



Published in final edited form as:

*Curr Epidemiol Rep.* 2017 December ; 4(4): 346–352. doi:10.1007/s40471-017-0130-z.

## Addressing Bias in Electronic Health Record-Based Surveillance of Cardiovascular Disease Risk: Finding the Signal Through the Noise

Julie K. Bower, PhD, MPH, FAHA<sup>1,2</sup>, Sejal Patel, MBBS, MPH<sup>1</sup>, Joyce E. Rudy, MS<sup>1</sup>, and Ashley S. Felix, PhD, MPH<sup>1</sup>

<sup>1</sup>Division of Epidemiology, College of Public Health, The Ohio State University, Columbus, OH

<sup>2</sup>Division of Cardiovascular Medicine, The Ohio State University College of Medicine, Columbus, OH

### Abstract

**PURPOSE OF REVIEW:** Use of the electronic health record (EHR) for CVD surveillance is increasingly common. However, these data can introduce systematic error that influences the internal and external validity of study findings. We reviewed recent literature on EHR-based studies of CVD risk to summarize the most common types of bias that arise. Subsequently, we recommend strategies informed by work from others as well as our own to reduce the impact of these biases in future research.

**RECENT FINDINGS:** Systematic error, or bias, is a concern in all observational research including EHR-based studies of CVD risk surveillance. Patients captured in an EHR system may not be representative of the general population, due to issues such as informed presence bias, perceptions about the healthcare system that influence entry, and access to health services. Further, the EHR may contain inaccurate information or be missing key data points of interest due to loss to follow-up or over-diagnosis bias. Several strategies, including implementation of unique patient identifiers, adoption of standardized rules for inclusion/exclusion criteria, statistical procedures for data harmonization and analysis, and incorporation of patient-reported data have been used to reduce the impact of these biases.

**SUMMARY:** EHR data provide an opportunity to monitor and characterize CVD risk in populations. However, understanding the biases that arise from EHR datasets is instrumental in planning epidemiological studies and interpreting study findings. Strategies to reduce the impact of bias in the context of EHR data can increase the quality and utility of these data.

### Keywords

electronic health record; bias; cardiovascular disease; risk factors; epidemiology

---

**Corresponding Author:** Julie K. Bower, PhD, 1841 Neil Avenue, 344 Cunz Hall, Columbus, OH 43210, Phone: 614-688-2148, Fax: 614-688-3533, jbower@cph.osu.edu.

**Potential Conflicts of Interest:** None declared.

## Introduction

Cardiovascular disease (CVD) is the leading cause of morbidity and mortality in the United States (U.S.).<sup>1</sup> CVD healthcare expenditures totaled \$316.1 billion in 2012–2013, accounting for 14% of total health care expenditures in those years.<sup>1</sup> Accordingly, epidemiological research has focused on reducing the burden of CVD through elucidating etiological factors, identifying early detection strategies, and informing prevention and disease management efforts. Much of this work has been facilitated with use of large prospective cohort studies, such as the Framingham Heart Study; however, substantial financial costs and logistical difficulties associated with developing and maintaining these resources have spurred investigators to seek more efficient ways to conducting this work. In particular, use of data from electronic health records (EHRs) has enabled epidemiologists to conduct cross-sectional and longitudinal investigations of CVD risk without the burdens imposed by assembling traditional cohort studies. Additionally, EHR data can be leveraged as an existing data source to conduct rapid and more efficient investigations into the population burden of CVD and its risk factors.<sup>2</sup> Although easier to obtain, EHR data present unique challenges to epidemiological research. These data are primarily collected for clinical purposes, and repurposing these data for surveillance research creates the potential for bias (defined as a systematic deviation of observed results or inferences from the “truth”)<sup>3</sup>.

In this paper, we review recent literature pertaining to the use of EHRs in epidemiological research aimed at surveillance of CVD risk in populations. We then explore and outline the most common types of bias that arise in EHR-based studies of CVD. Finally, we conclude with recommended strategies informed by the literature and our own work to reduce these biases in future research.

