12-17-2013

# Strategies for Handling Missing Data in Electronic Health Record Derived Data

Brian J. Wells
*Cleveland Clinic*, wellsb@ccf.org

Amy S. Nowacki
*Cleveland Clinic*, nowacka@ccf.org

Kevin Chagin
*Cleveland Clinic*, chagink@ccf.org

Michael W. Kattan
*Cleveland Clinic*, kattanm@ccf.org

## Recommended Citation

Wells, Brian J.; Nowacki, Amy S.; Chagin, Kevin; and Kattan, Michael W. (2013) "Strategies for Handling Missing Data in Electronic Health Record Derived Data," *eGEMs (Generating Evidence & Methods to improve patient outcomes)*: Vol. 1: Iss. 3, Article 7.
DOI: http://dx.doi.org/10.13063/2327-9214.1035
Available at: http://repository.academyhealth.org/egems/vol1/iss3/7

The Electronic Data Methods (EDM) Forum is supported by the Agency for Healthcare Research and Quality (AHRQ), Grant 1U18HS022789-01. eGEMs publications do not reflect the official views of AHRQ or the United States Department of Health and Human Services.

# Strategies for Handling Missing Data in Electronic Health Record Derived Data

**Abstract**

Electronic health records (EHRs) present a wealth of data that are vital for improving patient-centered outcomes, although the data can present significant statistical challenges. In particular, EHR data contains substantial missing information that if left unaddressed could reduce the validity of conclusions drawn. Properly addressing the missing data issue in EHR data is complicated by the fact that it is sometimes difficult to differentiate between missing data and a negative value. For example, a patient without a documented history of heart failure may truly not have disease or the clinician may have simply not documented the condition. Approaches for reducing missing data in EHR systems come from multiple angles, including: increasing structured data documentation, reducing data input errors, and utilization of text parsing / natural language processing. This paper focuses on the analytical approaches for handling missing data, primarily multiple imputation. The broad range of variables available in typical EHR systems provide a wealth of information for mitigating potential biases caused by missing data. The probability of missing data may be linked to disease severity and healthcare utilization since unhealthier patients are more likely to have comorbidities and each interaction with the health care system provides an opportunity for documentation. Therefore, any imputation routine should include predictor variables that assess overall health status (e.g. Charlson Comorbidity Index) and healthcare utilization (e.g. number of encounters) even when these comorbidities and patient encounters are unrelated to the disease of interest. Linking the EHR data with other sources of information (e.g. National Death Index and census data) can also provide less biased variables for imputation. Additional methodological research with EHR data and improved epidemiological training of clinical investigators is warranted.

**Acknowledgements**

The authors would like to acknowledge the editorial assistance of Stephanie S. Kocian.

**Keywords**

Methods, Informatics, Health Information Technology

**Disciplines**

Health Services Research | Statistical Methodology

# Strategies for Handling Missing Data in Electronic Health Record Derived Data

Brian J. Wells, MD, PhD; Kevin M. Chagin, MS; Amy S. Nowacki, PhD; Michael W. Kattan, PhD

## Abstract

Electronic health records (EHRs) present a wealth of data that are vital for improving patient-centered outcomes, although the data can present significant statistical challenges. In particular, EHR data contains substantial missing information that if left unaddressed could reduce the validity of conclusions drawn. Properly addressing the missing data issue in EHR data is complicated by the fact that it is sometimes difficult to differentiate between missing data and a negative value. For example, a patient without a documented history of heart failure may truly not have disease or the clinician may have simply not documented the condition. Approaches for reducing missing data in EHR systems come from multiple angles, including: increasing structured data documentation, reducing data input errors, and utilization of text parsing / natural language processing. This paper focuses on the analytical approaches for handling missing data, primarily multiple imputation. The broad range of variables available in typical EHR systems provide a wealth of information for mitigating potential biases caused by missing data. The probability of missing data may be linked to disease severity and healthcare utilization since unhealthier patients are more likely to have comorbidities and each interaction with the health care system provides an opportunity for documentation. Therefore, any imputation routine should include predictor variables that assess overall health status (e.g. Charlson Comorbidity Index) and healthcare utilization (e.g. number of encounters) even when these comorbidities and patient encounters are unrelated to the disease of interest. Linking the EHR data with other sources of information (e.g. National Death Index and census data) can also provide less biased variables for imputation. Additional methodological research with EHR data and improved epidemiological training of clinical investigators is warranted.

