HHS Public Access

Author manuscript

Curr Epidemiol Rep. Author manuscript; available in PMC 2022 May 27.

Published in final edited form as:

Curr Epidemiol Rep. 2018 December ; 5(4): 343–356. doi:10.1007/s40471-018-0164-x.

Measurement error and misclassification in electronic medical records: methods to mitigate bias

Jessica C. YOUNG, MSPH[*], **Mitchell M. CONOVER, PhD**[*], **Michele JONSSON FUNK, PhD**

[1]Department of Epidemiology, UNC Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Abstract

PURPOSE OF REVIEW: We sought to: 1) examine common sources of measurement error in research using data from electronic medical records (EMR), 2) discuss methods to assess the extent and type of measurement error, and 3) describe recent developments in methods to address this source of bias.

RECENT FINDINGS: We identified eight sources of measurement error frequently encountered in EMR studies, the most prominent being that EMR data usually reflect only the health services and medications delivered within the specific health facility/system contributing to the EMR data. Methods for assessing measurement error in EMR data usually require gold standard or validation data, which may be possible using data linkage. Recent methodological developments to address the impact of measurement error in EMR analyses were particularly rich in the multiple imputation literature.

SUMMARY: Presently, sources of measurement error impacting EMR studies are still being elucidated, as are methods for assessing and addressing them. Given the magnitude of measurement error that has been reported, investigators are urged to carefully evaluate and rigorously address this potential source of bias in studies based in EMR data.

Keywords

Electronic medical records; Measurement error; Misclassification; multiple imputation for measurement error; Pharmacoepidemiology; Comparative effectiveness; Real world evidence

Introduction

Medical record data offer key clinical details and include information on some aspects of health services that are not well-captured by other secondary data sources. Increasing adoption of electronic medical record (EMR) systems in both ambulatory and inpatient

Corresponding Author Michele Jonsson Funk, PhD, 104B Market Street, CB #7521, Chapel Hill, NC 27599, mfunk@unc.edu, Telephone: 919-843-0384, Fax: 919-843-7364.

[*]Both authors contributed equally to this work.

Conflict of Interest

Dr. Jonsson-Funk, Dr. Conover and Ms. Young each declare that they have no conflict of interest.

Human and Animal Rights and Informed Consent

This article does not contain any studies with human or animal subjects performed by any of the authors.

clinical care settings has provided health researchers greater access to these data. EMR data often cover large and diverse patient populations, allowing for healthcare research to be conducted in a timely manner. However, their primary function is to facilitate clinical care, and thus their secondary use for pharmacoepidemiology research requires careful consideration. Medical coding and documentation in EMR databases is often driven by factors outside of clinical care, such as required fields in the EMR system, insurance reimbursement policies, and automated importation of historical information. These factors can drive systematic bias in these data, adding complexity to their use for healthcare research. Given their increasing use in pharmacoepidemiology research, we sought to examine common sources of measurement error in EMR data, discuss methods to assess the extent and type of error, and describe recently published methodological developments intended to address resulting biases. Throughout this paper, we define misclassification and missing data as special cases of measurement error.

Sources of Measurement Error

We identified eight domains of commonly encountered sources of measurement error and/or missing data that may affect research using EMR data:

1. EMR data reflect only the health services and medications delivered within the specific health care setting that contributes to the EMR system.[1-6] This leads to both left and right censoring, and uncertainty regarding the person-time at risk. This is particularly problematic in inpatient EMRs.
2. Prescription records in an ambulatory EMR reflect clinician orders for medications, which may not be filled or consumed by the patient.[7-9]
3. In EMR studies, defining treatment episodes / treatment duration / cumulative exposure is complex and requires many decisions which have unpredictable influence on exposure misclassification.[10-12]
4. Automated data entry in EMR systems may forward-propagate erroneous data and/or carry forward information that is no longer clinically relevant.[13-15]
5. Recent advances in natural language processing (NLP), which automate extraction of information from unstructured data, may introduce systematic errors.[16-19]
6. Performance of EMR-based clinical prediction algorithms may vary widely between different health systems.[20]
7. Temporal changes in the recording of EMR data elements may produce systematic differences in classification and/or missingness over time.[21]
8. Horizontal linkage of populations captured by different EMR systems produce systematic differences in classification and/or missingness between the linked populations.[22]

In Table 1, we summarize studies that characterized various sources of measurement error that are commonly encountered in studies conducted using EMR data. For each source of

measurement error, we identify the source of measurement error being described, how they assessed the problem, and summarize the key findings and/or proposed solutions.

Since the data reflected in EMR are a complex function of factors such as clinical context, organization structure, health business relationships, and patient privacy relations, EMR data across health providers have systematic differences in implementation and structure. Additionally, ambulatory and inpatient EMR systems tend to vary on another level in complexity and types of data recorded. While the majority of this paper pertains generally to commonly observed aspects of EMR data, we highlight specific areas which may be more relevant in ambulatory or inpatient settings.

Assessing Measurement Error

Identifying Erroneous Values using Validation Studies

When key variables may have been measured with error in the full study population, investigators frequently acquire additional data in a subset of individuals that can serve as an *alloyed* gold standard: an imperfect but still useful/superior indicator of the true value. To date, the most common alloyed gold standard used to assess the validity of EMR data is manual abstraction of paper (or electronic) charts by a clinician. However, this method is expensive and sometimes subject to the same forms of measurement error affecting EMR.[23] Another common approach is to validate EMR data against self-report data (e.g. survey or interview data), though such data is rarely available for large patient populations.[24-27]

Common statistics used to quantify misclassification for dichotomous variables include sensitivity, specificity, positive predictive value, and negative predictive value.[28] Wang et al. used simulations to illustrate that under the assumption of no false positives, if misclassification of a variable is independent of its true value, measures of sensitivity are unbiased, however specificity will be underestimated.[1] If misclassification of an exposure or confounder is not independent of the outcome, then the bias in sensitivity and specificity are related to the association between misclassification and the outcome. The authors derive bias-corrected estimators of sensitivity and specificity under the condition that misclassification is independent of outcome status. However, these formulas require outcome prevalence and misclassification rate, statistics which may not be easily assessed if data is misclassified. Thus a priori knowledge is frequently used to guide sensitivity analyses and generate plausible ranges of values.

In practice, validation studies most frequently report the positive predictive value or the c-statistic (i.e. the area under the ROC curve) after sampling all potential cases.[29] However, unless accompanied by additional statistics and measurements (e.g. prevalence), both the c-statistic and the positive predictive value cannot be easily used to employ methods meant to repair or assess the impact that measurement error has on estimates (e.g. bias analysis).[30] Furthermore, estimates of positive and negative predictive value tend to be less transportable than estimates of sensitivity and specificity, since the former are functions of prevalence.

Identifying Erroneous Values using Vertical Linkage

Linked administrative billing claims are an alternative, complimentary source of information to the data available in EMR systems. While claims are frequently available for large patient populations, linkage can be logistically challenging. However claims have a number of complimentary advantages: 1) claims identify care that occurs outside of the health system that the EMR data was drawn from, 2) claims indicate whether the patient had insurance coverage that determines eligibility for certain medical services, and 3) claims are more likely to be standardized across facilities than EMR, which are primarily intended to facilitate care within a given facility or health system. Since 2010, multiple studies have explored linkage of various EMR sources (single-center EMR systems, OptumLabs Data Warehouse[31], Sentinel[32]) to large population-based claims sources, including commercial insurance providers [33], Medicare[3], PC-Rx [31], Medicaid [34, 35], and Sentinel [32]).

Identifying Erroneous Values using Repeated Measures

There are some special settings under which measurement error may be evaluated without a gold standard for comparison. Recently developed methods provide some useful, albeit limited, tools to investigators seeking to identify erroneous or outlying values in longitudinal data (i.e. settings where a continuous variable is repeatedly measured). In 2016, Yang et al. described the conditional growth percentile method, which flags outlier observations by comparing the actual value of a continuous variable (e.g. body mass) to an expected value estimated using time-dependent hierarchical models.[36] In 2018, Shi et al. developed a related method for identifying errors in longitudinal data which seeks to determine which of two measurements is erroneous when a clinically implausible change occurs between two consecutive measurements.[37] Though Shi's method out-performed Yang's in terms of both sensitivity and specificity, it requires investigators be able to define clear rules for what constitutes a clinically implausible change.