## Overview of sources of bias

Sources of bias that can threaten the validity of findings from CVD risk surveillance studies can broadly be divided into two categories: (1) selection bias, related to a discordancy between the representativeness of a study population and a particular target population to whom we hope to make inferences about; and (2) information bias, related to misclassification, mismeasurement, or nonrandom missingness of data used to characterize the population. Ideally, the perfect dataset would contain variables that directly capture the exact characteristic we aim to surveil, measured perfectly among all members of our target population. In reality, we nearly always end up with imperfectly measured variables among a biased subsample of the target population that are approximations of the true characteristic we hope to measure or are a surrogate for the characteristic we aim to capture (e.g., body mass index as a measure of obesity or C-reactive protein as a proxy for inflammation).

Bias affects findings from research and surveillance efforts in two key ways. First, if the population included in the study differs from the target population with regards to factors associated with cardiovascular health, we may incorrectly estimate the burden of a particular risk factor. For example, over-representation of older adults in an EHR system might result in an overestimate of the prevalence of hypertension (since blood pressure is positively

associated with aging). Second, if the data collected are of low or inconsistent quality, or have a high degree of missingness, we may misclassify individual patients' cardiovascular health. For example, we may overestimate the prevalence of hyperglycemia if fasting status is not confirmed before a glucose measurement is obtained in a clinical setting. Similarly, if a fasting glucose test is only ordered for patients with known diabetes risk factors, we may also inaccurately estimate the prevalence of undiagnosed diabetes if we ignore the nonrandom nature of the missing data (e.g., assume that all those with missing data are normoglycemic or simply exclude patients with missing data from calculations of hyperglycemia prevalence). While height, weight, and blood pressure are routinely measured at most clinical encounters, tests for other CVD risk factors such as hyperlipidemia or hyperglycemia may only be ordered when there is a clinical indication, potentially introducing bias. This limitation is unlike ascertainment of CVD risk factors in a traditional prospective cohort study where all participants uniformly undergo the same measurements, obtained in a standardized way, regardless of clinical indication.

In designing a surveillance study, we need to identify potential sources of bias so that we can: 1) prevent or reduce bias in the planning/design phase; 2) measure and quantify the remaining bias after data collection phase; and/or 3) adjust for this bias in the analysis phase. At a minimum, we should describe the magnitude and direction of bias after the data are collected so that we can project how our observed estimates may approximate or differ from the "truth." A thoughtful approach to designing surveillance studies can minimize bias by optimizing selection of patient populations, improving data collection approaches, addressing quality control issues, and capturing data to understand, quantify, and potentially adjust for bias.

## Selection Bias

Selection bias, a type of systematic error, is introduced into observational studies either by flawed recruitment/data extraction or by factors that affect subjects' participation in the study<sup>4</sup>. In a traditional prospective observational cohort study, this can occur in the design and implementation phases of research studies via inappropriately defined target populations and assembled sampling frames, lack of participation from eligible subjects, or both<sup>5</sup>. If the study population that is observed differs from the target population on key variables, then the study population is not considered to be "representative" of the target population and study validity may be compromised if the research questions of interest requires representativeness to produce generalizable findings<sup>6</sup>.

A major consequence of selection bias is its infringement on internal validity, where investigators make inferences unique to the sample that may not reflect the actual association in the intended target population<sup>4</sup>. Studies are considered internally valid when inferences are made in the context of minimal systematic error<sup>7</sup>. Internal validity precludes external validity<sup>4</sup>. The extent to which such inferences can be generalized beyond the sample in the study and pre-defined target population is encompassed by external validity<sup>7</sup>. Applications of external validity, particularly in the context of generating risk factor prevalence estimates, are only justified when the sample is representative of the population to which results are to be generalized.

Even though traditional health surveys and community-based studies are well-accepted approaches for conducting surveillance of CVD risk factors in a population, reductions in cost and resources as well as an interest in improving efficiency have motivated a shift towards utilizing EHRs and other existing administrative data sources over primary data collection<sup>7</sup>. However, selection bias from EHR-based surveillance of risk factors can inhibit the ability to accurately estimate CVD risk. A representative sample is fundamental to studies estimating disease burden in the general population<sup>8</sup>. Critics of the use of EHR data for characterizing population health assert that EHR datasets are composed of convenience samples, and that individuals accessing the health system (and thus populating EHR systems) are systematically different from those who abstain in a way that would bias findings obtained from such studies<sup>8</sup>.