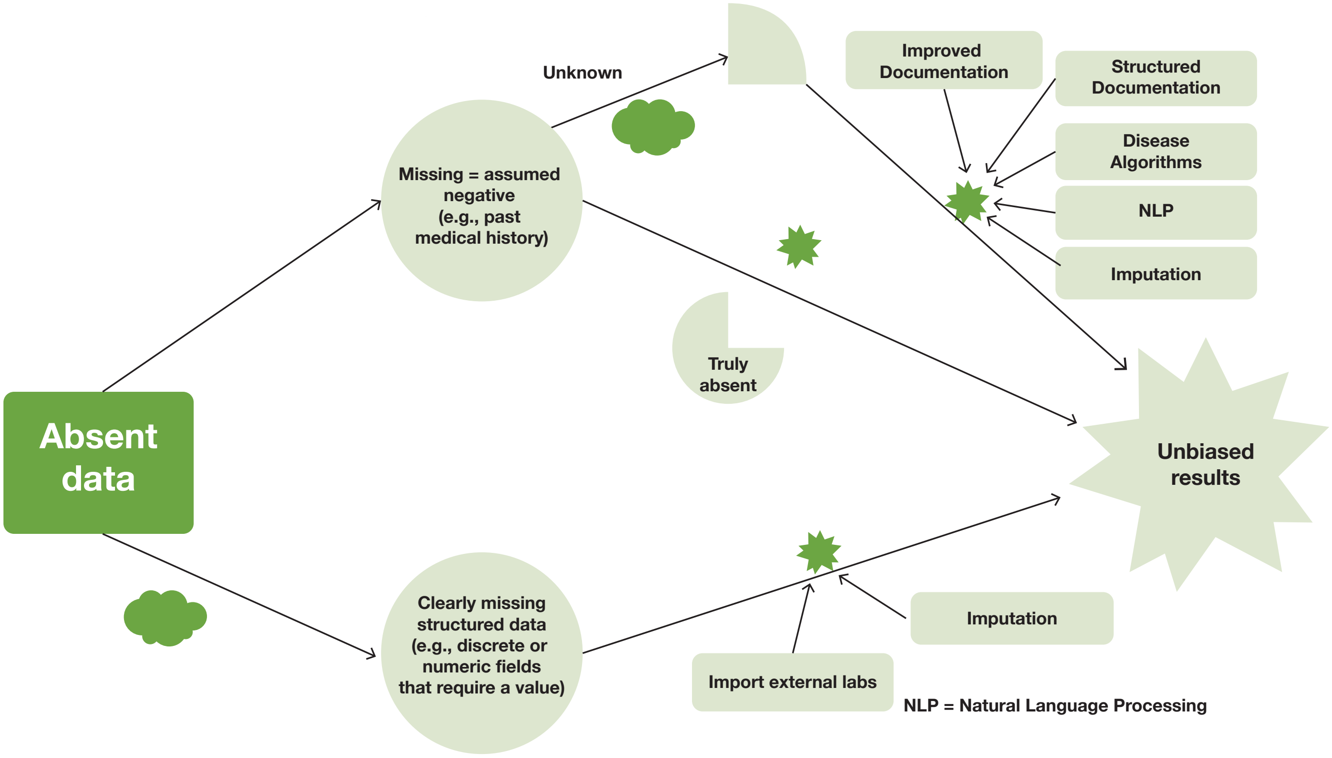
## Introduction

In the 2004 State of the Union Address, President Bush called for the adoption of Electronic Health Records (EHRs); and the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 incentivized the "meaningful use" of EHRs.[1] The use of at least partial EHRs by outpatient offices in the United States has increased from 16 percent in 2003 to 52 percent in 2010.[2] The increased use of EHRs has greatly enlarged the volume of data that are readily available to clinicians. The same EHR data are progressively used for quality improvement projects and research initiatives.

EHR data present significant challenges as a result of not having been collected specifically for research purposes and are subject therefore to considerable missing data. Improper handling of missing data can result in significant bias. These issues will become more relevant as EHRs become more widespread, as the volume of discrete electronic data continues to grow, and as the access to this data becomes increasingly available. Prior to the adoption of EHR systems many details of clinical practice were obscured, with the exception of relatively small chart reviews. Chart reviews, however, can be painstaking tasks and are complicated by poor handwriting, missing charts or pages, items being documented in inconsistent locations, and the inability for multiple people to review the chart simultaneously. The use of an EHR overcomes all of these obstacles, and the reduced cost of EHR data compared to traditional research is an attractive option. These features, coupled with an increased access to EHR data, have attracted more "evening and weekend" researchers lacking extensive biostatistical and epidemiological training. EHR data present significant challenges and—in order to reduce bias—the analyst must have a solid understanding of observational study designs, confounding, missing data, time-to-event analyses, and multiple regression. Increased epidemiological training of medical students, nurses, and other health care professionals who analyze EHR data is necessary, because past experience shows this training is frequently lacking. The goal for this paper is to highlight a number of key issues involving missing data in EHR research, to examine some of the unique aspects of missing data in EHR systems, to present some statistical advice about how to handle these issues according to "best practices," and to provide some suggested areas for future methodological research.