Addressing Measurement Error

Bias analysis

Historically, methods for addressing measurement error in epidemiologic studies have relied heavily on bias analysis, most notably quantitative bias analysis. A more thorough review of the diverse methods that make up the bias analysis literature is available elsewhere.[30] However, compared to other methods used to address measurement error, bias analysis has seen relatively little recent development. This may be a result of the aforementioned difficulty of obtaining measures of sensitivity and specificity to inform bias analyses. Further research developing and applying bias analysis methods within EMR studies is needed.

However, two recent bias analysis studies merit brief mention. First, 2017 Corbin et al. compared the application of various approaches to account for measurement error including: bias analysis (fixed parameter and probabilistic), direct imputation (a mixed method which incorporated priors on the sensitivity and specificity of measurement into imputation models), and Markov-Chain (Monte-Carlo) Bayesian analysis.[38] They advise

investigators to use quantitative bias analyses and Bayesian analyses in any settings where informative priors can be specified and where accounting for all sources of uncertainty is critical. Second, Rudolph & Stuart have adapted two existing methods, propensity score calibration [30, 39, 40] and VanderWeele & Arah's bias formulas [41, 42], which were originally developed to address unmeasured confounding, to now address imperfectly measured covariates.[43]

Extended Look-back Windows to Assess Medical History

Currently, it is common practice in database studies to assess medical histories within uniform or fixed look-back windows (e.g. 1-year). However, fixed look-backs require cohorts to be restricted to those meeting some definition of data continuity (e.g. continuous enrollment in claims studies) for the entire window, and potentially informative data occurring before the window are discarded.[44, 45] Observing all historical (pre-exposure) information available in a database while requiring only minimal baseline continuity has been proposed as an alternate approach which might improve capture of relevant medical history and selection of more inclusive, representative cohorts.[46, 47] Simulation studies indicated all-available look-backs may be superior to fixed look-backs in some settings.[47, 48] However, concerns remain that the method may be prone to bias if the completeness and longitudinal breadth of available data might vary informatively between exposure (e.g. when comparing users to non-users) or outcome groups.

Only two papers have been published exploring use of all-available look-backs in actual data with multiple interrelated covariates; however, both studies focused on claims, not EMR data.[49, 50] In both studies, control for confounding was not substantially affected by the look-back used to assess confounders. However, the second study indicates that eligibility criteria (e.g. history of exposures, outcomes) may be better assessed using all-available data or a long (3-year) fixed look-back, as opposed to a short (1-year) fixed look-back. [50] Further research is needed which explores the application of all-available look-back approaches in alternate data sources (e.g. EMR).

Follow-Up Contingent on Encounters

Loss to follow-up (or right censoring) is arguably one of the most common forms of missing data among studies using longitudinal data sources, especially when subjects are only observed periodically as they are in EMR. As mentioned above, censoring is particularly problematic in inpatient EMR settings. In 2018, Lewin et al. demonstrated that multiple imputation methods did no better than a complete-case analysis in a setting where the only missing data was non-ignorable right-censoring of time-to-event outcomes.[51] However, Lesko et al. provide new guidance to reduce bias affecting studies where events that occur after the last time a person was observed result in right-censoring and go unrecognized.[52] Based on both simulation and an applied example, they conclude that: 1) studies where events can only be recognized in an observed encounter should censor patients lost to follow-up at the time of their last encounter; 2) studies where events can be observed outside encounters (e.g. all-cause mortality obtained from a national death index) should censor when the patient meets the definition of loss to follow-up. They conclude that studies

conducted using EMR will be strongly affected by choice of censoring method, positing that bias is greatest when rates of the outcome and loss to follow-up are high.

Restricting to Patients with High Data Continuity

In response to the frequently encountered challenge in EMR research where some people have a high degree of data missingness for care delivered by health systems/providers not captured in the EMR, researchers have proposed methods of identifying cohorts with high data-completeness and medical record continuity.[2-4] These methods apply primarily to research using ambulatory EMRs, where continuity of coverage may be observed in patients returning for regular medical interactions. These methods require administrative claims data linked at the patient-level, allowing investigators to directly assess whether events observed in the linked claims are also observed in the medical record. For example, Lin et al. propose the statistic mean proportion of encounters captured (MPEC) which is equal to the average of two proportions: 1) proportion of outpatient visits recorded in claims that are also noted in EMR and 2) the proportion of inpatient admissions in claims that are also noted in EMR.[2] After restricting to patients in the top MPEC quintile, Lin et al. reported that misclassification for 40 different commonly used covariates was reduced by a factor of 3.5 to 5.8.[3]

However, restricting to patients with complete data is potentially problematic, since doing so conditions effect estimates on those variables, potentially impacting 1) the internal validity of effect estimates (by conditioning analyses on observed data), and 2) the external validity (by altering the composition of the study population). Weber et al. demonstrated that such restrictions selected cohorts that were older, sicker, and more likely to be female.[4] As an alternative, they proposed a more flexible approach which only seeks to eliminate people whose data is incomplete for the type of variables needed in a specific study. They propose various heuristic filters (e.g. demographics, data fact types [e.g. diagnoses, vital signs, lab tests, medications, or outpatient visits), and time spans (e.g. data in first and last study month) which may be necessary for some study designs but can be relaxed for others.

Maximum Likelihood Approaches and Inverse Probability Weighting

Inverse probability weighting (IPW) for complete case analyses, which is often classified as a maximum-likelihood approach, is a common alternative to imputing missing data or restricting to those without missing values.[53-56] The simple complete-case analysis (i.e. assessing only subjects with complete data) can be conceptualized as the extreme case of weighting where subjects with any missing data receive a weight of zero. More sophisticated IPW applications, such as the aforementioned approach by Weber et al, weight subjects according to their probability of missing data relevant to the analysis, as indicated by models fit within the observed data. Recent research indicates performance of maximum-likelihood approaches may be superior to multiple imputation when missing data is infrequent and when multiple variables are non-normal. [56]

Sun et al. recently proposed an extension to the inverse probability weighting approach, which is capable of yielding valid inferences in analyses with non-monotone missing data. [57] Their method requires that investigators specify the mechanisms of non-monotone

missingness. They outline procedures for discerning these mechanisms from the data itself, using either maximum likelihood estimation or constrained Bayesian estimation.[57] In settings with longitudinal missing data, Doidge et al. propose incorporating an indicator of previously observed responsiveness (i.e. likelihood of having missing data in prior study encounters) into IPW models predicting data missingness.[55] They assert that the method is likely to reduce bias when data is missing partially at random but caution that it may increase bias when data is missing completely at random.

Imputing Missing Data

In analyses where values can be explicitly identified as missing or misclassified, imputation can be used to assign corrected values based on conditional distributions assessed in the observed data. Multiple imputation, one of the most widely used methods to address missing data, refers to the practice of generating multiple hypothetical data sets containing various imputed values and then analyzing pooled results across them in order to appropriately incorporate the increase in variance due to imputing values.[58] Comprehensive reviews of the wide range of multiple imputation methods and their applications in pharmacoepidemiology are available.[59] Here, we summarize selected recent advancement in multiple imputation methods which are applicable to pharmacoepidemiologic studies. Given that imputation methods are often agnostic to temporal and analytical relationships between variables, most of these methods can be applied to impute variables of many different types (e.g. exposures, covariates, outcomes) and scales (e.g. categorical, continuous). When necessary, we will highlight when a method was intended to be applied in narrower setting.

General Multiple Imputation Developments—Currently, no clear guidance is available for investigators seeking to determine how much missing data is too much for imputations to be reliable. Such determinations are a complex function of the proportion of observations with missing data, the number of observations with non-missing data, the number of variables with missing data, and the covariance between the missing and observed values.[58, 60-63] For example, 95% missingness might not be problematic if imputation assumptions are satisfied and the 5% with complete data is comprised of a sufficient number of observations to inform the imputation.