Errors attributed to selection bias are minimized when each person in the target population has an opportunity to be selected for the study<sup>8</sup>. Obtaining truly random samples for EHR-based observational studies that measure CVD risk factors is difficult, though, because inclusion requires the patient to actively seek medical care to be included<sup>9</sup>. If information (e.g., demographic characteristics) is lacking on the population from which the patients arose, estimating the bias becomes challenging. Thus, identifying types of selection bias and the stages in which they occur in EHR-based surveillance studies is critical to understanding potential implications on effect estimates and generalizability of results.

Informed presence is the premise that patients do not appear randomly in an EHR data repository; rather, illness or symptoms may influence entrance into the healthcare system<sup>10</sup>. Thus, patients in the healthcare system are systematically different and more likely to be diagnosed with conditions that are also tracked in CVD risk factor surveillance than non-patients (i.e., healthcare system non-utilizers). When using EHR data for surveillance, we unintentionally condition on patients being ill for inclusion into the study. The exception to this are records that capture preventive care interactions, yet these too are subject to selection bias because factors such as education, health insurance coverage, and transportation might influence who uses these primary care services<sup>11,12</sup>.

The relationship between sufficiency of using EHRs for CVD risk factor surveillance and selection bias is well described<sup>10,13–15</sup>. CVD risk factor surveillance using EHRs hinges on records offering complete information<sup>16</sup>. Missing data in EHRs are considered missing at random (MAR) or not missing at random (NMAR), due to systematic biases from the clinical care process or to a key characteristic of the population. Missing data is also user-defined and meaningful only in reference to the data structure's ability to answer the research question of interest<sup>15</sup>. Criteria for meeting complete data requirements constrict inclusion into the study and exclude eligible subjects whose data may be relevant.

Patient wellness is correlated with breadth and frequency of information recorded in the EHR. Frequency of certain elements in the EHR such as laboratory results and medication orders are negatively associated with patient health<sup>16</sup>. Thus, those with more complete records are often patients with underlying health conditions that prompt more frequent visits with a healthcare provider<sup>16</sup>. We define "informed presence" as the notion that inclusion in an EHR is not random but rather indicates that the subject is more likely to be ill. It then

follows that persons represented in EHRs are systematically different from those not in EHRs. As other authors have noted, individuals contained within an EHR dataset tend to be non-representative of the larger population to whom results are meant to be generalized<sup>6,8</sup>. Since people within the EHRs are not observed randomly or in set intervals (but rather only when they have a medical encounter), there is the potential for bias in the collected data. One way this can manifest is that patients with more medical encounters have more opportunity to receive clinical diagnoses. By inflicting complete data requirements for inclusion into surveillance studies, the surveillance “system” may contain an overrepresentation of patients with poorer health. Risk factors for CVD may be seemingly more or less prevalent compared to the general population, and generalizing associations observed under such circumstances to healthier populations violates external validity. In fact, adults seeking medical care tend to have higher rates of diabetes, hypercholesterolemia, hypertension, and obesity, and a lower rate of smoking compared to adults in the general population<sup>17</sup>.

The populations in EHR data repositories are less heterogeneous than the target population; regardless, they are still being used for CVD risk factor surveillance. Inclusion of eligible patients into EHR-based surveillance is hampered by factors that influence healthcare utilization. For example, a study by Romo et. al using survey data from the 2013 New York City Community Health Survey (CHS) and the 2013–2014 New York City Health and Nutrition Examination Survey (NYC HANES) found that visiting a healthcare provider is more common among women, the unemployed, non-Hispanic Whites, and residents of neighborhoods with the lowest levels of poverty<sup>11</sup>. Negative perceptions about the healthcare system regarding cost, service availability, and culturally competent care also influence likelihood of visiting a provider<sup>11</sup>. Additionally, health insurance status is associated with entry into the healthcare system. Compared to those with health insurance, those who lack health insurance are more likely to be ill and less likely to receive medical care<sup>18</sup>. According to the U.S. Census Bureau, the highest uninsured rates are for young and middle-aged adults, those living below 100% of the poverty line, Hispanics, and noncitizens<sup>18</sup>.

