CONTINUING EDUCATION

CE 1.5 ANCC CONTACT HOURS

Challenges Frequently Encountered in the Secondary Use of Electronic Medical Record Data for Research

Meghan E. Edmondson, BSN, RN, Andrew P. Reimer, PhD, RN

The wide adoption of electronic medical records and subsequent availability of large amounts of clinical data provide a rich resource for researchers. However, the secondary use of clinical data for research purposes is not without limitations. In accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines, we conducted a systematic review to identify current issues related to secondary use of electronic medical record data via MEDLINE and CINAHL databases. All articles published until June 2018 were included. Sixty articles remained after title and abstract review, and four domains of potential limitations were identified: (1) data quality issues, present in 91.7% of the articles reviewed; (2) data preprocessing challenges (53.3%); (3) privacy concerns (18.3%); and (4) potential for limited generalizability (21.7%). Researchers must be aware of the limitations inherent to the use of electronic medical record data for research and consider the potential effects of these limitations throughout the entire study process, from initial conceptualization to the identification of adequate sources that can provide data appropriate for answering the research questions, analysis, and reporting study results. Consideration should also be given to using existing data quality assessment frameworks to facilitate use of standardized data quality definitions and further efforts of standard data quality reporting in publications.

KEY WORDS: Data preprocessing, Data quality, EMR data, Secondary use

Research using electronic medical record (EMR) data is a relatively new method of inquiry when compared to other types of research. Electronic medical record data impart several benefits, such as low cost, large volume of data available, and saved time because there is no need to recruit and retain participants,[1,2] although considered revisions to the US Common Rule and the European General Data Protection Requirements may affect future consent requirements.[3,4] However, there are valid concerns regarding the use of EMR data for research and the potential limitations it entails. Electronic medical record data, or any data that were not originally collected for the purpose of research, carry a risk of poor data quality, identified by van der Lei[5] as "the first law of informatics," which states that data should be used only for its originally intended purpose. More recent literature has countered that idea and instead embraced the concept of "fitness for purpose" or "fitness for use."[2,6–8] This approach assumes that a data set is appropriate depending on intended use; for researchers, it depends on the research question.[2,6,8] Understanding the limitations of EMR data and developing technologies and methodologies that mitigate these inherent limitations are a burgeoning aspect of informatics research. Therefore, the purpose of this article is to conduct a comprehensive and updated systematic review of the literature to identify and describe the known challenges of using EMR data for secondary research purposes and to provide a beginner's guide that summarizes the many aspects one should consider.

Author Affiliation: Frances Payne Bolton School of Nursing, Case Western Reserve University, Cleveland, Ohio.

The authors have disclosed that they have no significant relationships with, or financial interest in, any commercial companies pertaining to this article.

Corresponding author: Meghan E. Edmondson, BSN, RN, Frances Payne Bolton School of Nursing, Case Western Reserve University, 10900 Euclid Ave, Cleveland, OH 44106 (mee58@case.edu).

DOI: 10.1097/CIN.0000000000000609

METHODS

Search Strategy

In accordance with Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, an exhaustive literature search was performed using MEDLINE and CINAHL databases for articles published up to the date on which the search was conducted, June 26, 2018.[9,10] The search was not limited by date, because use of EMR data for research purposes is a relatively new and foundational concept discussed in older articles could have been missed if the search had been limited by date. The topic of secondary use of EMR data has only become relevant since the mid-1990s as primary use of EMRs for clinical documentation increased in popularity. MEDLINE and CINAHL databases were searched using the following terms: "EMR data + research + limitations," "EMR data + research + challenges," "EMR data + quality + research," "EMR data + research + pitfalls," "using EMR data for research," "EMR data + suitability for research," and "EMR data + data quality + research."

Inclusion Criteria

Studies published in English that discussed the limitations of EMR data in the context of secondary use for research on a conceptual level were included in the analysis.

Exclusion Criteria

Articles were excluded if their discussion of EMR use was limited to primary clinical use, such as rates of or barriers to EMR implementation for use in clinical documentation. Studies addressing quality improvement or development of clinical decision support applications or other purely clinical applications were also excluded. Articles related to development of data remediation methodologies aimed specifically at circumventing barriers, such as natural language processing applications or other preprocessing software, rather than identification and description of the barriers that necessitate such applications, were considered beyond the scope of this article and thus were also excluded.