In highly dimensional data, modeling the full joint distribution of a large covariate set may be infeasible. Multiple imputation by chained equations (MICE) can efficiently address this problem by imputing values for missing variables sequentially in different orders in each of the imputed data sets.[64, 65] Kunkel and Kaizar recently compared imputation approaches with full joint models against MICE and conclude that in scenarios with multivariate normal missing data, MICE models are easier to implement and often produce results similar to the fully specified joint model. However, the authors caution that the choice of prior distributions strongly affect results and advise testing them in sensitivity analyses.[66] Kline et al. also compared the two approaches for imputing longitudinal data at the person-level and found that MICE was only comparable to full joint models when the covariance structure of the missing variable was homogenous and correlations were exchangeable.[67] In high-dimensional data settings, covariates may be balanced using a

summary score (e.g. propensity score, disease risk score). However, methodologists have debated whether investigators imputing propensity scores should 1) average the propensity scores themselves across the multiple imputed datasets then estimate a single effect, or 2) estimate effects in each of the multiple imputed datasets then average across them to produce a single effect.[68-70] In a recent paper, Leyrat et al. provide new guidance, recommending that investigators pool effect estimates, not propensity scores produced within the various imputed datasets, so long as imputation models include the outcome.[70]

Zahid et al. propose another approach to enable imputation in settings with a large number of missing covariates: multiple imputation with sequential penalized regression. The method is an extension of MICE which allows each imputed variable to take on a different distributional form using models specified using various ridge penalties.[71] The authors demonstrate via simulation that this method can be applied to both normal and non-normal response models, and performs well even in scenarios with large number of missing covariates and few observations. An R package (mispr) is also provided.

Another method that relaxes parametric assumptions is predictive mean matching (PMM).[72] Under PMM, imputed values are drawn at random from a matched set of people with complete data and similar expected values generated by the model. Many assert that PMM leads to more realistic distributions than standard imputation approaches. Recent research indicates that predictive mean matching is particularly useful when imputing values for continuous variables with non-normal distributions and when plausible bounds can be placed on missing variables.[73, 74] However, further research comparing PMM to alternate approaches is needed.

Other miscellaneous developments in the imputation literature merit brief mention. In 2017, Sullivan et al. demonstrated that using standard imputation approaches based on logistic regression may produce biased / attenuated estimates in studies with missing data on binary outcomes.[75] It is plausible to expect that in some datasets, performance of imputation may depend on the value of the variable being imputed. For example, imputation may be relatively accurate when imputing patient incomes in the low to medium range but inaccurate when imputing extremely high incomes. Bak et al. introduce a machine-learning approach to multiple imputation which estimates an expected error for each imputed value.[76] This method allows investigators to selectively impute values that fall below some minimum threshold for error.

Imputing Longitudinal or Time-Varying Data—In settings with high-dimensional longitudinal data, imputation models can quickly become complex, difficult to accurately specify, and frequently fail to converge. There are methods available which adapt the MICE procedure to impute data in time-to-event studies using Cox proportional hazards model, for example multiple imputation for joint modeling (MIJM). Moreno-Betancur et al. recently described the method in detail and distributed statistical programming tools for implementation in R.[77] In some instances, methods and programming tools have been specifically tailored for the imputation of specific clinical constructs, for instance longitudinal measures of body-mass-index).[78]

An important consideration when imputing longitudinal data is leveraging information from between-person variation vs. within-person or longitudinal variation. Gottfredson et al. used multilevel multiple imputation (MMI), to impute values of missing data as a function of models fit between people and longitudinal models fit within-person.[79] Similar to their IPW method incorporating predictors of non-responsiveness in models of missing data, Doidge et al. also propose a corresponding imputation method which incorporates non-responsiveness observed in earlier data as a predictor.[55] This approach can be conceptualized as a simplified MMI model.

Forward Bridging—As the medical system continues to advance over time, data generation system and electronic health records dynamically evolve to meet the needs of healthcare providers and patients. Migrating data from older systems to newer systems is necessary to enable longitudinal analyses of healthcare data and poses a common challenge to investigators working with EMR. Thompson et al. introduce a forward bridging method using multiple imputation with multinomial logistic regression.[80]

Latent Variables and External Calibration Data—While missing data is often discussed in the context of vertical linkage, it is becoming increasingly common for researchers to use horizontal linkage, harmonizing data across different study populations to enable larger scale studies with improved generalizability and greater potential for detailed subgroup analyses (see Table 1). When two populations have a variable (e.g. functional status) measured in a similar but not directly comparable way, the combined data are subject to measurement error. As a solution Gu and Gutman propose latent variable matching, a novel method which draws upon both multiple imputation and item response theory. Latent variable matching imputes values for a third, hypothetically unmeasured variable representing the underlying truth that each of the differently measured variables indicate. Their method is an adaptation of predictive mean matching imputation and is appropriate for non-longitudinal data that is missing partially at random. [81] Using simulation, they demonstrate the method's ability to provide valid inference with smaller bias than other methods.[82] An alternate solution has been proposed by Siddique et al., using as an example a meta-analysis of two studies with different outcome definitions.[83] Their method uses external calibration data, or a population where both outcome definitions were measured, to provide information on the relationship between the two measures. The authors propose a multivariate random-effect model that leverages the external calibration data, and jointly models the missing outcome measures, allowing estimation of the effect of time and treatment on the outcome.[83]

Confidence Intervals—Bootstrap estimation and multiple imputation are increasingly common in causal inference research. Shomaker et al. introduce methods to construct confidence intervals in scenarios where bootstrap estimation techniques are used in conjunction with multiple imputation, providing recommendations for calculating valid confidence intervals consistent with randomization.[84] Van Walraven compared the plausible estimate ranges produced by bootstrap imputation to those produced by quantitative bias analyses. They found that while more computationally demanding, bootstrap imputation more effectively decreased misclassification bias compared to

quantitative bias analysis, the latter of which is highly dependent on accurate parameters of bias estimates.[85]

Doubly-Robust Estimation Methods

Conceptually, conventional multiple imputation is chiefly concerned with modeling the values of missing data, while IPW methods are concerned with modeling the probability of missing data. Doubly-robust methods for analyzing data with missing or misclassified values fit both models, implementing imputation and weighting approaches in parallel. These methods are described as doubly-robust since they are more robust to misspecification of the models and their link functions, requiring that only one of the two be appropriately specified. In health services research, investigators tend to be more confident in their ability to correctly specify models predicting missing data than models predicting the values of missing data, particularly in high-dimensional analyses with many interrelated variables.[55]

One such doubly-robust method, augmented inverse probability of treatment weighting (AIPW), requires that the analyst separately fits two parametric working models.[58, 86] The model for the probability of having missing data is used to create inverse probability weights while the model for the values of missing data is incorporated into effect estimation models as an augmentation term. Zhou et al. and Hsu & Yu both proposed doubly-robust extensions of the PMM method, with the first model predicting the missing variable of interest and the second model predicting the probability of missingness.[87, 88] As stated earlier, incorporating PMM has the advantage of relaxing parametric assumptions in the imputation phase. Hsu & Yu show through simulation that this approach is more robust to misspecification of either model compared to other common models.[87] Zhou et al. discuss these methods specifically in the context of missing data in categorical variables with more than two categories, using a combination of multinomial logistic regression and binary logistic regression.[88] Addressing the setting of propensity score analyses, Shu et al. proposed a doubly-robust estimator (allows for misspecification in either the treatment or the outcome model) to address the bias due to measurement error in a binary outcome when a gold standard validation subset is available.