Cleveland Clinic

**Figure 1. Overview of the missing data problem with electronic health records.**


Absent data
Missing = assumed negative (e.g., past medical history)
Unknown
Truly absent
Clearly missing structured data (e.g., discrete or numeric fields that require a value)
Improved Documentation
Structured Documentation
Disease Algorithms
NLP
Imputation
Unbiased results
Imputation
Import external labs
NLP = Natural Language Processing

## Understanding Data Missingness in EHR Systems

EHR data elements can be divided into two broad categories: structured and textual data. Structured data primarily consists of quantifiable numeric values (e.g., systolic blood pressure) and discrete elements made up of predefined categories (e.g., diagnoses codes based on ICD-9). Textual data (narrative data) are the free text areas of the patient chart (e.g., physician progress note) that are difficult to analyze quantitatively due to the breadth of human expression, grammatical errors, the use of acronyms and abbreviations, and the potential for different interpretations of the same phrase depending on context. Structured data is very amenable to statistical analyses because it can be stored in structured databases that allow for reliable data retrieval. Analysis of textual data involves a specialized branch of computer science called "Natural Language Processing" (NLP). The authors envision NLP as a tool for augmenting information in structured data, but argue that the capture of structured data, whenever possible, would maximize the potential impact of EHR systems on research and patient care activities.

Whether an EHR based analysis utilizes structured data or textual data, missing data will likely be of concern and can come in a variety of forms. Missing data can be due to a lack of collection (e.g., patient was never asked about asthma) or a lack of documentation (e.g., patient was asked about asthma but the response was never recorded in the medical record). Lack of documentation is particularly common when it comes to a patient not having a symptom/comorbidity. Instead of recording a negative value for each potential symptom/comorbidity, all data fields are left blank (missing) and only the positive values are recorded. Thus it can be impossible to differentiate between the lack of a comorbidity, the lack of documentation of a comorbidity and the lack of data collection regarding the comorbidity. In order to conduct research using EHR data, it is typically necessary to assume that missing data elements indicate a negative value. This assumption may be substantially violated, especially for diseases that are frequently not documented. Figure 1 is an overview of the missing data issue in EHR systems and provides several options for mitigating this issue. From a research perspective, the preferred situation is to improve the completeness and accuracy of discrete data documentation in EHR records. Some clinical care providers complain that more complete discrete electronic documentation requires additional work; sometimes without immediate, obvious benefits to patient care. However, it seems a poor argument to suggest that it is acceptable to have incomplete information in the EHR even for clinical purposes. Much of the data referred to in this paper are elements that should be collected in the course of routine patient care (e.g., past medical history, family history, current medications, etc.) One simple example where structured data could help diabetes research at our institution involves the length of time that patients fast prior to having blood drawn for a metabolic panel. This information is frequently collected from patients but is recorded on paper. This makes it very difficult to determine if a blood sugar value meets the American Diabetes Association criteria for a diagnosis of diabetes. This information is also important for patient care, and recording it in the EHR as a numeric value could make it easier for clinicians to access it.

DOI: 10.13063/2327-9214.1035

Absent data in EHR records can decrease the ability to create accurate predictions. Methods for mitigating missing data are difficult because many "NULL" values are assumed to be negative. Improving documentation, employing disease algorithms (phenotypes), and natural language processing are methods for decreasing the impact of potential biases.