Longitudinal surveillance studies — such as the Atherosclerosis Risk in Communities (ARIC) surveillance study that was designed to monitor trends in coronary heart disease and associated mortality — are often used in parallel with national health surveys to draw inferences about population cardiovascular health<sup>19</sup>. While migration bias is less problematic with diligent patient follow-up and tracking, additional design features of ARIC make migration bias even less likely. For example, participants were recruited for a cohort study from four clearly defined catchment areas and the subset chosen for the surveillance component were those residing in a geographic location with low migration and a single hospital servicing the medical needs of most participants. Sampling from specific geographic locations may limit generalizability, however, and participant follow-up at these sites may not reflect the typical healthcare setting. Further, healthcare systems use different EHR software and reporting is inconsistent among the programs. Tracking patients longitudinally also introduces challenges as the same measurements may not be conducted at regular intervals for all patients<sup>20</sup>.

A logical next step is to combine EHR data from multiple healthcare systems and repositories to increase the diversity of EHR datasets and address migration bias. However, simply pooling EHR datasets does not necessarily solve the issue, since high revenue healthcare systems with capabilities for big data analytics are the most common contributors to these collaborative efforts<sup>15</sup>. Consequently, patient diversity and representativeness in such data repositories may still be a concern.

## Information bias

Information bias occurs when data that appear in the EHR are inaccurate due to missing data, data entry errors, or measurement errors. The majority of bias arising from coding inaccuracies due to data entry errors is considered a form of non-differential misclassification, meaning that the misclassification is not systematically an over- or underestimate of the truth in the case of a continuous numeric variable (e.g., systolic blood pressure) and, therefore, is not considered a true bias. However, an important factor driving coding inaccuracies with diagnostic (International Classification of Diseases, or ICD) codes<sup>21</sup> is a preference for recording conditions that are likely to be reimbursed higher on the list for billing incentives; this type of bias is indeed considered “differential” because it results in a systematic over-reporting of procedures or conditions with more favorable reimbursement structures. Behavioral history information such as alcohol consumption or smoking may also be differentially misclassified if self-reported by the patient rather than directly observed; patients are more likely to underreport substance abuse and smoking behaviors. Of note, this bias due to inaccuracies in self-reported data, often termed “social desirability bias,” is similar to that seen in most observational research studies.

Differential misclassification often occurs when data from multiple EHR systems are aggregated for research purposes. Some EHR systems do not have compatible interfaces to simply merge data, leading to systematic missing values. Therefore, the investigator must address data harmonization issues. In the aggregated data setting, another challenge is to identify hospital-specific policies that might impact surveillance of a particular disease. Examples include enhanced screening for deep vein thrombosis, recording adverse effects of certain drugs, and monitoring specific types of complications of interest after procedures<sup>22</sup>. Additionally, use of certain ICD codes may vary between providers and across time<sup>23</sup>. Furthermore, medical equipment and laboratory tests may define different ranges for “normal” values, and using strict cutoffs for defining abnormal values in aggregated data may introduce bias in the estimates. Finally, data may be missing if non-standardized terminology is used, technical problems occur with data capture, or similar data fields are not uniform across EHR systems<sup>24</sup>.

Diagnostic suspicion bias, also known as over diagnosis bias, leads to higher estimated prevalence rates and occurs when symptomatic or high risk patients are more likely to undergo screening that subsequently leads to higher likelihood of diagnosis and receipt of treatment<sup>22</sup>. This may also be labeled as “surveillance bias” if there is an increase in disease burden measures due to policies around quality control conditions like adverse effects of drugs<sup>22</sup>.

Missing data in EHRs arise from multiple sources including values for measurements or laboratory tests that fall outside the detectable range, the frequency in which specific ICD codes/diagnoses are entered, errors in coding during data extraction, and missed deaths that occur outside the medical system. When a patient leaves a particular EHR system or is not seen regularly, it is difficult to know if the patient is healthy, receiving treatment from another provider, or is sick but not seeing a provider<sup>23</sup>. These missing values are NMAR and appropriate techniques are warranted to handle missing values<sup>25</sup>. As patients within the EHR system represent a dynamic cohort, calculation of at-risk person time is also a challenge<sup>23</sup>.