Study Selection

Duplicate articles were removed, and titles were reviewed by one author using the previously discussed inclusion and exclusion criteria. The abstracts of those articles were further reviewed. After abstract review, full-text articles were reviewed. An ancestor search was then conducted on the references of the included articles, following the same title review, abstract review, and full-text review process.

RESULTS

A total of 3462 articles were returned in the search (Figure 1). Duplicates were removed, reducing the total to 1590 articles. After title review, 197 articles remained. Abstract review reduced the total number of articles to 81, and 32 of those articles were included in the analysis after full-text review. The ancestor search produced another 28 articles that were included in the review, bringing the total number of articles to 60. Of these 60 articles, 22 were author manuscripts, seven were reviews, two were case studies, and two used qualitative methods including workshops and stakeholder interviews. There were 16 articles reporting original research findings, one of which was a prospective randomized controlled trial.

Review of the articles yielded four domains of limitations on the secondary use of EMR data for research: privacy concerns, data extraction and transformation challenges, problems with data quality, and the potential for limited generalizability. Problems with data quality were further divided into subdomains of completeness, correctness, and currency. We adopted the terminology proposed by Weiskopf et al[2,6,11] due to their thorough definitions and widespread use, and to enable consistent use of existing terms and continued development of standardized frameworks for data quality assessment. An overwhelming majority of studies were found to include discussion related to more than one domain. Data quality was discussed in 91.7% of the articles included in the review. Issues encountered during data preprocessing, such as issues with data extraction and transformation, were explored in 53.3% of articles reviewed, and privacy and generalizability were mentioned in 18.3% and 21.7% of articles, respectively.

Privacy

Privacy was noted to be a challenge in EMR-based research in 18.3% of articles. The secondary use of clinical data for research does not require consent in many cases, but it does require data to be collected and managed in a way that does not include identifiable personal health information to minimize the risk of reidentification.[12] There are two ways to achieve this. First is using limited data sets, which can retain dates and ages, geographic locations, and unique patient identifiers that cannot be used to reidentify the patient.[13] A second option is developing a deidentified data set, from which all identifying information are removed, including dates, geographic locations, and unique patient identifiers.[13] However, even when using exempt data sets, or deidentified data sets, institutional review is still required to determine that the data and proposed study meet the criteria.[13] The institutional review board will determine if, and at what level, consent is required from participants.[13] The type of data set a researcher chooses largely depends on what information is necessary to answer the research question of interest. Whenever possible, research should be conducted using the minimum amount of necessary personal information.

Another strategy to preserve privacy when using EMR data is to use aggregated or pooled data.[12,14,15] When individual-level data are not necessary to answer a particular research question, data can be combined to form population-level data, an approach that is often used in genomic studies.[13] Population data are used when patients are grouped by provider, hospital or unit, or some other grouping factor in which individual patient data are not needed. These methods ensure compliance with requirements to protect patients' privacy.[12,13]

Data Preprocessing: Extraction and Transformation

Extraction of EMR data falls under the umbrella of data preprocessing, which also includes cleaning, transforming data into a statistically interpretable format, and loading the transformed data into applications that enable statistical analysis. Extraction is the process by which data are located within an EMR. Preprocessing is noted to be a "laborious" process consuming a majority of the project time[12,13,16–19] that requires merging data from multiple sources, in multiple different formats in a database, into a single format that