## Understanding Ways Missing Data Affect our Understanding of Patient Care and Quality

Do EHRs improve patient care? In most studies, electronic prescription ordering has resulted in a reduction of medication errors and adverse drug events for both inpatient and outpatient practice.[3-4] For quality measures, the literature on this topic is mixed. Initial analyses of the National Ambulatory Medical Care Survey found no difference in quality between paper- and EHR-based practices and found no improvement in quality with the use of decision support systems.[5-6] However, more recent studies have shown that EHRs are associated with better care for patients with diabetes in Cleveland[7] and better overall quality of care on nine quality measures in New York state.[8] Cebul et al. hypothesize that one reason for a difference in results may be due to newer, more robust EHR systems.[7]

Regardless of its impact on quality, complete and structured EHR data can make patient care easier by allowing previously entered data to be reused and easily found by other providers. In the "National Priorities for Research" publication, the Patient-Centered Outcomes Research Institute has placed a priority on research that "seeks to improve the volume, completeness, comprehensiveness, accuracy, and efficiency of use of clinical data collected across healthcare systems…"[9] Complete, structured EHR data also make administrative duties, quality improvement projects, and research projects much simpler to conduct. Structured data entry can also lead to more complete and accurate documentation. Electronic data entry forms can require fields to be completed, prompt the user for more information, alert the user about numerical values that are outside of a reference range, prevent typographical errors with spellchecking, and enforce standardized nomenclature by using drop-down lists. Structured data entry is becoming easier for clinicians. Applications like Epic's NoteWriter (Epic Systems Corporation, Verona, WI) help build progress notes using a point-and-click interface, and new practice models are tasking medical assistants and nurses to assist with data gathering and visit note documentation.[10]

In order to improve the accuracy of disease classification using structured EHR data (e.g., laboratory values, medications, orders, diagnoses) standardized disease definitions and numerous disease classification algorithms (phenotypes) have been developed (e.g., hypertension can be established not only by diagnosis code, but also through office-based blood pressure readings and prescriptions for antihypertensive medications). Disease definitions based on clinical records are becoming more formalized with initiatives like the Chronic Conditions Warehouse definitions created by the Centers for Medicare and Medicaid.[11] Alternatively, an information-technology-based solution includes automatically generated algorithms based on structured data[12] as well as natural language processing (NLP) for textual elements of the EHR.

It is anticipated that discrete coding and, hence, completeness and accuracy will continue to improve as documentation requirements continue to expand for quality-based reimbursement**,** EHR software becomes increasingly sophisticated, and voice recognition software becomes more universal. "Smart" computer-assisted data entry tools that would prompt clinicians to agree to automatic discrete data entry based on NLP are envisioned for use in the future. For example, as the clinician creates a free text note using voice recognition software the tool could ask, "Would you like me to add diabetes to the patient's past medical history list?" Another underutilized source for discrete data elements is through direct patient-to-EHR communication. For example, the Cleveland Clinic EHR system based on Epic includes a patient portal called "My Chart" that allows patients to view medical records, schedule appointments, receive health maintenance reminders, and have limited direct communication with their physician. A new initiative at the Cleveland Clinic, called "MyFamily," is piloting a tool that gathers family medical history information directly from patients (http://my.clevelandclinic.org/cph/education/newsletter/cph-winter2013.aspx#2). Physicians or other health care professionals would likely have to review and approve any patient entered information prior to acceptance of this data into the EHR, but questionnaires regarding the presence or absence of disease like the National Health and Nutrition Examination Survey have found that patient report is fairly reliable, at least for hypertension.[13]

## Strategies to Address Missing Data

While improvements can be made in the data collection process, it is unreasonable to believe that this will eliminate all missing data. Therefore, it is important to understand the kind of missing data present in EHRs and the methodologies that are available to address this. Statistically speaking, missing data falls into one of three categories: missing-completely-at-random (MCAR), missing-at-random (MAR), and not-missing-at-random (NMAR).[14-15] MCAR refers to a situation in which the probability that a data point is missing is unrelated to the value of that data point or to the value of any other variable(s). Thus, if a patient's data were missing because he missed his appointment (i.e., data collection opportunity) after his car was struck by a meteor, his data would presumably be MCAR. Here, any piece of data is just as likely to be missing as any other piece of data and, while one may lose power for the analysis, the estimated parameters are not biased by absence of the data. MAR refers to a situation in which the probability that a data point is missing does not depend on the value of that data point after controlling for all other known variable(s). For example, healthier patients are less likely to utilize the healthcare system and may be more likely to have missing data such as systolic blood pressure readings. Therefore, systolic blood pressure may have a direct, positive relationship with the number of office visits in a univariate analysis, but adequate adjustment for current health status would make this relationship between office visits and systolic blood pressure disappear. NMAR refers to the extreme situation in which the probability that a data point is missing depends entirely on the value of that data point or on the value of other unmeasured variable(s). For example, if a rogue laboratory technician refuses to enter a test result into the records of patients with green eyes this would represent a NMAR situation. Obtaining unbiased estimates of the parameters in the face of NMAR is difficult. However, it is difficult to imagine situations in real life where 100 percent of the variability in missing information is either NMAR or MCAR. Methods for handling missing data that fall into each of these categories are described next.

