International Journal of Engineering Research & Management Technology
ISSN: 2348-4039
March-April-2023 Volume 10, Issue-2
Email: editor@ijermt.org www.ijermt.org

THE USE OF ELECTRONIC HEALTH RECORD DATA AND METHODOLOGICAL ADVANCES IN AGING EPIDEMIOLOGY

Deebya Mukherjee

Assistant Professor, ABS Academy of Management and Health Science (Affiliated to Maulana Abul Kalam Azad University Of Technology, West Bengal)

Sagarbhanga, Durgapur, West Bengal

ABSTRACT

Over the past ten years, there has been a significant rise in the functionality and use of electronic health records (EHRs). While EHRs are primarily utilized for clinical purposes, researchers have also employed them for epidemiologic investigations. These investigations have taken the form of longitudinal studies on patients who are geographically dispersed or cross-sectional studies conducted within a single institution. In this article, we define EHRs, look at how they're used in community health studies, and contrast them with conventional epidemiologic techniques. We outline a variety of research uses that profit from the broad patient populations and big sample sizes that EHRs make possible. These have included the reassessment of earlier results, a variety of illnesses and their subgroups, epidemiology of the environment and society, conditions that are stigmatized, prediction modeling, and assessment of natural trials. EHR-based studies are less costly and take less time to complete, even if primary data gathering methods may yield more dependable data and higher population retention. Future EHR epidemiology could increase clinical care and population health by integrating future technologies like personal sensing, linking with vital data, and collecting more social and behavioral measures.

Keywords: *electronic health records, EHR, environmental epidemiology, social epidemiology, geographic information systems, health determinants*

INTRODUCTION

The prevalent ideas, the risk factor measures that are now accessible, and the expense of acquiring pertinent data all influence the design and conclusion of epidemiologic research. Vital statistics were widely employed by researchers to perform cross-sectional and time series investigations of noninfectious disease prior to the 1950s. Longitudinal data were lacking, which hindered causal inference. Thanks to financing in the second part of the 20th century, researchers were able to create cohorts of people who were tracked over time. However, in the twenty-first century, conducting traditional expensive and time-consuming prospective studies is made more difficult by diminishing research support and participation rates.

A timely substitute is provided by the recent increase in the usage of electronic health records, or EHRs. For epidemiologic research, these databases offer an inexpensive way to obtain comprehensive longitudinal data on huge populations. EHRs are more than just digital copies of paper records; they may answer complicated

network of causality questions by combining self-reported data with contextual data sourced from geographic information systems (GIS). The work in question has promise for the advancement of epidemiologic theory in the 21st century.

We outline the characteristics of EHRs and their use in epidemiological research in this paper. Ever since its debut, EHR data have significantly enriched a wide range of public health scholarship, spanning from social epidemiology to infectious disease research. In order to guide future research, we first outline this body of literature and then compare and contrast traditional and EHR-based studies to emphasize the relative advantages and disadvantages of each.

OBJECTIVES

1. To determine the use of electronic health record data in aging epidemiology.
2. To construct epidemiologic cohorts from electronic health record data

METHODOLOGY

Selecting Research Participants

Research participants in a prospective cohort study are usually recruited volunteers who satisfy predetermined eligibility requirements, are willing to participate in a variety of study procedures, and consent to active follow-up at predetermined intervals, sometimes up to several years following enrollment. The only people who can participate in an EHR-based retrospective study are those who are patients, or people who are seeking services from healthcare institutions. Significantly, some estimates indicate that in any one year, up to 50% of the population may not come into contact with the healthcare sector. This fact might have an impact on extrapolating results from epidemiologic research based on electronic health records to a larger target group. After that, apart from inclusion criteria pertaining to illnesses unique to the study, the main challenge is identifying the right research subset from a potentially large EHR patient database.