Data entry errors may occur when patients are treated for multiple conditions by different providers using different EHRs; this may be because a thorough medical history is not verified or comorbid conditions do not appear in discrete fields so may be missed in data extraction or pooling efforts<sup>23,26</sup>. Demographic information regarding factors such as race/ethnicity, geographic location, socioeconomic status, and education are either self-reported, assigned by the provider without input from the patient, or not queried at every encounter. In addition, data regarding domains of social or behavioral health such as psychological stress, physical activity, and social isolation are also commonly missing or incomplete.

## Recommendations for Assessing and Reducing Bias in EHR Studies

EHRs must meet five criteria to be considered valid data sources for CVD risk factor surveillance: 1) coverage of the EHR system(s) must include the entire population or a representative subset of the population, 2) cardiovascular health measures should be obtained in standardized way, 3) measures should be recorded in the EHR in a standardized way, 4) records need to be linked such that equivalent data are correctly merged, and 5) legal authority for data sharing needs to be in place<sup>27,28</sup>. Efforts to meet or address these requisites can also address concerns about bias. For example, selection bias would be eliminated if coverage of the EHR system(s) captures the entire target population or a simple random sample of that population. Universal healthcare coverage can increase entry into an EHR data repository, as evidenced by the observed increased testing for diabetes and hypercholesterolemia due to expansion of government insurance type<sup>13</sup>.

### Implementation of unique patient identifiers

Tracking patients over time and across systems is a challenge to conducting EHR research. Improvement of interoperability between record systems can be accomplished through legislation or agreements that require a unique identifier to be assigned to each record<sup>29,30</sup>, allowing for easier tracking of individual patients if they move between EHR systems over time. For example, the National Institutes of Health implemented the Global Rare Diseases Patient Registry Data Repository, in which de-identified records of a data repository are assigned a global unique identifier<sup>28</sup>. This process enables data from patients to be “integrated; tracked over time; and linked across projects, databases, and biobanks<sup>31</sup>.” The National Institutes of Health has also created an approach to provide unique identifiers (GUID-Global Unique Identifier) that can be used to link records across different systems<sup>30</sup>. Mandating collection of specific metrics for population-based studies can also enhance measurement standardization in EHRs<sup>31</sup>. A goal of the Query Health Project, for example, is

to validate a standard strategy for clinics to capture quality measures that can then be used for public health research<sup>28</sup>.

### **Adoption of standardized rules for inclusion/exclusion criteria**

In estimating the prevalence of CVD risk factors, decisions need to be made systematically to determine which patients within an EHR dataset are included in the denominator (i.e., the population total). For example, if quantifying the prevalence of hyperlipidemia in an EHR dataset, criteria need to be defined regarding whether patients with missing blood cholesterol information in the EHR are considered to be lost to follow-up (and therefore should be excluded from the prevalence calculation) or whether it can be assumed that because the test was not ordered that they are likely healthy (and therefore can be included in the population total and be considered to have values in the optimal range). If the latter, additional rules need to be applied for how patients with missing values will be categorized with regards to CVD risk. Adoption of clear guidelines on how missing data will be handled is important for consistency, transparency, and assurance so that important subpopulations of interest are not excluded from the analytic dataset.

### **Application of statistical procedures**

Statistical approaches can be used to describe and/or reduce the impact of bias after data collection has occurred. In the analysis phase, use of external data sources can help to evaluate and quantify bias in the study population. Several methods, described below, can be considered to integrate external data to reduce the impact of bias on study findings.

Describing how the study sample differs from the target population can be achieved by leveraging publicly-available data sources. For example, U.S. Census or state-level vital statistics data can be used to quantify differences in demographic and clinical characteristics of patients in the EHR dataset compared to the general population. Comparing distributions of these characteristics between the study sample to the target population can help inform generalizability of results<sup>15,32</sup>. Further, information that includes data on birth, death, pregnancies, and cancer can be validated through linkage with centralized databases and registries that when combined also serve to enhance EHR datasets. In some countries, individual-level data can be linked to population health and lifestyle surveys and data collected by other sectors regarding social factors<sup>33</sup>.