The simplest and most common approach to handling missing data is to omit the cases with missing data and to run the analysis on what remains. This is referred to as "complete case analysis" or "listwise deletion." Results obtained from complete case analysis are unbiased when the data are MCAR, however, there is a loss in power. The primary argument against complete case analysis is that patients with missing information are systematically different than patients with complete data (i.e., data that is not MCAR). This argument is almost always plausible when working with EHR data (e.g., compliant patients, patients with good insurance, and patients with more severe disease tend to have less missing data). Therefore, excluding patients with missing data will introduce bias.

Instead of omitting patients with missing data, one can fill in or impute missing data points and include all patients in the analysis. Imputation of missing data in the setting of MCAR will increase power but should not change the point estimates from those obtained with a complete case analysis. Obtaining unbiased estimates in the face of NMAR requires the creation of a model that accounts for the missing data that is incorporated into the imputation process.[16] In most EHR instances, other variables collected would be expected to explain some, but not all, of the variation in missingness between patients. Therefore, it seems reasonable to perform imputations using EHR data under the assumption of MAR. Standard imputation of missing EHR data that qualifies as NMAR without an NMAR model might produce biased estimates but the bias should be small. This position is supported by a paper by Lin (2006) that demonstrated that prediction models created with EHR data that included information regarding missingness performed better.[17] In other words, there was an association between the simple presence or absence of data with the outcome. It is difficult to test the MAR assumption in retrospective EHR data without contacting the patients directly. However, Rubin et al. emphasize that as the number of covariates utilized to perform the imputation increase, the MAR assumption becomes increasingly plausible.[18] Thus, MAR missing data can generally be imputed with a certain degree of accuracy as long as adequate covariates are included in the regression equations used in the imputation process.[19] Since the amount of exposure to the health care system is associated with missingness (patients with more visits have had more opportunities for documentation) it is recommended that the degree of health care utilization (number of encounters) and disease severity (comorbidity index) be included as covariates in the imputation process. Validated comorbidity indices have been created that are based solely on ICD-9 and ICD-10 coding schemes.[20] In addition, the inclusion of follow-up time and outcome information have also been recommended as covariates and seem to improve the accuracy of results obtained using multiple imputation.[21-22]

The most popular method for imputing missing values is through a process called "multiple imputation using chained equations" (MICE).[23] What makes MICE popular is its ability to impute different variable types (i.e., continuous, unordered and ordered categorical, etc.) that may reside in the EHR, since each variable is imputed by its own model. MICE does this by approximating the posterior predicted distribution of each variable by regressing it on all other remaining variables. The first variable with missing observations, $x_1$, is regressed on all remaining variables within the EHR data set, $x_2, \ldots, x_k$, where k is the total number of variables in the EHR data set. The missing values for the variable $x_1$ are replaced with the predicted values produced by the regression model. The imputation process is continued by creating regression models for each variable sequentially and inserting predicted values into the missing data slots until all missing values have been imputed exactly once for the first iteration. It should be noted that imputed data are included in the regression equations for subsequent imputations. Successive iterations are performed to re-impute and replace imputed values from previous iterations in order to obtain a stable estimate for each missing data point. As long as a sufficient number of iterations have been performed, the order in which the variables are imputed is irrelevant.[24] The MICE package in R uses five iterations for each imputation according to its default setting. This whole process is then repeated *m* times to give you *m* imputed data sets. Although other methods besides multiple imputation exist for handling missing data, they are outside the scope of this paper and it can be argued that multiple imputation is the best method for handling missing data in most instances.[13]

One statistical methodology question regarding the use of imputation involves the number of imputed data sets. How many imputed data sets are required to obtain stable estimates? In order to estimate the number, it is necessary to understand the difference between "missing data" and "missing information." "Missing data" refers to the specific number of missing data points for any particular parameter, while "missing information" is estimated by comparing the variation in results obtained across multiple imputed data sets. Sometimes there are a lot of missing data but other related variables are available, so the results of imputed data sets are quite stable (little variability); thus, there is trivial missing information. For example, Pantalone et al. examined information derived from the Cleveland Clinic EHR about mortality differences among a cohort of patients with type 2 diabetes who were receiving metformin plus one of several possible sulfonylureas. The statistical model contained four variables with >50 percent missing values that included hemoglobin A1c. Multiple imputation was performed using regression equations that were built with all other variables in the data set, some of which had very few or no missing values (e.g. age, gender, race, smoking). The hazard ratios estimating the mortality risk according to sulfonylurea type were found to be missing < 1.5 percent of the total information expected with complete data.[25] Alternatively, there may be a lot of missing data but no related variables are available, so the results of imputed data sets are unsteady (large variability); thus, there is considerable missing information. Rubin's formula for estimating missing information has been used to determine the number of imputed data sets required to achieve a given level of efficiency (typically 95 percent) for each specific coefficient of interest.[26]