People engage with healthcare organizations in a variety of ways, and these variations always affect the quality and accessibility of data obtained from an electronic health record. Additionally, patients frequently receive care from several healthcare organizations, and it is frequently impossible to integrate data between them, which further impairs the quality of the information. Two overlapping but separate characteristics of information quality in EHR-based research are accuracy and completeness. Although data (in)correctness—such as incorrectly recorded diagnostic codes or quantitative errors—is a real risk in EHR-based research, it is frequently difficult for researchers to identify and fix in a retrospective setting. On the other hand, data completeness is more within the control of the researcher because criteria can be used to improve completeness in the final research subjects selected. Sadly, when evaluating retrospective EHR data where actual "completeness" is ill-defined, data completeness—that is, the ability of EHR data to properly characterize an individual's medical state—becomes an ill-defined construct. On the other hand, it is evident that completeness positively correlates with both the types and frequency of interactions with a healthcare institution (more interaction equates to better completeness; for example, primary care visits usually yield more information than specialized visits). There is therefore a justification for using a "information completeness" metric of some kind when choosing research participants from an EHR database—and only

including patients who surpass a certain threshold—but these criteria are generally untestable and can be challenging to define in an appropriately objective manner. Moreover, selecting research participants based on the incompleteness of the data may lead to an overselection of less well-off patients, as they are more likely to require medical attention. Accounting for such uneven data completeness is, in fact, a major continuing analytical difficulty in EHR-based epidemiology.

When a study's objectives justify it, a reasonable place to start when establishing inclusion criteria is primary care services received through the healthcare organization, since this will help achieve the data-completeness objective. A larger percentage of a patient's medical record can usually be obtained during primary care visits, and primary care frequently acts as a gateway to more specialized services, which are then frequently provided by the same healthcare organization when necessary (assuming the facility provides such services). Study inclusion criteria should be carefully considered because only a small percentage of patients in an EHR database may have acceptable information quality. Tradeoffs between sample size and information quality must be made because adding more patients to a study only benefits the study at the expense of the additional patients' lower information quality (Figure 1). In the T2DM-HFH example, roughly 500,000 patients who had received primary care services from the study institution for at least two years following the initial EHR-documented interaction made up the base group, from which T2DM patients were selected. Merely one-third of the total patients in the EHR database were included in this research subset.

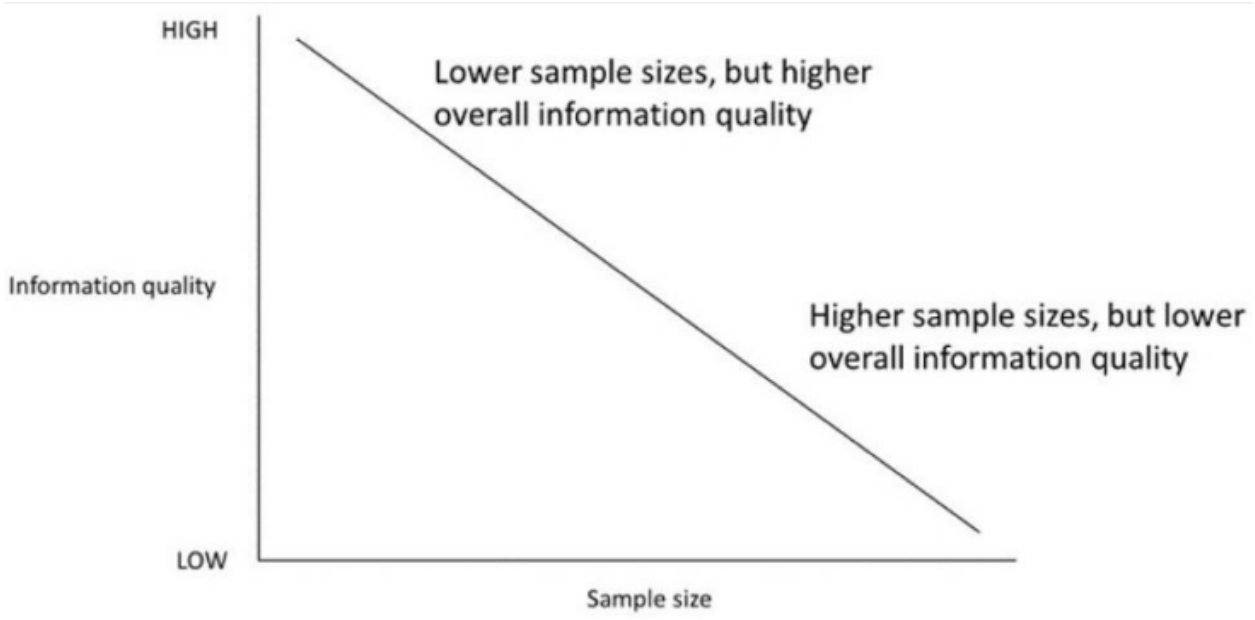

Figure 1 Electronic health record data quality spectrum. Depending on a variety of circumstances, including the frequency and nature of interactions with healthcare institutions, the information quality of patients' electronic health records can vary greatly. Increasing the number of patients included in epidemiologic research is possible by moving down the information quality spectrum, but this comes at the cost of data quality.