Inferring treatment duration using the reverse waiting time distribution

Defining prescription duration in pharmacoepidemiologic studies can be difficult, and researchers typically make simplifying assumptions or pre-specified decisions based on prescription and patient characteristics.[11, 12, 89] In 2013, Pottegard adapted the waiting time distribution (WTD) method, used to estimate prevalence of drug exposures in databases lacking data on prescription days supply, dosage, and refills, to also infer duration of drug exposures.[90, 91] In companion papers published in 2017, Støvring et al propose an adapted method, the reverse WTD, which models the distribution of time from last prescription to the end of a pre-specified time window, as a function of patient and prescription characteristics.[92, 93] Regardless of whether detailed prescription data is available, the method outputs plausible estimates of prescription duration, customized to the patient and their prescription, providing a scalable, data-driven alternative to pre-specifying decision rules.[92] Hallas illustrates the use of this method, and found that use of the reverse

WTD may reduce misclassification of exposure, while being more statistically efficient than alternative methods.[94]

Misclassification in instrumental variable (IV) analyses

Another recently developed method, published in 2017 by Ertefaie, adjusts IV analyses for confounders of the treatment effect that are associated with the IV (termed IV-confounders), which have non-independent or non-ignorable missing data.[95] The procedure has two critical steps: 1) the instrumental variable value (e.g. provider preference) is estimated using a model including all IV-confounders, among subjects with complete data; and 2) among all subjects, estimate treatment effect using model fit with only IV-confounders that have no missing values. The authors assert that the method is only valid when three assumptions are met for each unmeasured confounder: 1) provider-level missingness cannot be related to unmeasured confounders (although person-level missingness can); 2) the effects of any unmeasured confounders on treatment allocation must be the same for all physicians; and 3) positivity (i.e. each physician sees patients with different values of that variable).

Conclusions

In this paper, we provide an overview of recent advancements in the published literature describing and addressing measurement error in EMR studies. Presently, sources of measurement error impacting EMR studies are still being elucidated, as are the methods to address them. Here, we emphasized methods which seek to repair or reduce the impact of measurement error in the analysis phase. However, investigators may find that repairing the data is more complex than drawing on alternative data sources in which the critical events are more accurately assessed or an alternative study design that eliminates the need for the imperfect data elements (e.g instrumental variable analyses or self-controlled observational study designs). [96, 97] Regardless of the approach, investigators conducting studies that primarily rely on EMR would be well advised to thoroughly consider how measurement error might affect their data as well as their study findings.

Acknowledgments

Dr. Jonsson-Funk, Dr. Conover and Ms. Young report grant support from NIH National Institute on Aging, grants from NIH National Heart Lung and Blood Institute, grants from NIH National Center for Advancing Translational Sciences.

References

1. Wang LE, Shaw PA, Mathelier HM, Kimmel SE, French B. Evaluating risk-prediction models using data from electronic health records. Ann Appl Stat. 2016;10(1):286–304. doi:10.1214/15-aoas891 [PubMed: 27158296]

2••. Lin KJ, Glynn RJ, Singer DE, Murphy SN, Lii J, Schneeweiss S. Out-of-system Care and Recording of Patient Characteristics Critical for Comparative Effectiveness Research. Epidemiology. 2018;29(3):356–63. doi:10.1097/ede.0000000000000794 [PubMed: 29283893] The authors use EMR data from two medical care networks linked with Medicare insurance claims to develop and assess data capture in EMR for 40 research-relevant variables. They report reporting surprisingly low capture proportions (16-27%), and propose a method to restrict EMR studies to patients with sufficiently informative data continuity.

3. Lin KJ, Singer DE, Glynn RJ, Murphy SN, Lii J, Schneeweiss S. Identifying Patients With High Data Completeness to Improve Validity of Comparative Effectiveness Research in Electronic Health Records Data. Clin Pharmacol Ther. 2018;103(5):899–905. doi:10.1002/cpt.861 [PubMed: 28865143]

4••. Weber GM, Adams WG, Bernstam EV, Bickel JP, Fox KP, Marsolo K, et al. Biases introduced by filtering electronic health records for patients with "complete data". J Am Med Inform Assoc. 2017;24(6):1134–41. doi:10.1093/jamia/ocx071 [PubMed: 29016972] Using EMR data from 7 (PCORNet) hospitals and health systems and (un-linked) Aetna insurance claims, the authors assess the impact of applying combinations of 16 different “complete-data” filters within EMR and claims populations. The authors demonstrate how missing data restrictions can be tailored to study-specific needs allowing for optimization of trade-offs between bias and generalizability.

5. Wells BJ, Chagin KM, Nowacki AS, Kattan MW. Strategies for handling missing data in electronic health record derived data. EGEMS (Wash DC). 2013;1(3):1035. doi: 10.13063/2327-9214.1035 [PubMed: 25848578]

6. Mooney SJ. Invited Commentary: The Tao of Clinical Cohort Analysis-When the Transitions That Can Be Spoken of Are Not the True Transitions. Am J Epidemiol. 2017; 185(8):636–8. doi: 10.1093/aje/kww236 [PubMed: 28338912]

7. Fischer MA, Stedman MR, Lii J, Vogeli C, Shrank WH, Brookhart MA, et al. Primary medication non-adherence: analysis of 195,930 electronic prescriptions. Journal of general internal medicine. 2010;25(4):284–90. [PubMed: 20131023]

8. Li X, Cole SR, Westreich D, Brookhart MA. Primary non-adherence and the new-user design. Pharmacoepidemiol Drug Saf. 2018;27(4):361–4. doi: 10.1002/pds.4403 [PubMed: 29460385]

9. Hampp C, Greene P, Pinheiro SP. Use of Prescription Drug Samples in the USA: A Descriptive Study with Considerations for Pharmacoepidemiology. Drug Saf. 2016;39(3):261–70. doi: 10.1007/s40264-015-0382-9 [PubMed: 26798052]

10. Bijlsma MJ, Janssen F, Hak E. Estimating time-varying drug adherence using electronic records: extending the proportion of days covered (PDC) method. Pharmacoepidemiol Drug Saf. 2016;25(3):325–32. doi:10.1002/pds.3935 [PubMed: 26687394]

11••. Pye SR, Sheppard T, Joseph RM, Lunt M, Girard N, Haas JS, et al. Assumptions made when preparing drug exposure data for analysis have an impact on results: An unreported step in pharmacoepidemiology studies. Pharmacoepidemiol Drug Saf. 2018. doi: 10.1002/pds.4440 Intended to clarify complex decision-making when defining drug treatment episodes in longitudinal data, the authors lay out a detailed algorithm/framework comprised of 10 decision nodes and 54 possible assumptions. They explore how variation in different decisions can impact effect estimates in an applied analysis conducted within UK CPRD data.

12. Pazzagli L, Linder M, Zhang M, Vago E, Stang P, Myers D, et al. Methods for time-varying exposure related problems in pharmacoepidemiology: An overview. Pharmacoepidemiol Drug Saf. 2018;27(2): 148–60. doi: 10.1002/pds.4372 [PubMed: 29285840]

13. Devarakonda MV, Mehta N, Tsou CH, Liang JJ, Nowacki AS, Jelovsek JE. Automated problem list generation and physicians perspective from a pilot study. Int J Med Inform. 2017;105:121–9. doi:10.1016/j.ijmedinf.2017.05.015 [PubMed: 28750905]

14. Zhang R, Pakhomov SVS, Arsoniadis EG, Lee JT, Wang Y, Melton GB. Detecting clinically relevant new information in clinical notes across specialties and settings. BMC Med Inform Decis Mak. 2017;17(12):68. doi:10.1186/s12911-017-0464-y [PubMed: 28699564]

15. Hubbard RA, Johnson E, Chubak J, Wernli KJ, Kamineni A, Bogart A, et al. Accounting for misclassification in electronic health records-derived exposures using generalized linear finite mixture models. Health Serv Outcomes Res Methodol. 2017;17(2):101–12. doi: 10.1007/s10742-016-0149-5 [PubMed: 28943779]

16. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. J Biomed Inform. 2017;73:14–29. doi: 10.1016/j.jbi.2017.07.012 [PubMed: 28729030]

17. McTaggart S, Nangle C, Caldwell J, Alvarez-Madrazo S, Colhoun H, Bennie M. Use of text-mining methods to improve efficiency in the calculation of drug exposure to support pharmacoepidemiology studies. Int J Epidemiol. 2018;47(2):617–24. doi: 10.1093/ije/dyx264 [PubMed: 29420741]

18. Munkhdalai T, Liu F, Yu H. Clinical Relation Extraction Toward Drug Safety Surveillance Using Electronic Health Record Narratives: Classical Learning Versus Deep Learning. JMIR Public Health Surveill. 2018;4(2):e29. doi:10.2196/publichealth.9361 [PubMed: 29695376]

19. Hamon T, Grabar N. Linguistic approach for identification of medication names and related information in clinical narratives. J Am Med Inform Assoc. 2010;17(5):549–54. doi: 10.1136/jamia.2010.004036 [PubMed: 20819862]

20•. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol. 2015;68(3):279–89. doi: 10.1016/j.jclinepi.2014.06.018 [PubMed: 25179855] The authors propose a three-step framework for validating prediction models / algorithms which characterizes model performance in context of differences between the validation and development populations. Transportability is indicated by strong performance, maintained across heterogeneous validation and development populations. They walk through the framework in an applied example.