Since the amount of exposure to the health care system is associated with missingness (patients with more visits have had more opportunities for documentation), it is recommended that the degree of health care utilization (e.g., number of encounters), markers of disease severity, comorbid conditions, and social economic status be included as covariates in the imputation process. Comorbid conditions may simply include diagnoses codes for relevant disease entities or formal, validated comorbidity scales that have been adapted for use with ICD-9 and ICD-10 coding schemes in administrative data.[20] In addition, the inclusion of follow-up time and outcome information have also been recommended as covariates and seem to improve the accuracy of results obtained using multiple imputation.[21-22] Markers of disease severity depend on the condition being studied but could include items like serum hemoglobin A1c (HbA1c) values for patients with diabetes or serum cholesterol levels for patients with hyperlipidemia. SES markers already in the EHR such as primary insurance type (e.g., private, Medicare, Medicaid) are generally relatively crude indicators of SES. A more precise estimate of SES can be obtained by using Geographical Information Systems (GIS) software like ArcGIS (Esri, Redlands, CA) to geocode patient addresses and link this with data in the American Community Survey (ACS) that is available at the Census Block Group level.[27] ACS data types include median income, education level, and home value. Once addresses are available, more sophisticated GIS analyses can be performed as necessary to determine items like distance to emergency department. Once the address has been linked to the ACS data the address can be removed and the data set can be de-identified. This process would not work if the institution de-identified data prior to making it available for research unless there is a mechanism for re-identifying patients to obtain addresses. Obviously, ACS data is not available for patients with addresses outside of the United States. In addition, some addresses may not be found during the geocoding process due to inaccurate address information, incorrect formatting, or gaps in the GIS address coverage. If geocoding does not find an exact address match it is still possible in many cases to use the less accurate income based on a five-digit ZIP code in the ACS. In our experience at the Cleveland Clinic, 80–90 percent of patients can be mapped to their exact address, and an additional 5–10 percent of patients can be mapped using a five-digit ZIP. This generally leaves <15 percent missing income data that requires imputation.

## Measuring Outcomes

EHR systems present particular challenges when it comes to assessing outcomes for a variety of reasons. First, outcomes may not be discretely entered into a diagnosis section of the patient's chart in a timely fashion. This can occur because the patient sought care outside of the EHR's health care system, did not seek treatment for the condition, the provider did not enter the information or the patient expired. Our general principles for measuring outcomes are as follows:

(1). Seek outcomes that are ascertained outside of the EHR system in a fashion that is not biased toward any subset of the population. A favorite example of this is the Social Security Death Index (SSDI), which determines the date or fact of death for any individual in the United States who has a social security number even if they relocate. Unfortunately, the Social Security Administration changed its policy in 2011 such that it will no longer use state death records to identify deaths. This policy shift means that SSDI is a less sensitive method for identifying mortality after November 1, 2011. This issue does not have an impact on the National Death Index (NDI) maintained by the Centers for Disease Control and Prevention, but the NDI data is not free. Medicare and Medicaid data that has been linked to the NDI by the National Center for Health Statistics is also unaffected. Death data can still be obtained from individual state records in some instances. Other external sources of outcome information include state and local tumor registries, transplant registries, and disease registries maintained within a respective health care institution.

(2). Utilize as many different sources of data as possible to ascertain outcome information to increase sensitivity (i.e., an event is counted if the outcome is identified in any of the sources). For example, Kumbhani et al. used EHR data to create a model for predicting outcomes following Acute Coronary Syndromes. In order to improve the sensitivity for capturing cardiac outcomes the researchers included any of the following findings: new encounter diagnosis for myocardial infarction using ICD-9, serum troponin-T $\geq$0.08 mcg/L, or new current procedural terminology (CPT) code for a coronary revascularization procedure.[28] This increases the sensitivity for capturing the outcome, and none of these items are mutually exclusive. Similarly, we have found that some office procedures are not always documented in the order/procedure section of our EHR. For example, a procedure performed by the physician (e.g., excision of a skin lesion) does not require an order to be placed. In this instance, the only structured location for this item is via CPT code in billing records. In some instances it may be necessary to manually review diagnostic test results in order to determine if the outcome occurred. The simple presence of a diagnostic test may determine the pool of patients who may have developed the outcome (i.e., if a condition requires a specific test in order to make the diagnosis, then it may be safe to assume that patients who did not undergo the test did not develop the outcome of interest). This can help to limit the extent of any manual chart review.