Defining "Baseline" and Assembly of Baseline Characteristics

A precise study start date is predetermined in prospective cohort studies. This date may be the diagnosis of a particular condition, the completion of a particular procedure, or the acquisition of informed permission in the case of a study including volunteers in general good health. This date is the beginning point for the follow-up of study results and is commonly referred to as the "baseline date" (or just "baseline"). A typical prospective study uses a variety of methods, including blood samples, questionnaires, and possibly even more innovative assessment technologies like imaging, to measure a wide range of baseline parameters. Notably, prospective studies by design assess baseline variables for each research participant in a consistent, protocol-specified manner.

Each patient's electronically recorded journey through a healthcare organization can be represented as a timeline in an EHR-based retrospective cohort study, with the first and last EHR-documented encounters serving as bookends. The encounters in between can be multiple, unevenly spaced, and qualitatively distinct in nature (Figure 2). Any professional interaction between a patient and a healthcare organization, including visits to the emergency room, speciality care, laboratory tests, primary care, hospital admissions, and others, is widely referred to here as an encounter. When patient-provider communications like emails and phone calls yield pertinent data for a research project, they might be classified as interactions even though they are less direct. Remarkably, a healthcare institution might provide a restricted range of encounters (e.g., a hospital that operates independently), which could provide limitations for study. Any EHR-based retrospective study can choose a baseline date that falls on or between the first and last encounters, but there are some guidelines that need to be followed. Firstly, baseline dates should be chosen as close as possible to the first EHR-documented encounter (when this makes sense) in order to optimize post-baseline follow-up time. The first encounter, however, is typically an unattractive baseline date since it rarely offers enough information for a meaningful baseline characteristic assessment. A more thorough baseline evaluation can be achieved by giving more time and encounters to accumulate before the baseline date assignment, but this comes at the cost of a smaller sample size and shorter follow-up periods. Indeed, it is normal for retrospective studies utilizing electronic data sources to require pre-baseline encounter data spanning at least six months, but usually two years or more. Due to the nature of the data-generating process, healthier individuals often need fewer interactions to offer a complete medical picture, but worse patients often need more.

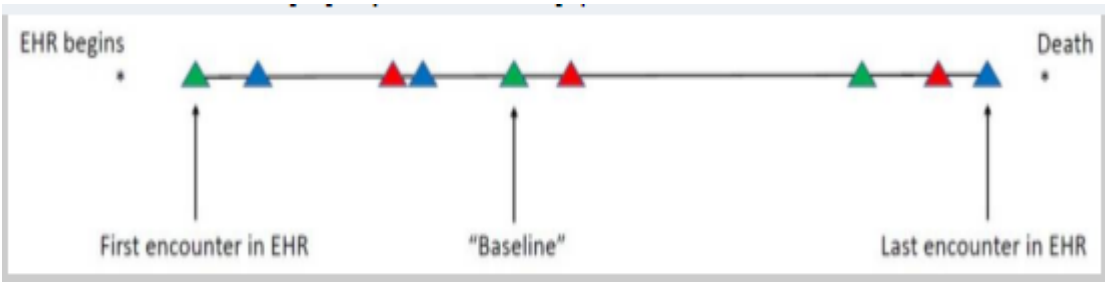

Figure 2 A patient's electronic health record (EHR)-documented journey through a healthcare organization is shown as a timeline, with the first and last EHR-documented encounters at the beginning and conclusion, and several further encounters of varying spacing and quality between. Each form of encounter is represented by a colored triangle.