Post-stratification adjustment standardizes crude estimates according to variables implicating the selection bias. In the context of EHR data, these variables might include demographic factors such as sex, race/ethnicity, insurance status, and poverty level<sup>15</sup>. Sample weights can be generated to adjust for over- and under-representation of key population subgroups in the EHR dataset in comparison to the target population. Since inclusion into EHR is non-random, controlling for the number of health encounters also accounts for systematic differences between those who regularly or irregularly visit their provider<sup>17</sup>.

Additional frameworks have also been validated for dealing with selection bias specifically<sup>10</sup>. Propensity score adjustment/matching can be employed to account for systematic differences in health system “users” versus “nonusers” (i.e., to control for



selection bias). Propensity scores can then be used in analyses to create inverse probability weights to balance observed differences in the two populations with the goal of mimicking a scenario where individuals would be randomized to be included versus excluded from the EHR dataset.<sup>34,35</sup> Inverse probability weighting to achieve representativeness is still debated, but provides the ability to address factors that are associated with inclusion or exclusion in the dataset<sup>8,15,36</sup>.

Finally, efforts to reduce missing data at the onset should be explored when possible. For example, use of open source natural language processing tools can help to incorporate CVD risk factor data that may not appear in discrete fields in the EHR (e.g., family history or behavioral factors that may appear in clinic notes as free text)<sup>37</sup>. Additionally, imputation methods can be applied in scenarios where data are NMAR<sup>13</sup>. For example, in the case where a fasting glucose test is only ordered for patients with known diabetes risk factors, missing values may be imputed based on observed glucose data from these patients. In other words, it would be problematic to assume either 1) that all patients with missing glucose data have normal values, or 2) that all patients with missing glucose data were lost to follow-up.

### **Incorporation of patient-reported data**

Integration of patient-reported outcomes and other contextual information could also reduce the impact of missing data and should be considered for inclusion in EHR systems and surveillance efforts moving forward. In the case of tracking the prevalence of use of tobacco products over time, relying only on providers to accurately record and update this information affects data quality. The addition of standardized questionnaires to ascertain behavioral factors, for example, can improve EHR record completeness and accuracy.

### **Conclusion**

From a public health perspective, understanding the strengths and limitations of using the EHR for surveillance of CVD risk can inform more thoughtful design of epidemiological studies and interpretation of findings that utilize these data sources. Several strategies can also be incorporated in the data collection and analysis phases to reduce the impact of selection or information bias. Acknowledging and addressing its limitations, the EHR offers a powerful platform for monitoring and characterizing cardiovascular health on a large scale in an efficient and meaningful way, with the ultimate goal of advancing efforts to prevent, detect, and treat CVD to improve population health.

### **References**

\*Of importance

\*\*Of major importance

1. Benjamin EJ, Blaha MJ, Chiuve SE, et al. Heart Disease and Stroke Statistics-2017 Update: A Report From the American Heart Association. *Circulation* 2017;135(10):e146–e603. [PubMed: 28122885]