21. Lanes S, Brown JS, Haynes K, Pollack MF, Walker AM. Identifying health outcomes in healthcare databases. Pharmacoepidemiol Drug Saf. 2015;24(10): 1009–16. doi: 10.1002/pds.3856 [PubMed: 26282185]

22. Lesko CR, Jacobson LP, Althoff KN, Abraham AG, Gange SJ, Moore RD, et al. Collaborative, pooled and harmonized study designs for epidemiologic research: challenges and opportunities. Int J Epidemiol. 2018;47(2):654–68. doi:10.1093/ije/dyx283 [PubMed: 29438495]

23. Lin KJ, Garcia Rodriguez LA, Hernandez-Diaz S. Systematic review of peptic ulcer disease incidence rates: do studies without validation provide reliable estimates? Pharmacoepidemiol Drug Saf. 2011;20(7):718–28. doi:10.1002/pds.2153 [PubMed: 21626606]

24. Koller KR, Wilson AS, Asay ED, Metzger JS, Neal DE. Agreement Between Self-Report and Medical Record Prevalence of 16 Chronic Conditions in the Alaska EARTH Study. J Prim Care Community Health. 2014;5(3):160–5. doi:10.1177/2150131913517902 [PubMed: 24399443]

25. Nakamura Y, Sugawara T, Kawanohara H, Ohkusa Y, Kamei M, Oishi K. Evaluation of estimated number of influenza patients from national sentinel surveillance using the national database of electronic medical claims. Jpn J Infect Dis. 2015;68(1):27–9.doi: 10.7883/yoken.JJID.2014.092 [PubMed: 25420664]

26. Stewart AL, Lynch KJ. Identifying discrepancies in electronic medical records through pharmacist medication reconciliation. J Am Pharm Assoc (2003). 2012;52(1):59–66. doi:10.1331/JAPhA.2012.10123

27. Wright A, McCoy AB, Hickman TT, Hilaire DS, Borbolla D, Bowes WA 3rd, et al. Problem list completeness in electronic health records: A multi-site study and assessment of success factors. Int J Med Inform. 2015;84(10):784–90. doi:10.1016/j.ijmedinf.2015.06.011 [PubMed: 26228650]

28. Rothman KJ, Greenland S, Lash TL. Clinical Epidemiolgy. In: Seigafuse S, Bierig L, editors. Modern Epidemiology. 3rd ed. Philadelphia (PA): Lippincott Williams & Wilkins; 2008. p. 643.

29. Gini R, Schuemie MJ, Mazzaglia G, Lapi F, Francesconi P, Pasqua A, et al. Automatic identification of type 2 diabetes, hypertension, ischaemic heart disease, heart failure and their levels of severity from Italian General Practitioners' electronic medical records: a validation study. BMJ Open. 2016;6(12):e012413. doi:10.1136/bmjopen-2016-012413

30. Funk MJ, Landi SN. Misclassification in administrative claims data: quantifying the impact on treatment effect estimates. Curr Epidemiol Rep. 2014;1(4):175–85. doi:10.1007/s40471-014-0027-z [PubMed: 26085977]

31. Rowan CG, Flory J, Gerhard T, Cuddeback JK, Stempniewicz N, Lewis JD, et al. Agreement and validity of electronic health record prescribing data relative to pharmacy claims data: A validation study from a US electronic health record database. Pharmacoepidemiol Drug Saf. 2017;26(8):963–72. doi:10.1002/pds.4234 [PubMed: 28608510]

32. Flory JH, Roy J, Gagne JJ, Haynes K, Herrinton L, Lu C, et al. Missing laboratory results data in electronic health databases: implications for monitoring diabetes risk. J Comp Eff Res. 2017;6(1):25–32. doi:10.2217/cer-2016-0033 [PubMed: 27935320]

33. Patorno E, Gopalakrishnan C, Franklin JM, Brodovicz KG, Masso-Gonzalez E, Bartels DB, et al. Claims-based studies of oral glucose-lowering medications can achieve balance in critical clinical

parameters only observed in electronic health records. Diabetes Obes Metab. 2017. doi:10.1111/dom.13184

34. Heintzman J, Bailey SR, Hoopes MJ, Le T, Gold R, O'Malley JP, et al. Agreement of Medicaid claims and electronic health records for assessing preventive care quality among adults. J Am Med Inform Assoc. 2014;21(4):720–4. doi:10.1136/amiajnl-2013-002333 [PubMed: 24508767]
35. Devoe JE, Gold R, McIntire P, Puro J, Chauvie S, Gallia CA. Electronic health records vs Medicaid claims: completeness of diabetes preventive care data in community health centers. Ann Fam Med. 2011;9(4):351–8. doi:10.1370/afm.1279 [PubMed: 21747107]
36. Yang S, Hutcheon JA. Identifying outliers and implausible values in growth trajectory data. Ann Epidemiol. 2016;26(1):77–80.e1-2. doi:10.1016/j.annepidem.2015.10.002 [PubMed: 26590476]
37. Shi J, Korsiak J, Roth DE. New approach for the identification of implausible values and outliers in longitudinal childhood anthropometric data. Ann Epidemiol. 2018;28(3):204–11.e3. doi:10.1016/j.annepidem.2018.01.007 [PubMed: 29398298]
38. Corbin M, Haslett S, Pearce N, Maule M, Greenland S. A comparison of sensitivity-specificity imputation, direct imputation and fully Bayesian analysis to adjust for exposure misclassification when validation data are unavailable. Int J Epidemiol. 2017;46(3):1063–72. doi:10.1093/ije/dyx027 [PubMed: 28338966]
39. Sturmer T, Schneeweiss S, Rothman KJ, Avorn J, Glynn RJ. Performance of propensity score calibration--a simulation study. Am J Epidemiol. 2007;165(10):1110–8. doi:10.1093/aje/kwm074 [PubMed: 17395595]
40. Sturmer T, Schneeweiss S, Rothman KJ, Avorn J, Glynn RJ. Propensity Score Calibration and its Alternatives. Am J Epidemiol. 2007;165(10):1122–3. [PubMed: 24518589]
41. Arah OA. Bias Analysis for Uncontrolled Confounding in the Health Sciences. Annu Rev Public Health. 2017;38:23–38. doi:10.1146/annurev-publhealth-032315-021644 [PubMed: 28125388]
42. Vanderweele TJ, Arah OA. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. Epidemiology.2011;22(1):42–52. doi:10.1097/EDE.0b013e3181f74493 [PubMed: 21052008]
43•. Rudolph KE, Stuart EA. Using Sensitivity Analyses for Unobserved Confounding to Address Covariate Measurement Error in Propensity Score Methods. Am J Epidemiol. 2018; 187(3):604–13. doi: 10.1093/aje/kwx248 [PubMed: 28992211] The authors propose adaptations of two prominent methods for assessing the impact of unobserved confounders (propensity score calibration, VanderWeele and Arab's bias formulas) to instead assess the impact of measurement error They illustrate the methods in an applied example.
44. Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, et al. Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. Journal of the American Medical Informatics Association : JAMIA. 2007; 14(1): 1–9. doi:M2273 [pii] [PubMed: 17077452]
45. Velentgas P, Dreyer NA, Nourjah P, Smith SR, Torchia MM. Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide. 2013. Contract No.: Report.
46. Rothman KJ, Greenland S, Lash TL. Case-Control Studies. In: Seigafuse S, Bierig L, editors. Modern Epidemiology. 3rd ed. Philadelphia (PA): Lippincott Williams & Wilkins; 2008. p. 121–7.
47. Brunelli SM, Gagne JJ, Huybrechts KF, Wang SV, Patrick AR, Rothman KJ, et al. Estimation using all available covariate information versus a fixed look-back window for dichotomous covariates. Pharmacoepidemiology and drug safety. 2013;22(5):542–50. doi: 10.1002/pds.3434 [PubMed: 23526818]
48. Conover MM, Jonsson Funk M, editors. Uniform vs. all-available look-backs to identify exclusion criteria in observational cohort studies [abstract]2015.
49. Nakasian SS, Rassen JA, Franklin JM. Effects of expanding the look-back period to all available data in the assessment of covariates. Pharmacoepidemiol Drug Saf. 2017;26(8):890–9. doi: 10.1002/pds.4210 [PubMed: 28397352]
50. Conover MM, Sturmer T, Poole C, Glynn RJ, Simpson RJ Jr., Pate V, et al. Classifying medical histories in US Medicare beneficiaries using fixed vs all-available look-back approaches. Pharmacoepidemiol Drug Saf. 2018. doi: 10.1002/pds.4435