EHR Data: What Is Available

Any EHR-based epidemiologic study can take into account data elements that are recorded during routine healthcare service delivery. These can be broadly categorized into categories such as demographics, vital signs, diagnoses, procedures, medications, and laboratory tests. The default measuring procedures put in place within the clinical business and the recording strategies used by certain healthcare professionals unquestionably influence the quantitative (how much) and qualitative (how good) characteristics of EHR data. An estimated 20% of doctors' professional time is devoted to documenting clinical encounters. While there are incentives to document as much as possible in order to maximize reimbursement, but not too much in order to prevent fraud, it is impossible to assess how well these guidelines were followed in actual practice in retrospect. Every epidemiologic study has issues about missing data, misclassification of categorical variables, and measurement error of continuous factors. These concerns are exacerbated in EHR-based research due to the nature of the data-generating process. Sadly, from the standpoint of a researcher, these data restrictions cannot be easily fixed at the source, but steps can be done to lessen their effects.

Opportunity for Information

The baseline date acts as the reference point by which study variables are allocated a present/absent status for dichotomous variables or a numerical value for continuous variables when calculating baseline features in an EHR-based retrospective research. Baseline data is compiled from interactions that happen on or before the baseline date and, when applicable, from interactions that happen soon after the baseline (e.g., 90 days). An EHR-based retrospective study really cannot standardize the process of organizing baseline characteristics in a way that is entirely acceptable, in stark contrast to the prospective study environment. In fact, it is more likely that no two patients will have had their baseline characteristics obtained in the same way when all the possible combinations of the quantitative (number of encounters) and qualitative (e.g., primary care, ED visits) ways patients could interact with healthcare organizations are taken into account. Opportunity for Information (OFI), a new concept, is presented here to characterize the gathering of interactions that may yield useful baseline data (Table 1). Reiterating the initial issue about the OFI, any EHR-based research study is likely to find significant inter-patient variability in the OFI. The number of encounters from the first EHR-documented encounter to the baseline encounter can be used to quantify OFI in terms of time. It may also be thought of implementing an OFI metric depending on specific encounter kinds (such as the quantity of primary care visits). In the T2DM-HFH example, the baseline date was determined by the initial office visit that followed a primary care visit and two years had passed since the first EHR-documented interaction. The OFI time ranged from 2.0 to 14.8 years, with a mean (SD) of 4.6 (3.3) years using this baseline definition. Pre-existing (at baseline) and newly diagnosed T2DM were included in the study, which accounts for part of the broad variation in OFI duration; as predicted, pre-existing T2DM had a significantly lower mean OFI time than new diagnoses (3.5 vs. 6.7 years). Once more, a broad dispersion in OFI is seen when characterizing OFI variability in terms of the encounter frequency (Figure 3). The range of encounters available to determine baseline characteristics was 2 to 1437, with a median (IQR) of 22 (11, 41). The most informative encounter type in Figure 3 is the number of office visits, which is shown separately.