2. Kite BJ, Tangasi W, Kelley M, Bower JK, Foraker RE. Electronic Medical Records and Their Use in Health Promotion and Population Research of Cardiovascular Disease. *Current Cardiovascular Risk Reports* 2015;9(1):1–8.
3. Porta M A dictionary of epidemiology. Oxford University Press; 2014.
4. Tripepi G, Jager KJ, Dekker FW, Zoccali C. Selection Bias and Information Bias in Clinical Research. *Nephron Clinical Practice* 2010;115(2):c94–c99. [PubMed: 20407272]
5. Delgado-Rodríguez M, Llorca J. Bias. *Journal of epidemiology and community health* 2004;58(8):635–641. [PubMed: 15252064]
6. Bower JK, Bollinger CE, Foraker RE, Hood DB, Shoben AB, Lai AM. Active Use of Electronic Health Records (EHRs) and Personal Health Records (PHRs) for Epidemiologic Research: Sample Representativeness and Nonresponse Bias in a Study of Women During Pregnancy. *eGEMs* 2017;5(1).
7. Zaccai JH. How to assess epidemiological studies. *Postgraduate Medical Journal* 2004;80(941):140–147. [PubMed: 15016934]
8. Flood TL, Zhao Y-Q, Tomayko EJ, Tandias A, Carrel AL, Hanrahan LP. Electronic Health Records and Community Health Surveillance of Childhood Obesity. *Am J Prev Med* 2015;48(2):234–240. [PubMed: 25599907]
9. Kandola D, Banner D, O’Keefe-McCarthy S, Jassal D. Sampling Methods in Cardiovascular Nursing Research: An Overview. *Canadian Journal of Cardiovascular Nursing* 2014;24(3):15–18. [PubMed: 27356461]
- 10\*\*. 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.* 2016;184(11):847–855. [PubMed: 27852603] Coins the term informed presence, meaning that patient’s presence in EHR repository is not random; rather, it signifies that the patient is ill. Uses a simulation study to test the assumption that more visits to a clinician, means the the higher probability that disease will be detected. When number of health care visits is conditioned on, the selection bias that informed presence induces gets eliminated.
11. Romo ML, Chan PY, Lurie-Moroni E, Perlman SE, Newton-Dame R, Thorpe LE. Characterizing Adults Receiving Primary Medical Care in New York City: Implications for Using Electronic Health Records for Chronic Disease Surveillance. *Preventing Chronic Disease* 2016;13:E56. [PubMed: 27126554]
12. Christopher AS, McCormick D, Woolhandler S, Himmelstein DU, Bor DH, Wilper AP. Access to care and chronic disease outcomes among medicaid-insured persons versus the uninsured. *Am J Public Health* 2016;106(1):63–69. [PubMed: 26562119]
- 13\*\*. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. *Journal of Biomedical Informatics.* 2013;46(5):830–836. [PubMed: 23820016] Highlights the issue of lack of a formal definition for EHR completeness. Defines EHR completeness according to 4 elements: documentation, breadth, density, and predictive capability. Tested this definition against representative data from EHRs in New York-Presbyterian Hospital’s clinical data warehouse, and found that only half of the approximately 4 million records satisfied at least one component of completeness.
14. Haneuse S, Bogart A, Jazic I, et al. Learning about missing data mechanisms in electronic health records-based research: a survey-based approach. *Epidemiology (Cambridge, Mass).* 2016;27(1):82–90.
- 15 \*. Haneuse S, Daniels M. A General Framework for Considering Selection Bias in EHR-Based Studies: What Data Are Observed and Why? *eGEMs* 2016;4(1):1203. [PubMed: 27668265] Assesses the missing at random assumption in EHR data using data of patients with depression and weight change among these patients was investigated using a survey to delineate factors responsible for missing data for weight. They also described strategies for handling missing data in EHR research.
- 16 \*. 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 medical informatics and decision making* 2014;14(1):51. [PubMed: 24916006] Applies criteria of The American Society of Anesthesiologists Physical Status Classification System to 10,000 records from Columbia University Medical Center and determined that records with laboratory and

medication information belonged to the most ill patients. EHR data sufficient for research purposes is inadvertently contingent on sicker patients.