51. Lewin A, Brondeel R, Benmarhnia T, Thomas F, Chaix B. Attrition Bias Related to Missing Outcome Data: A Longitudinal Simulation Study. Epidemiology. 2018;29(1):87–95. doi: 10.1097/ede.0000000000000755 [PubMed: 28926372]

52•. Lesko CR, Edwards JK, Cole SR, Moore RD, Lau B. When to Censor? American Journal of Epidemiology. 2018;187(3):623–32. doi:10.1093/aje/kwx281 [PubMed: 29020256] Informative loss to follow-up is an extremely common form of measurement error affecting time-to-event EMR studies. The authors provide needed guidance on how to appropriately right-censor follow-up time for outcomes that can be identified only during observed encounters vs. outside of observed encounters.

53. Little RJ, Rubin DB. Statistical analysis with missing data: John Wiley & Sons; 2014.

54. Wooldridge JM. Inverse probability weighted estimation for general missing data problems. Journal of Econometrics. 2007;141(2):1281–301.

55. Doidge JC. Responsiveness-informed multiple imputation and inverse probability-weighting in cohort studies with missing data that are non-monotone or not missing at random. Stat Methods Med Res. 2018;27(2):352–63. doi: 10.1177/0962280216628902 [PubMed: 26984909]

56. Shin T, Davison ML, Long JD. Maximum likelihood versus multiple imputation for missing data in small longitudinal samples with nonnormality. Psychol Methods. 2017;22(3):426–49. doi: 10.1037/met0000094 [PubMed: 27709974]

57. Sun B, Perkins NJ, Cole SR, Harel O, Mitchell EM, Schisterman EF, et al. Inverse-Probability-Weighted Estimation for Monotone and Nonmonotone Missing Data. American Journal of Epidemiology. 2018;187(3):585–91. doi:10.1093/aje/kwx350 [PubMed: 29165557]

58. Rubin DB. Multiple imputation for nonresponse in surveys: John Wiley & Sons; 2004.

59. Harel O, Mitchell EM, Perkins NJ, Cole SR, Tchetgen Tchetgen EJ, Sun B, et al. Multiple Imputation for Incomplete Data in Epidemiologic Studies. American Journal of Epidemiology. 2018;187(3):576–84. doi:10.1093/aje/kwx349 [PubMed: 29165547]

60. Schafer JL. Analysis of incomplete multivariate data: Chapman and Hall/CRC; 1997.

61. Schafer JL. Multiple imputation: a primer. Stat Methods Med Res. 1999;8(1):3–15. doi: 10.1177/096228029900800102 [PubMed: 10347857]

62. Wei R, Wang J, Su M, Jia E, Chen S, Chen T, et al. Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. Sci Rep. 2018;8(1):663. doi: 10.1038/s41598-017-19120-0 [PubMed: 29330539]

63. Dong Y, Peng CY. Principled missing data methods for researchers. Springerplus. 2013;2(1):222. doi: 10.1186/2193-1801-2-222 [PubMed: 23853744]

64. Rubin DB. Statistical matching using file concatenation with adjusted weights and multiple imputations. Journal of Business & Economic Statistics. 1986;4(1):87–94.

65. Rawlings AM, Sang Y, Sharrett AR, Coresh J, Griswold M, Kucharska-Newton AM, et al. Multiple imputation of cognitive performance as a repeatedly measured outcome. Eur J Epidemiol. 2017;32(1):55–66. doi: 10.1007/s10654-016-0197-8 [PubMed: 27619926]

66. Kunkel D, Kaizar EE. A comparison of existing methods for multiple imputation in individual participant data meta-analysis. Stat Med. 2017;36(22):3507–32. doi: 10.1002/sim.7388 [PubMed: 28695667]

67. Kline D, Andridge R, Kaizar E. Comparing multiple imputation methods for systematically missing subject-level data. Res Synth Methods. 2017;8(2): 136–48. doi: 10.1002/jrsm.1192 [PubMed: 26679326]

68. Hill J Reducing bias in treatment effect estimation in observational studies suffering from missing data. 2004.

69. Mitra R, Reiter JP. A comparison of two methods of estimating propensity scores after multiple imputation. Stat Methods Med Res. 2016;25(1):188–204 doi: 10.1177/0962280212445945 [PubMed: 22687877]

70. Leyrat C, Seaman SR, White IR, Douglas I, Smeeth L, Kim J, et al. Propensity score analysis with partially observed covariates: How should multiple imputation be used? Stat Methods Med Res. 2017:962280217713032. doi: 10.1177/0962280217713032

71. Zahid FM, Heumann C. Multiple imputation with sequential penalized regression. Stat Methods Med Res. 2018:962280218755574. doi: 10.1177/0962280218755574

72. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. Bmj. 2009;338:b2393. [PubMed: 19564179]

73. Lee KJ, Carlin JB. Multiple imputation in the presence of non-normal data. Stat Med. 2017;36(4):606–17. doi:10.1002/sim.7173 [PubMed: 27862164]

74. Geraci M, McLain A. Multiple Imputation for Bounded Variables. Psychometrika. 2018. doi: 10.1007/s11336-018-9616-y

75. Sullivan TR, Lee KJ, Ryan P, Salter AB. Multiple imputation for handling missing outcome data when estimating the relative risk. BMC Med Res Methodol. 2017;17(1): 134. doi: 10.1186/s12874-017-0414-5 [PubMed: 28877666]

76•. Bak N, Hansen LK. Data Driven Estimation of Imputation Error-A Strategy for Imputation with a Reject Option. PLoS One. 2016;11(10):e0164464. doi: 10.1371/journal.pone.0164464 [PubMed: 27723782] The authors describe a novel imputation method that selectively imputes values when they fall below a maximum error threshold. The method assesses imputation error among those with complete data, then assigns the error value to a person with missing data, who is non-parametrically matched using machine-learning.