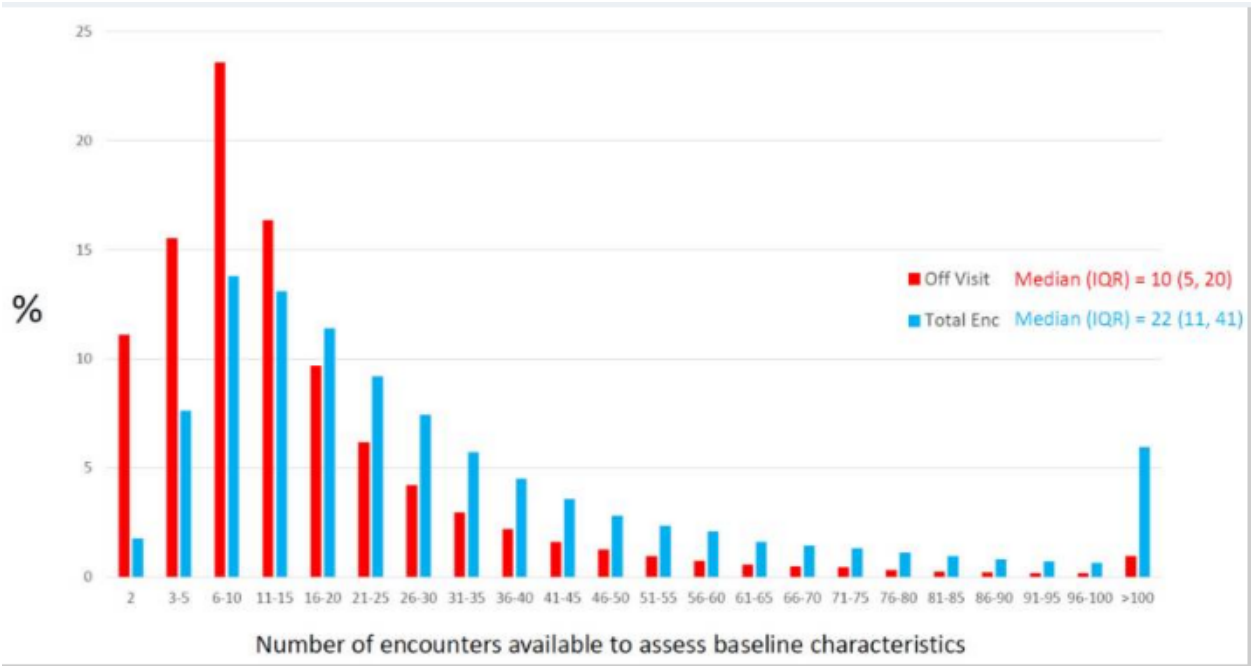

Figure 3 A case study including a hospitalization for heart failure and diabetes provide an opportunity for information. Baseline characteristics for all interactions (blue) and office visits only (red) are determined by relative frequency histograms of the number of encounters.

Table 1 Definitions of terms and phrases.

Term or Phrase	Definition
Encounter	Any professional contact between a patient and healthcare organization, including primary care, specialty care, laboratory testing, emergency department visits, hospital admissions, etc.
Opportunity for Information	The collection of pre-baseline encounters that could provide usable research information. Can be expressed in units of time (days from first encounter to baseline encounter) or as number of encounters (between first and baseline encounters).
Creating Rules for the 99%	When assembling baseline characteristics for an EHR-based retrospective study, rules must be created for determining presence/absence of qualitative characteristics and values for quantitative characteristics. This informal expression implies that imperfect rules must be implemented that work well for the majority but rarely universally.
Looking for Yes	An expression applied when determining the presence/absence of a binary characteristic, denoting how rules typically only look for positive affirmations of the characteristic and rarely negative affirmations.
Hidden Missingness	A phrase describing the scenario where a qualitative condition (e.g., diagnosis) is labeled "absent" but was never queried nor investigated in clinical practice. Thus, the condition's true status as present/absent is actually undetermined despite being labeled "absent".
Weak No	A scenario where a qualitative condition (e.g., a diagnosis) is labeled absent based on weak information.
Strong No	A scenario where a qualitative condition (e.g., a diagnosis) is labeled absent based on strong information.

The aforementioned instance illustrates how the quantity of pertinent data required to compile baseline characteristics might differ significantly amongst research participants, even in a group of patients who meet a minimal set of requirements. The main issue with inter-patient variability in OFI is that it appears that there is a correlation between the rate of misclassification for specific binary features and the reported presence of certain baseline parameters on the OFI. Specifically, there is a positive correlation between the OFI and EHR documentation of traits that are intermittent, transient, and/or more subjectively decided (e.g., depression, shortness of breath); whereas, permanent, common, and more objective aspects (e.g., hypertension) have a weaker correlation. Compared to higher OFI, lesser OFI always results in a higher false-negative classification of baseline features. Without a doubt, the OFI will have a role in the accurate identification of pre-existing medical issues at baseline, as research participants with higher OFI will naturally have more opportunities to document a medical condition. When there is a significant relationship between two OFI-dependent variables, there is a particular risk since it is possible to create an artificially inflated correlation. When the right conditions are present, the phenomena behaves quantitatively like standard confounding and can be managed using standard confounding-correction strategies (by using an OFI metric). Nevertheless, putting these strategies into reality can be difficult, especially when there are multiple covariates—some of which are reliant on OFI while others are not.

Several studies using electronic data sources (EHRs, insurance claims) have applied a fixed, pre-baseline time interval for baseline assessment, which ignores encounter information prior to the interval, in an attempt

to control for inter-patient variability in OFI and standardize the baseline information-gathering process. While the strategy will standardize OFI duration (SD = 0), encounter-based OFI measures will still exhibit inter-patient variability in the end. Applying a fixed, 2-year pre-baseline time constraint in the T2DM-HFH example results in a reduction in the number of encounters evaluated for baseline evaluation, as shown in Figure 4 (mean OFI duration drops from 4.6 to 2.0 years). About one-third of patients use only two visits to assess baseline characteristics as a result of the restriction, which lowers the median number of total encounters considered from 22 to 4 (Figure 4). The rate of false-negative misclassification that is increased by this standardization attempt also leads to a new issue. In fact, the method purposefully transforms what is assumed to be accurate information into false information. When a 2-year time constraint is applied, Table 2 illustrates the decline in the prevalence of a subset of baseline features from the T2DM-HFH case. Applying a pre-baseline time restriction may not have much of an effect on how aggregate numbers are interpreted, as the table indicates; but, comparing counts with and without the restriction shows how much misclassification was caused. While there are always going to be limitations, it seems best to choose a strategy that includes all pre-baseline interactions that are available.