17. McVeigh KHN-DR, Perlman S, Chernov C, Thorpe L, Singer J, Greene C. Developing an electronic Health Record-Based Population Health Surveillance System. New York: New York City Department of Health and Mental Hygiene 2013.
18. Healthy People 2020: Access to Health Services. <https://www.healthypeople.gov/2020/topics-objectives/topic/Access-to-Health-Services>. Accessed September 28, 2017.
19. White AD, Folsom AR, Chambless LE, et al. Community surveillance of coronary heart disease in the Atherosclerosis Risk in Communities (ARIC) Study: methods and initial two years' experience. *Journal of Clinical Epidemiology* 1996;49(2):223–233. [PubMed: 8606324]
20. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. *Am J Epidemiol* 1989;129(4):687–702. [PubMed: 2646917]
21. Jha AK, DesRoches CM, Campbell EG, et al. Use of electronic health records in US hospitals. *New England Journal of Medicine* 2009;360(16):1628–1638. [PubMed: 19321858]
- 22 \*\*. Chioloro A, Santschi V, Paccaud F. Public health surveillance with electronic medical records: at risk of surveillance bias and overdiagnosis. *Eur J Public Health* 2013;23(3):350–351. [PubMed: 23599219] Describes surveillance and overdiagnosis bias with examples.
- 23 \*\*. Casey JA, Schwartz BS, Stewart WF, Adler NE. Using Electronic Health Records for Population Health Research: A Review of Methods and Applications. *Annu Rev Public Health* 2016;37:61–81. [PubMed: 26667605] Describes the content of EHR data, how study variables are constructed, study population is set up, manifold uses of EHR data including epidemiological, environmental and social epidemiology study and compares the traditional epidemiological study with EHR studies.
- 24 \*. Cowie MR, Blomster JI, Curtis LH, et al. Electronic health records to facilitate clinical research. *Clin Res Cardiol* 2017;106(1):1–9. Describes the use of EHR data in research including epidemiological research, challenges associated with use of EHR in clinical studies and potential solutions.
25. Haneuse S, Bogart A, Jazic I, et al. Learning About Missing Data Mechanisms in Electronic Health Records-based Research: A Survey-based Approach. *Epidemiology (Cambridge, Mass)*. 2016;27(1):82–90.
26. 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):S30–37. [PubMed: 23774517]
- 27 \*\*. Nair S, Hsu D, Celi LA. Challenges and Opportunities in Secondary Analyses of Electronic Health Record Data Secondary Analysis of Electronic Health Records. Cham: Springer International Publishing; 2016:17–26. Reviews biases that may occur when using pooled data from multiple healthcare systems for public health research. Uncovers barriers to development of a clinical data warehouse from multiple healthcare systems, including monetary and collaboration constraints. Considers that current pooled EHR data sets are mainly comprised of records from high revenue healthcare systems, and that the promising opportunity of EHR data for secondary analyses may have drawbacks such as data security issues.
- 28 \*\*. Friedman DJ, Parrish RG, Ross DA. Electronic Health Records and US Public Health: Current Realities and Future Promise. *Am J Public Health* 2013;103(9):1560–1567. [PubMed: 23865646] Recognizes the potential contributions of EHR to population health research and provides recommendations to make EHR data less biased for secondary analyses. These included increased population coverage of EHRs, standardized collection and recorded on measures, and legislature that would enable more effective pooling of EHR from multiple data sources.
29. Sommers BD, Gawande AA, Baicker K. Health Insurance Coverage and Health — What the Recent Evidence Tells Us. *New England Journal of Medicine* 2017;377(6):586–593. [PubMed: 28636831]
30. NIH. GUID (Global Unique Identifier). 2017; <https://data-archive.nimh.nih.gov/guid>.
31. Resouces. TNNPG. Global Unique Identifier (GUID). 2017; <https://ncats.nih.gov/grdr/guid>. Accessed September 29, 2017.

32. Casey JA, Schwartz BS, Stewart WF, Adler NE. Using Electronic Health Records for Population Health Research: A Review of Methods and Applications. *Annual Review of Public Health* 2016;37(1):61–81.
33. Wood R, Clark D, King A, Mackay D, Pell J. Novel cross-sectoral linkage of routine health and education data at an all-Scotland level: a feasibility study. *The Lancet* 2013;382:S10.
34. Deb S, Austin PC, Tu JV, et al. A Review of Propensity-Score Methods and Their Use in Cardiovascular Research. *Can J Cardiol* 2016;32(2):259–265. [PubMed: 26315351]
35. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behav Res* 2011;46(3):399–424. [PubMed: 21818162]
36. Thompson CA, Arah OA. Selection bias modeling using observed data augmented with imputed record-level probabilities. *Annals of epidemiology* 2014;24(10):747–753. [PubMed: 25175700]
37. Khalifa A, Meystre S. Adapting existing natural language processing resources for cardiovascular risk factors identification in clinical notes. *J Biomed Inform* 2015;58 Suppl:S128–132.