(3). Validate the results both formally and informally: Do the number of events make sense clinically, and are the number of events consistent with publications in the scientific literature? Manual chart reviews should also be performed on a random subset of patients to determine the sensitivity and specificity of the outcome ascertainment.

(4). Patients should be censored at the time of their last follow-up in the EHR system unless the outcome is determined by regional or national data sets that are not dependent upon the EHR. Censoring helps to decrease potential bias caused by not capturing events that may occur at another health care institution or in folks who die suddenly. National data sets such as the NDI should also reduce bias by capturing deaths that occur outside of the local EHR institution. The NDI has the added benefit of allowing longer follow-up times in the survival analysis and thus increasing statistical power. Previous studies have found that the NDI captures 87–98 percent of deaths in the United States.[29]

## Discussion

EHR data have obvious limitations when utilized for research purposes. These limitations include inaccurate information and lack of study specific variables. Thus, EHR data will never eliminate the need for research involving primary data collection. However, EHR research can still lead to important knowledge from investigations that might otherwise be unfeasible or too costly to perform. Using a limited number of common discrete data types available in the Cleveland Clinic EHR (e.g., age, gender, race, smoking, labs, vitals, medications, diagnoses) researchers are successfully publishing a variety of research articles.[28,30-33] And it's important to note that despite the flaws in these data they have succeeded in creating accurate risk prediction models.

The increasing volume of EHR data has greatly expanded the availability of data for research purposes, and methodologies for handling these data are rapidly changing. Data in the EHR provide an opportunity to improve patient outcomes through research and the development of clinical decision support tools, but the issue of missing data is raised.  Increased use of disease phenotyping and natural language processing can decrease missing EHR data used for research. Improved documentation methods that decrease missing data without overburdening clinicians should improve patient care and research simultaneously. Researchers using these data need to be well versed in biostatistical methods regarding missing data. Fortunately, EHRs tend to contain many data points and frequently can be linked with outside sources that can help to reduce possible confounding, can improve imputation, and can help to capture the outcome(s) of interest.

## Acknowledgements

The authors would like to acknowledge the editorial assistance of Stephanie S. Kocian.

## References

1. Blumenthal D. Launching HITECH, *N Engl J Med* 2010;362:382-385.
2. Kokkonen EW, Davis SA, Lin HC, Dabade TS, Feldman SR, Fleischer AB,Jr. Use of electronic medical records differs by specialty and office settings, *J Am Med Inform Assoc* 2013;20:e33-8.
3. Ammenwerth E, Schnell-Inderst P, Machan C, Siebert U. The effect of electronic prescribing on medication errors and adverse drug events: a systematic review, *J Am Med Inform Assoc* 2008;15:585-600.
4. Devine EB, Hansen RN, Wilson-Norton JL, Lawless NM, Fisk AW, Blough DK, Martin DP, Sullivan SD. The impact of computerized provider order entry on medication errors in a multispecialty group practice, *J Am Med Inform Assoc* 2010;17:78-84.
5. Romano MJ, Stafford RS. Electronic health records and clinical decision support systems: impact on national ambulatory care quality, *Arch Intern Med* 2011;171:897-903.
6. Linder JA, Ma J, Bates DW, Middleton B, Stafford RS. Electronic health record use and the quality of ambulatory care in the United States, *Arch Intern Med* 2007;167:1400-1405.
7. Cebul RD, Love TE, Jain AK, Hebert CJ. Electronic health records and quality of diabetes care, *N Engl J Med* 2011;365:825-833.
8. Kern LM, Barron Y, Dhopeshwarkar RV, Edwards A, Kaushal R, HITEC Investigators. Electronic health records and ambulatory quality of care, *J Gen Intern Med* 2013;28:496-503.
9. Patient-Centered Outcomes Research Institute (PCORI). National Priorities for Research and Research Agenda [Internet]. Washington, D.C.: Patient-Centered Outcomes Research Institute (PCORI) May 21, 2012 cited December 21, 2013]; [21]. Available from: http://www.pcori.org/assets/PCORI-National-Priorities-and-Research-Agenda-2012-05-21-FINAL1.pdf English.
10. Anderson P, Halley MD. A new approach to making your doctor-nurse team more productive, *Fam Pract Manag* 2008;15:35-40.
11. Schneider KM, O'Donnell BE, Dean D. Prevalence of multiple chronic conditions in the United States' Medicare population, *Health Qual Life Outcomes* 2009;7:82.
12. Wright A, Pang J, Feblowitz JC, Maloney FL, Wilcox AR, Ramelson HZ, Schneider LI, Bates DW. A method and knowledge base for automated inference of patient problems from structured data in an electronic medical record, *J Am Med Inform Assoc* 2011;18:859-867.
13. Vargas CM, Burt VL, Gillum RF, Pamuk ER. Validity of self-reported hypertension in the National Health and Nutrition Examination Survey III, 1988-1991. *Prev Med* 1997;26:678.
14. Schafer JL, Graham JW. Missing data: our view of the state of the art, *Psychol Methods* 2002;7:147-177.
15. Rubin DB. Inference and missing data, *Biometrika* 1976;63:581-592.
16. Dunning T, Freedman D. Handbook of Social Science Methodology: Modeling selection effects, 2008.
17. Lin JH, Haug PJ. Exploiting missing clinical data in Bayesian network modeling for predicting medical problems, *J Biomed Inform* 2008;41:1-14.
18. Rubin DB, Stern HS, Vehovar V. Handling "don't know" survey responses: the case of the Slovenian plebiscite, *Journal of the American Statistical Association* 1995;90:822-828.
19. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures, *Psychol Methods* 2001;6:330-351.
20. Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi JC, Saunders LD, Beck CA, Feasby TE, Ghali WA. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data, *Med Care* 2005;43:1130-1139.
21. Royston P. Multiple imputation of missing values, *Stata Journal* 2004;4:227-241.