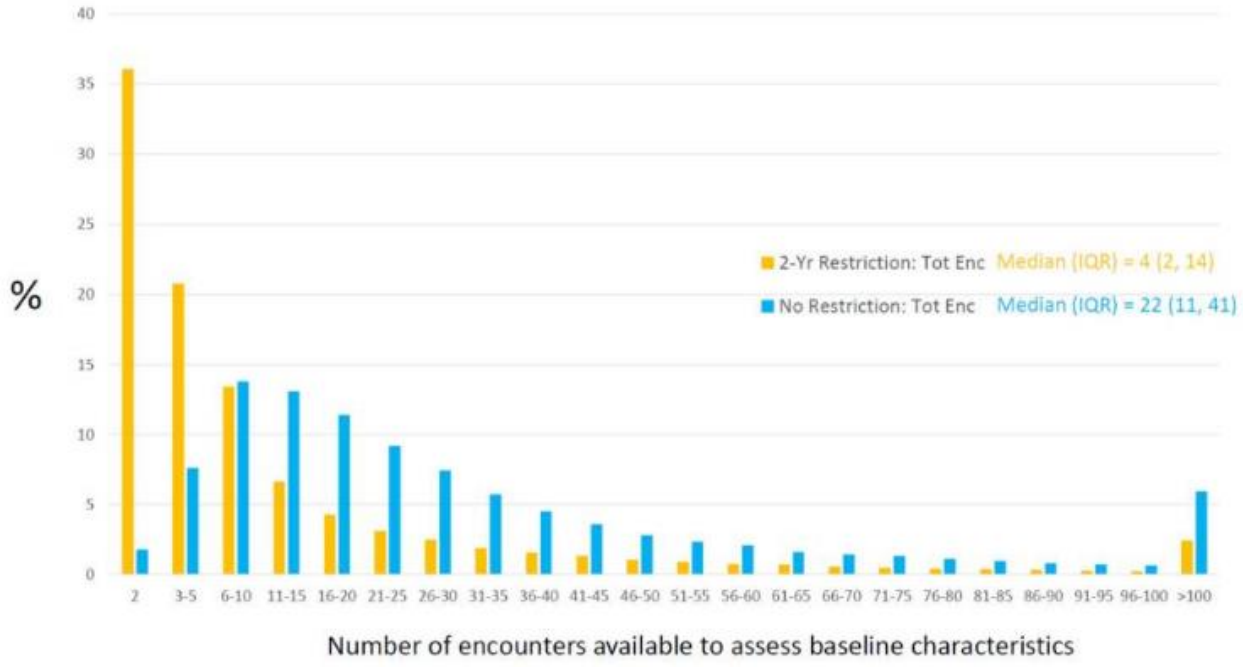

Figure 4 Informational Opportunity: the impact of imposing a two-year pre-baseline time restriction on hospitalizations for heart failure and diabetes. The blue bars represent the relative frequency histograms of all encounters used to establish baseline characteristics, whereas the orange bars show the same data restricted to encounters that occurred within two years of baseline. The blue bars do not restrict the encounters' timing from before baseline.

Table 2 Loss of information when restricting pre-baseline time intervals for assessment of baseline characteristics.