77. Moreno-Betancur M, Carlin JB, Brilleman SL, Tanamas SK, Peeters A, Wolfe R. Survival analysis with time-dependent covariates subject to missing data or measurement error: Multiple Imputation for Joint Modeling (MIJM). Biostatistics. 2017.doi: 10.1093/biostatistics/kxx046

78. Kontopantelis E, Parisi R, Springate DA, Reeves D. Longitudinal multiple imputation approaches for body mass index or other variables with very low individual-level variability: the mibmi command in Stata. BMC Res Notes. 2017;10(1):41. doi: 10.1186/s13104-016-2365-z [PubMed: 28086961]

79. Gottfredson NC, Sterba SK, Jackson KM. Explicating the Conditions Under Which Multilevel Multiple Imputation Mitigates Bias Resulting from Random Coefficient-Dependent Missing Longitudinal Data. Prev Sci. 2017;18(1):12–9. doi: 10.1007/s11121-016-0735-3 [PubMed: 27866307]

80. Thompson CA, Boothroyd DB, Hastings KG, Cullen MR, Palaniappan LP, Rehkopf DH. A Multiple-Imputation "Forward Bridging" Approach to Address Changes in the Classification of Asian Race/Ethnicity on the US Death Certificate. Am J Epidemiol. 2018; 187(2):347–57. doi: 10.1093/aje/kwx215 [PubMed: 29401361]

81. Little RJ. Missing-data adjustments in large surveys. Journal of Business & Economic Statistics. 1988;6(3):287–96.

82. Gu C, Gutman R. Combining item response theory with multiple imputation to equate health assessment questionnaires. Biometrics. 2017;73(3):990–8. doi: 10.1111/biom.12638 [PubMed: 27936287]

83. Siddique J, Reiter JP, Brincks A, Gibbons RD, Crespi CM, Brown CH. Multiple imputation for harmonizing longitudinal non-commensurate measures in individual participant data meta-analysis. Stat Med. 2015;34(26):3399–414. doi: 10.1002/sim.6562 [PubMed: 26095855]

84. Schomaker M, Heumann C. Bootstrap inference when using multiple imputation. Stat Med. 2018. doi: 10.1002/sim.7654

85. van Walraven C Improved Correction of Misclassification Bias With Bootstrap Imputation. Med Care. 2017. doi:10.1097/mlr.0000000000000787

86. Wang C, Chen HY. Augmented inverse probability weighted estimator for Cox missing covariate regression. Biometrics. 2001;57(2):414–9. [PubMed: 11414564]

87•. Hsu CH, Yu M. Cox regression analysis with missing covariates via nonparametric multiple imputation. Stat Methods Med Res. 2018:962280218772592. doi: 10.1177/0962280218772592 The authors develop a novel method for addressing missing data for multiple covariates in time-to-event analysis that combines two existing methods: augmented inverse probability of treatment weighting (AIPW) and predictive mean matching imputation. The method is doubly-robust to model misspecification and is non-parametric, so suitable for non-normally distributed data.

88. Zhou M, He Y, Yu M, Hsu CH. A nonparametric multiple imputation approach for missing categorical data. BMC Med Res Methodol. 2017;17(1):87. doi:10.1186/s12874-017-0360-2 [PubMed: 28587662]

89. Gardarsdottir H, Souverein PC, Egberts TC, Heerdink ER. Construction of drug treatment episodes from drug-dispensing histories is influenced by the gap length. J Clin Epidemiol. 2010;63(4):422–7. doi:10.1016/j.jclinepi.2009.07.001 [PubMed: 19880282]

90. Hallas J, Gaist D, Bjerrum L. The waiting time distribution as a graphical approach to epidemiologic measures of drug utilization. Epidemiology. 1997;8(6):666–70. [PubMed: 9345667]

91. Pottegard A, Hallas J. Assigning exposure duration to single prescriptions by use of the waiting time distribution. Pharmacoepidemiol Drug Saf. 2013;22(8):803–9. doi: 10.1002/pds.3459 [PubMed: 23703741]

92••. Støvring H, Pottegård A, Hallas J. Refining estimates of prescription durations by using observed covariates in pharmacoepidemiological databases: an application of the reverse waiting time distribution. Pharmacoepidemiology and Drug Safety. 2017;26(8):900–8. doi:doi: 10.1002/pds.4216 [PubMed: 28466973] The authors develop and apply a novel method, adapted from the reverse-waiting time distribution method, to estimate prescription durations in longitudinal data, modeled as a function of patient characteristics. Their data-driven method is more scalable and may be more accurate than the existing practice of specifying decision rules.

93. Støvring H, Pottegård A, Hallas J. Estimating medication stopping fraction and real-time prevalence of drug use in pharmaco-epidemiologic databases. An application of the reverse waiting time distribution. Pharmacoepidemiology and Drug Safety. 2017;26(8):909–16. doi:doi: 10.1002/pds.4217 [PubMed: 28474439]

94. Hallas J, Pottegard A, Stovring H. Using probability of drug use as independent variable in a register-based pharmacoepidemiological cause-effect study-An application of the reverse waiting time distribution. Pharmacoepidemiol Drug Saf. 2017;26(12): 1520–6. doi: 10.1002/pds.4326 [PubMed: 29024218]

95. Ertefaie A, Flory JH, Hennessy S, Small DS. Instrumental Variable Methods for Continuous Outcomes That Accommodate Nonignorable Missing Baseline Values. American Journal of Epidemiology. 2017;185(12):1233–9. doi:10.1093/aje/kww137 [PubMed: 28338946]

96. Ertefaie A, Small DS, Flory JH, Hennessy S. A tutorial on the use of instrumental variables in pharmacoepidemiology. Pharmacoepidemiol Drug Saf. 2017;26(4):357–67. doi: 10.1002/pds.4158 [PubMed: 28239929]

97. Gault N, Castaneda-Sanabria J, De Rycke Y, Guillo S, Foulon S, Tubach F. Self-controlled designs in pharmacoepidemiology involving electronic healthcare databases: a systematic review. BMC Med Res Methodol. 2017;17(1):25. doi: 10.1186/s12874-016-0278-0 [PubMed: 28178924]

Table 1.

Sources of measurement error and missing data relevant to pharmacoepidemiology research using electronic medical records (EMR).

Source of measurement error	Research setting and approach	Key findings and proposed solutions / methods	Citation
EMR data reflect only the health services and medications delivered within the specific health care setting that contributes to the EMR system.	• Setting: U Pennsylvania Health System EMR, linked data from state mandated readmissions reporting system • Approach: Used state registry to assess unidentified 30-day readmissions among hospital patients with a primary diagnosis of heart failure in EMR data.	• 29% of 30-day readmissions went to another hospital • Informative loss to follow up may substantially bias effects estimated within EMR • Biases are not easily addressed using standard methods	[1]
	• Setting: EMR data from two medical care networks linked with Medicare claims data • Approach: Quantify continuity of EMR capture for 40 indicator variables relevant to comparative effectiveness research compared to claims (using standardized differences).	• Mean capture proportion in a single electronic health record system was 16%–27%. • Reasonable variable classification achieved when ≥60% of claims encounters also appear in EMR system.	[2]
	• Setting: EMR data from two medical care networks linked with Medicare claims data • Approach: Estimated mean proportion of encounters captured (MPEC) as the average of two proportions: 1) proportion of outpatient visits recorded in claims that are also noted in EMR, 2) proportion of inpatient admissions in claims that are also noted in EMR.	• Misclassification of 40 indicator variables was 3.5 to 5.8 times lower among patients in the top 20% of MPEC compared to the rest of the study population • Recording of diagnoses made during inpatient visits is complete and accurate than outpatient visits but right censoring / loss to follow-up is common	[3]
	• Setting: EMR from 7 (PCORNet) hospitals and health systems; Nationwide Aetna claims database (no patient-level linkage) • Approach: Assessed the impact of applying different combinations of 16 complete data filters within EMR and claims populations	• EMR completeness varies widely between sites • Applying complete data filters selects sicker, older patients more likely to be female and may exclude patients with complete data • Complete data filters should be applied to only data elements that are critical for analysis.	[4]
	• Approach: Literature review / commentary	• Missing/misclassified data is associated disease severity and healthcare utilization • Whenever possible, imputation within EMR should 1) include predictors for overall health status and utilization and 2) be conducted using linked (non-EMR) data.	[5]
		• Patient behaviors during periods that are well-captured by EMR likely differ from those during periods not well-captured by EMR (when few health encounters are happening).	[6]