22. Moons KG, Donders RA, Stijnen T, Harrell FE,Jr. Using the outcome for imputation of missing predictor values was preferred, *J Clin Epidemiol* 2006;59:1092-1101.
23. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice, *Stat Med* 2011;30:377-399.
24. Buuren S, Groothuis-Oudshoorn K. MICE: Multivariate imputation by chained equations in R, *Journal of statistical software* 2011;45.
25. Pantalone KM, Kattan MW, Yu C, Wells BJ, Arrigain S, Nutter B, Jain A, Atreja A, Zimmerman RS. The risk of overall mortality in patients with Type 2 diabetes receiving different combinations of sulfonylureas and metformin: a retrospective analysis, *Diabet Med* 2012;29:1029-1035.
26. Rubin DB. Multiple Imputation for Nonresponse in Surveys. Wiley: New York, 1987:258.
27. Alexander C. American Community Survey, *Encyclopedia of the US Census* 2000:26-28.
28. Kumbhani DJ, Wells BJ, Lincoff AM, Jain A, Arrigain S, Yu C, Goormastic M, Ellis SG, Blackstone E, Kattan MW. Predictive models for short- and long-term adverse outcomes following discharge in a contemporary population with acute coronary syndromes, *Am J Cardiovasc Dis* 2013;3:39-52.
29. Cowper DC, Kubal JD, Maynard C, Hynes DM. A Primer and Comparative Review of Major U.S. Mortality Databases, *Ann Epidemiol* 2002;12:462-468.
30. Pantalone KM, Kattan MW, Yu C, Wells BJ, Arrigain S, Jain A, Atreja A, Zimmerman RS. Increase in overall mortality risk in patients with type 2 diabetes receiving glipizide, glyburide or glimepiride monotherapy versus metformin: a retrospective analysis, *Diabetes Obes Metab* 2012;14:803-809.
31. Wells BJ, Roth R, Nowacki AS, Arrigain S, Yu C, Rosenkrans Jr WA, Kattan MW. Prediction of morbidity and mortality in patients with type 2 diabetes, *PeerJ* 2013;1:e87.
32. Navaneethan SD, Jolly SE, Sharp J, Jain A, Schold JD, Schreiber MJ,Jr, Nally JV,Jr. Electronic health records: a new tool to combat chronic kidney disease? *Clin Nephrol* 2013;79:175-183.
33. Navaneethan SD, Jolly SE, Schold JD, Arrigain S, Saupe W, Sharp J, Lyons J, Simon JF, Schreiber MJ,Jr, Jain A, Nally JV,Jr. Development and validation of an electronic health record-based chronic kidney disease registry, *Clin J Am Soc Nephrol* 2011;6:40-49.