Baseline Characteristic	No Restriction	2-Year Restriction
Hypertension	71% (*n* = 56,653)	67% (*n* = 53,350)
High cholesterol	69% (*n* = 54,652)	64% (*n* = 51,003)
Coronary bypass surgery	7% (*n* = 5293)	6% (*n* = 4673)
Heart failure	11% (*n* = 9026)	10% (*n* = 8170)
Acute myocardial infarction	8% (*n* = 6516)	7% (*n* = 5362)
Chest pain	22% (*n* = 17,179)	15% (*n* = 12,141)
Shortness of breath	16% (*n* = 12,993)	12% (*n* = 9784)
Depression	25% (*n* = 19,812)	21% (*n* = 16,901)

Creating Rules for the 99%

Putting together baseline characteristics from the wide range of elements found in EHRs means developing rules for the 99%, a colloquial way of saying that we need to apply imperfect rules that are effective for most people but not always for everyone. The argument against is that there are frequently situations in which a proposed rule applied strictly results in wrong information; yet, modifying the rule to account for the circumstance incorrectly alters correct information for a large number of additional study subjects and may therefore be detrimental. For example, in T2DM, demanding documentation of a high hemoglobin A1c as part of the diagnostic criteria of a study may lead to the incorrect classification of patients as diabetes based solely on diagnosis codes (but without the A1c). When creating a rule for qualitative variables, one typically observes the relevant structured data elements at encounters within predetermined time frames, possibly with additional criteria based on temporal proximity (e.g., requiring <1 year between separate code occasions), frequency (e.g., requiring >1 occasion of a code), and/or context (e.g., primary diagnoses given precedence over secondary). In the end, this noise must be tolerated by researchers, and faulty standards must be put into place. Since EHR data sets are usually too big for thorough manual validation of every data element, rules must be relied upon. Thankfully, researchers may examine EHR-based rules closely by directly observing electronic medical documents. This enables them to "pull back the curtain" to identify and improve subpar rules, an opportunity not afforded by the majority of insurance-claims-based studies.

Hidden Missingness

EHR-based rules for binary characteristics are primarily limited to positive affirmations for defining disease "presence" (i.e., observing a documented code) and the absence of positive affirmations for defining disease "absence" (i.e., not observing a code) due to the nature of clinical documentation processes (see phrase "looking for yes"—Table 1). In other words, negative affirmations—documentation that a certain disease was sought but not found—are uncommon in organized EHR data, which would support the absence of the condition in real life. Because of this, it is intrinsically impossible for EHR-based research to distinguish between "no disease" and "missing disease status"—the former being defined as a clinical scenario in which a particular disease was sought but not discovered, and the latter as a disease that was not sought in any clinical setting. Due to the fact that some diagnoses that are recorded in EHR studies rely on the OFI, there will eventually be some concealed missingness in the form of qualitative diagnoses that are labeled as "not present" based on rule criteria but were not really examined in routine clinical settings. The degree of

concealed missingness leads to a misclassification issue and is correlated with the quality of the information (and the OFI). poor no and strong no are two similar ideas introduced here; the former describes disease-absence labels based on poor knowledge, while the latter is based on strong information.

Quantitative Data: Measurement Error and Missing Data

Comparing EHR data to other electronic data sources like insurance claims, one frequently mentioned strength of EHR data is the quantitative data that is typically available within them (e.g., vital signs, laboratory test results). When tying a single numerical value to a baseline date for a quantitative data element, the process should be hierarchical in nature, favoring values measured on the baseline date first, values measured closest in time prior to baseline (possibly with a limit on how far back in time is allowed), and values measured closest in time after baseline (possibly with a definite limit on how far forward in time is allowed). It may also be important to take into account the context of the assessment (e.g., inpatient vs. outpatient). There will inevitably be a lot of missing data and measurement errors. Quantitative measurements from an EHR are mostly uncorrectable and subject to random fluctuation that is frequently larger than equivalent measurements made under a standardized, prospective research technique (e.g., blood pressure). Moreover, missing data in an EHR is rarely the result of data loss. For example, in one study, different quantitative health indicators were measured according to treatment status, degree of chronic illness, and demographic variables. Missing data frequently suggest improved health in unreported ways. The use of common imputation techniques is complicated by the shakiness of the missing-at-random assumption.