Source of measurement error	Research setting and approach	Key findings and proposed solutions / methods	Citation
		• Left and right-censored data before and after periods of intense medicalization (e.g. hospitalization) are common. • EMR-capture may be a function of patients' self-perceived risk	
Prescription records in an ambulatory EMR typically reflect clinician orders for medications, which may not be filled or consumed by the patient.	• Setting: Community-practice based eRx Collaborative in Massachusetts (Blue Cross Blue Shield of Massachusetts, Tufts Health Plan and Zix Corporation). Linked three data sources: e-prescribing transactions, pharmacy claims files, and provider characteristic files. • Approach: Assessed proportion of medications ordered that generated pharmacy/fill claims as an indicator of treatment adherence.	• 78% of all prescriptions and 72% of new prescriptions generated claims for prescription fills • Medication class was the strongest predictor of adherence • Non-adherence common for newly prescribed medications treating chronic conditions: hypertension (28.4%), hyperlipidemia (28.2%), and diabetes (31.4%)	[7]
	• Approach: Literature review / commentary	• Investigators must explicitly decide which is of analytical interest (prescribing, medication disbursement, or consumption) then determine whether variable measured in their data is a misclassified form of the true variable of interest	[8]
	• Setting: National physician survey (Encuity Research Treatment Answers) with encounter-level prescribing data including: 1) whether drug samples, prescriptions, or both were provided and 2) whether they were new or ongoing treatment. • Approach: Quantified exposure time accounted for by (typically unobservable) free prescription drug samples for various drug classes.	• The average days supply accounted for by samples ranged from 13.4 days (dabigatran) to 25.3 (exenatide) among chronic condition medications. • Identified meaningful differences in sample use between frequently used active comparators (e.g. sitagliptin: 44.7%, metformin: 3.6 %) • Measurement error due to sample use affects pharmacy/fill data more than provider-sourced data (e.g. EMR).	[9]
In EMR studies, defining treatment episodes / treatment duration / cumulative exposure is complex and requires many decisions which have unpredictable influence on exposure misclassification.	• Setting: 100 randomly sampled patients from a nationally representative pharmacy dispensing database (IADB.nl) in the Netherlands • Approach: Over the course of one year, estimated for a cohort of 100 patients initiating statin therapy, the difference between a time-fixed definition of proportion of days covered (PDC) and a time-varying definition of PDC.	• Assessing PDC as a static or time-fixed construct in EMR may lead to bias since patients with irregular consumption patterns may be classified with patients who have steady drug consumption. • EMR studies assessing adherence (as either a causal factor or outcome) should use a time-varying definition of PDC	[10]
	• Setting: Clinical Practice Research Datalink (CPRD), a national UK primary care EMR database. • Approach: Developed various algorithms for processing assembling treatment episodes from raw prescription data in the Clinical Practice	• Many of the different approaches to defining treatment episodes produced similar estimates of the hazard ratio and standard errors. • Findings appear sensitive to the definition of episode end date, particularly when constructed using data elements that are frequently missing or unobserved – HRs range from 1.77 (1.56, 2.00) to 2.83 (1.59, 5.04).	[11]

Curr Epidemiol Rep. Author manuscript; available in PMC 2022 May 27.

Source of measurement error	Research setting and approach	Key findings and proposed solutions / methods	Citation
	Research Datalink (CPRD) and tested them in an applied study of cardiovascular risk.		
	• Approach: Literature review / commentary	• Different methods of constructing treatment episodes can lead to substantial differences in the estimates of the associations or effects of interest.	[12]
Automating data entry in EMR systems may forward-propagate erroneous data and/or carry forward information that is no longer clinically relevant.	• Setting: 15 patients selected from 15 conveniently sampled internal medicine attendings at Cleveland Clinic, each of which had medical records data with ≥ 3 encounters and ≥ clinical notes. • Approach: Used three methods to assemble summary "problem lists" describing the general clinical status of 15 patients: 1) existing EMR list, 2) Watson-system natural language processing, 3) manual physician-curated list. Physicians then assessed and scored the three approaches.	• Manual physician-curated lists received the highest scores, followed by Watson, with the existing (automatic) EMR lists being scored lowest. • Natural language processing and cognitive computing may be used to characterize patients in the EMR as a function of data previously observed in the EMR, potentially resulting in forward propagation of error.	[13]
	• Setting: 40 patients with multiple comorbidities in EMR from U of Minnesota affiliated Fairview Health Services hospitals and clinics clinical data repository; manual chart review. • Approach: Used natural language processing algorithms to assess clinical information in patient charts and assessed whether semantic similarity measures reduced the amount of redundant information.	• Mean redundancy was similar in inpatient (68.3%) vs. outpatient (60.7%) settings • Mean redundancy varied by medical specialty (75%, 66%, 57%, and 55% observed in pediatric, internal medicine, psychiatry and surgical notes, respectively)	[14]
	• Setting: EMR data from Group Health, an integrated health system in Washington State; linked to Puget Sound Surveillance Epidemiology and End Results (SEER) registry • Approach: Used simulation to compare the bias and efficiency of various methods to classifying tests/labs/exams as either diagnostic vs. screening in a comparative study of the two on cancer outcomes.	• Differentiating between tests/labs/exams meant for screening vs. surveillance, vs. diagnostic purposes is challenging in EMR data • Various approaches to classifying tests/labs/exams as screening vs. diagnostic have important impacts on effect estimates.	[15]
Recent advances in natural language processing (NLP), which automate extraction of information from unstructured data, may introduce systematic errors.	• Approach: Literature review / commentary	• Natural language processing is powerful but there is no one-size-fits-all solutions capable of solving all of the nuanced problems encountered when applying these tools in clinical sub-domains.	[16]

Source of measurement error	Research setting and approach	Key findings and proposed solutions / methods	Citation
		• When searching for natural language processing tools, searching online code repositories is preferred to searching published literature, which is often incomplete or outdated.	
	• Setting: Free text prescriptions in the NHS Scotland Prescribing Information System; manual chart review • Approach: Designed and tested a natural language processing algorithm to codify medication dosing instructions stored in unstructured text notes	• Key challenges identified: misspellings, acronyms and abbreviations, structural ambiguity • Natural language processing algorithms may need to be built within therapeutic-specific areas.	[17]
	• Setting: 791 EHR text notes from randomly sampled patients with hematological malignancy and cancer drug exposures; manual chart review • Approach: Use supervised machine-learning approaches to natural language processing to identify data entities (e.g. medication, dose, indication, severity) and relations between them (e.g. medication-dose, medication-indication).	• Classical learning models (SVM) remains advantageous over deep learning models for relation identification • Relation identification in unstructured text notes is challenging, regardless of the method selected.	[18]
	• Setting: Hospital discharge summaries from Partners Healthcare; manual chart review • Approach: Designed and tested a natural language processing algorithm to extract medication and medication-related information from unstructured clinical narratives provided by Partners Healthcare.	• Performance of natural-language processing algorithms is highly contextual, varying depending on: – The clinical purpose of the text notes being processed (e.g. a discharge vs. admission summary) – The primary language of the text's author (e.g. English vs. English as a second language).	[19]
Performance of EMR-based clinical prediction algorithms may vary widely between different health systems.	• Setting: Individual-level participant data from four distinct datasets with varying case-mix • Approach: Tested an algorithm to predict deep vein thrombosis assessed differences between the testing population and the original development population including: case mix, predictive accuracy of the algorithm.	• Health systems will often have differences in case-mix, with different underlying patient populations leading to differences in algorithm validity across systems • Blind application of algorithms developed in one EMR environment to EMR from different health systems may lead to unexpected measurement error. • Researchers should quantify and/or understand differences between populations algorithms are developed/tested in vs. those they are applied in.	[20]
Temporal changes in EMR data elements recorded may produce systematic differences in classification and/or missingness over time.	• Approach: Literature review / commentary	• Consult with relevant data and clinical experts to consider the performance/transportability of existing algorithms and/or data elements across study: 1) populations, 2) time periods, 3) databases. • Simple exclusion of patients with missing data may cause bias if missingness is non-random.	[21]

Source of measurement error	Research setting and approach	Key findings and proposed solutions / methods	Citation
		• If necessary and feasible, validate algorithms/definitions within sub-groups, and use quantitative bias analyses to assess the impact of missing data • Note: A “forward-bridging” imputation method for temporal data changes is described in the “Addressing Measurement Error” section. (Thompson, 2018)	
Horizontal linkage of populations captured by different EMR systems produce systematic differences in classification and/or missingness between the linked populations.	• Approach: Literature review / commentary	• Methodological advantages to horizontal data linkage include increased power for sub-group analyses, studying rare events, and detecting heterogeneity • Horizontal linkage produces systematic biases including misclassification, missing data, and residual confounding • Unclear whether horizontal linkage should be conducted before or after imputation.	[22]