0	0.9
Missing (%)	18.6	24.2	34.7	24.2	33.7	40.4
Ethnicity						
Non-Hispanic	81.3	76.3	66.8	79.8	73.6	66.2
Hispanic	1.3	1.3	1.4	0.7	0.8	0.8
Missing (%)	17.4	22.3	31.8	19.5	25.7	33.0
Medical insurance coverage						
VA only (%)	47.7	51.7	54.6	52.9	63.1	65.2
VA plus private (%)	16.5	15.0	15.8	16.1	11.5	12.4
VA plus other government (%)	35.8	33.2	29.6	31.0	25.4	22.4
Smoking status						
Current (%)	19.9	18.7	16.9	12.0	18.2	16.4
Former (%)	26.0	29.7	23.9	7.4	11.2	9.2
Never (%)	18.9	20.4	18.7	10.8	22.7	18.9
Unknown (%)	35.3	31.2	40.6	69.9	47.9	55.6
Baseline CVD (%)	33.2	30.3	25.7	25.3	29.0	22.9
Baseline diabetes (%)	20.6	18.9	14.6	20.1	21.9	15.8
Hypertension medications (%)	63.6	55.1	45.0	61.6	56.1	45.7
Cholesterol medications (%)	46.1	40.6	30.7	39.2	38.2	26.0

	New England Network			Southeast Network		
	Group 1	**Group 2**	**Group 3**	**Group 1**	**Group 2**	**Group 3**
	BP at t=0	**BP between t=0 and t=90th percentile (+/−11d)**	**No BP at t≤90th percentile (+/−11d)**	**BP at t=0**	**BP between t=0 and t=90th percentile (+/−5d)**	**No BP at t≤90th percentile (+/−5d)**
Systolic BP (SD), mmHg	133.3 (14.0)	132.8 (14.2)	131.8 (15.3)	134.3 (14.7)	133.6 (14.4)	132.5 (15.2)
Diastolic BP (SD), mmHg	75.5 (8.4)	75.1 (8.4)	74.7 (9.2)	76.5 (9.0)	76.1 (8.9)	75.7 (9.4)
Total cholesterol (SD), mg/dL	184.6 (38.0)	183.2 (36.4)	183.5 (38.9)	187.3 (38.6)	184.0 (38.6)	184.7 (39.0)
LDL cholesterol (SD), mg/dL	108.3 (31.7)	109.7 (30.7)	109.5 (33.6)	111.0 (32.8)	111.9 (38.5)	110.6 (33.2)
HDL cholesterol (SD), mg/dL	45.5 (12.4)	45.4 (12.1)	46.0 (12.8)	45.4 (14.9)	44.9 (14.5)	46.2 (15.4)
BMI (SD), kg/m^2	29.2 (5.5)	28.9 (5.2)	28.6 (5.2)	28.8 (5.7)	28.7 (5.5)	28.3 (5.4)
eGFR (SD), mL/min	74.5 (21.0)	73.4 (20.0)	73.8 (22.0)	79.1 (22.7)	77.0 (21.9)	77.9 (22.7)
Outpatient encounters, median (IQR)	13 (6, 28)	10 (5, 21)	7 (3, 16)	13 (7, 24)	12 (7, 23)	9 (5, 16)

Baseline characteristics of 3 patient groups in the New England and Southeast Networks of the Veterans Health Administration, categorized by availability of blood pressure (BP) measurements. BMI, body-mass index; CVD, cardiovascular disease; eGFR, estimated glomerular filtration rate; HDL, high-density lipoprotein: IQR, interquartile range; LDL, low-density lipoprotein; SBP, systolic blood pressure; SD, standard deviation.

* Includes American Indian, Alaska Native, Asian, Native Hawaiian, Pacific Islander.