CONCLUSIONS

EHR-based retrospective studies that draw valid conclusions and estimate minimally biased effects rely heavily on identifying the most useful patients from an EHR database while maintaining generalizability. Because people use healthcare services in different ways and to varying degrees, it is likely that a significant portion of patients in an EHR system will have incomplete or incomplete data, which means they do not fulfill the minimal requirement for data completeness and should not be included in research studies. Although it makes sense, selecting patients with the goal of maximizing data completeness defies a strictly objective definition and is likely to overselect a population of less healthy patients. Enforcing primary care services through the EHR-bearing organization as part of the study's inclusion criteria should increase confidence in a more thorough determination of baseline characteristics and future study outcomes, but the organization's overall service spectrum and the degree to which clients utilize it are also crucial from the standpoint of data completeness. Data completeness can also be impacted by other factors, such as the standing of a healthcare organization and the quantity of rival healthcare organizations in the region. Mature EHR databases offer vast patient populations and unprecedented longitudinal detail, but they also often need some compromise in data quality to fully utilize them. Although hidden missingness cannot be prevented and cannot be measured, care can be taken to reduce its effects by carefully choosing patients and developing rules. Regretfully, it is simple to conduct an EHR study that yields credible results but is riddled with missing data and avoidable misclassification. Because EHR study results are exceedingly exact and come from large sample numbers, they may appear credible, but they may also be seriously biased. Notwithstanding these drawbacks, as EHR databases multiply, EHR-based retrospective studies will probably gain popularity. Given the importance of EHRs in the current healthcare setting, epidemiologists need to keep refining the techniques of EHR-based epidemiology.

References

1. Adler NE, Prather AA. 2015. Risk for Type 2 diabetes mellitus: person, place, and precision prevention. *JAMA Intern. Med*. 175:1321–22

2. Adler NE, Stead WW. 2015. Patients in context—EHR capture of social and behavioral determinants of health. *N. Engl. J. Med*. 372:698–701

3. Adler-Milstein J, DesRoches CM, Furukawa MF, Worzala C, Charles D, et al. 2014. More than half of US hospitals have at least a basic EHR, but stage 2 criteria remain challenging for most. *Health Aff*. 33:1664–71

4. Alonso A, Jick SS, Hernán MA. 2006. Allergy, histamine 1 receptor blockers, and the risk of multiple sclerosis. *Neurology* 66:572–75

5. Alwahaibi A, Zeka A. 2015. Respiratory and allergic health effects in a young population in proximity of a major industrial park in Oman. *J. Epidemiol. Community Health*. doi: 10.1136/jech-2015-205609.

6. Anderson HD, Pace WD, Brandt E, Nielsen RD, Allen RR, et al. 2015. Monitoring suicidal patients in primary care using electronic health records. *J. Am. Board Fam. Med*. 28:65–71

7. Armstrong-Wells J, Johnston SC, Wu YW, Sidney S, Fullerton HJ. 2009. Prevalence and predictors of perinatal hemorrhagic stroke: results from the Kaiser Pediatric Stroke Study. *Pediatrics* 123:823–28

8. Baillargeon J, Paar D, Wu H, Giordano T, Murray O, et al. 2008. Psychiatric disorders, HIV infection and HIV/hepatitis co-infection in the correctional setting. *AIDS Care* 20:124–29

9. Berkowitz SA, Karter AJ, Lyles CR, Liu JY, Schillinger D, et al. 2014. Low socioeconomic status is associated with increased risk for hypoglycemia in diabetes patients: the Diabetes Study of Northern California (DISTANCE). *J. Health Care Poor Underserved* 25:478–90

10. Betancourt T, Scorza P, Kanyanganzi F, Fawzi MC, Sezibera V, et al. 2014. HIV and child mental health: a case-control study in Rwanda. *Pediatrics* 134:e464–72

11. Bild DE, Bluemke DA, Burke GL, Detrano R, Diez Roux AV, et al. 2002. Multi-ethnic study of atherosclerosis: objectives and design. *Am.J. Epidemiol*. 156:871–81

12. Birkhead GS, Klompas M, Shah NR. 2015. Uses of electronic health records for public health surveillance to advance public health. *Annu. Rev. Public Health* 36:345–59

13. Black MH, Smith N, Porter AH, Jacobsen SJ, Koebnick C. 2012. Higher prevalence of obesity among children with asthma. *Obesity* 20:1041–47

14. Blumenthal D, Tavenner M. 2010. The “meaningful use” regulation for electronic health records. *N. Engl. J. Med*. 363:501–4

15. Botros N, Concato J, Mohsenin V, Selim B, Doctor K, Yaggi HK. 2009. Obstructive sleep apnea as a risk factor for type 2 diabetes. *Am. J. Med*. 122:1122–27

16. Brown JL. 2012. A piece of my mind: the unasked question. *JAMA* 308:1869–70