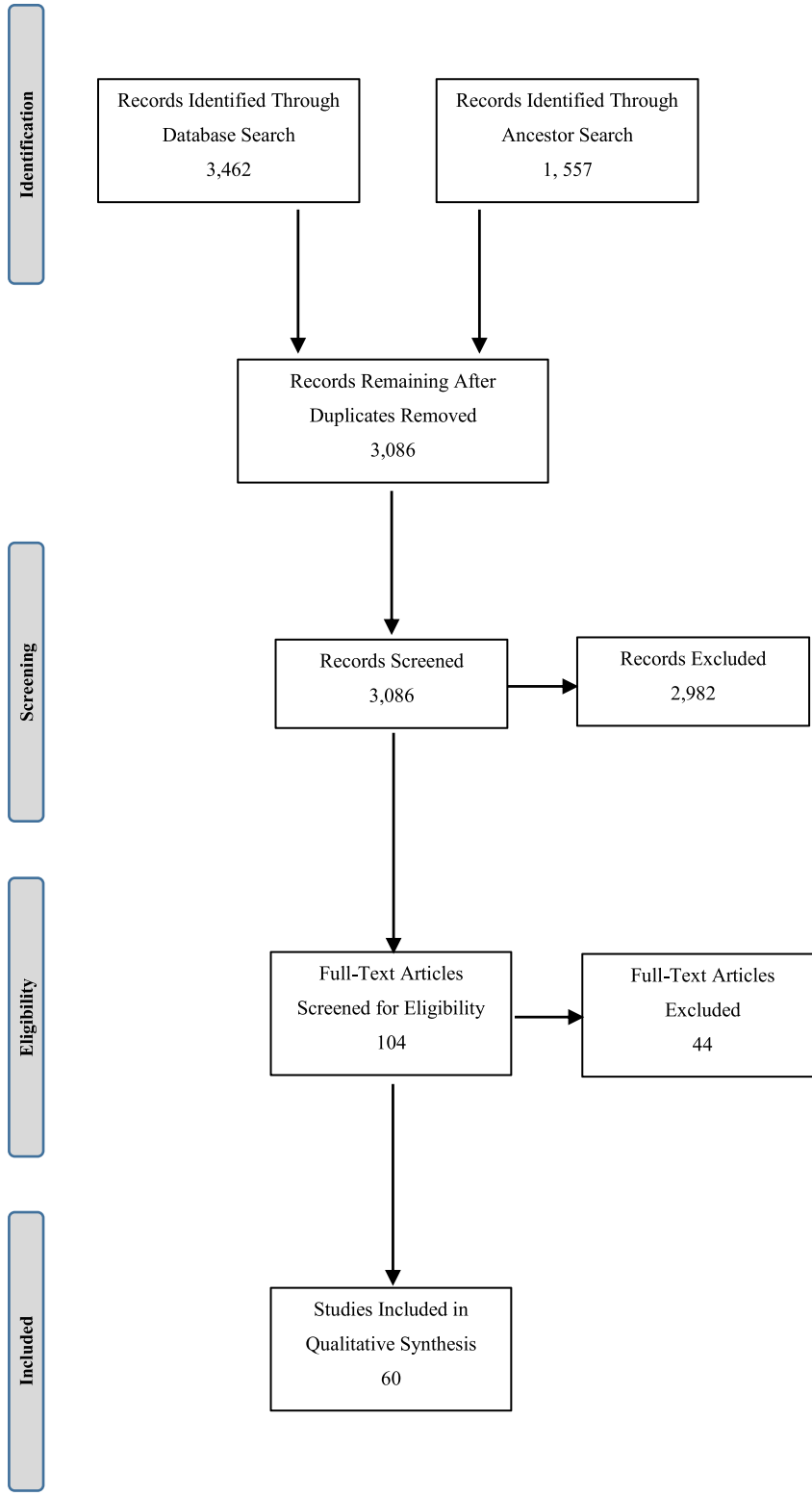

FIGURE 1. Flowchart summarizing article inclusion and exclusion filtering, as outlined by the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) 2009 guidelines.[9,10]

is analyzable with statistical software. Relevant research information can be documented in an EMR in a variety of formats, such as drop-down menus, free-text narrative notes, radiographic images, and laboratory values.[1,6,12,15–17,19–33] Each of these may be documented within a different software system. All relevant data must be merged into a single format,[6,12,13,16–18,20,24,26,34,35] which is often complicated because the various software systems are usually not interoperable.[6,12,14,17,27,36] Interoperability, as defined by the Healthcare Information and Management Systems Society, is "the ability of different information technology systems and software applications to communicate, exchange data, and use the information that has been exchanged."[37] Semantic interoperability, the most complex level, requires that systems be able to exchange data and interpret the data for further use.[37,38] In the context of data extraction and transformation for EMR-based research, better interoperability can improve preprocessing, so that more data can be merged and interpreted by receiving systems more efficiently, thereby decreasing the time required to extract and transform data. Here, we encounter the barrier that the need to maintain anonymity places on data preprocessing. In order for multiple formats of data from multiple sources to be merged, one needs a way to identify separate pieces of data belonging to the same patient.[12,16,17,34] One approach to transforming these data is to develop randomly generated unique identifiers that are substituted for patient identifiers like the medical record. Another approach, hashing, uses a mathematical formula to code data such that it cannot be decrypted to reveal protected health information, but can still be used to link data from different sources. Hashing is emerging as a method of ensuring that privacy is protected during data extraction.[12,13] Tamersoy et al[39] provide an example of this technique.

Another challenge in data extraction is proper identification of eligible records.[18,20,22,23,29,32,40] This task is undertaken based on a priori inclusion criteria, such as the diagnosis of a particular illness, age within a certain range, or a record generated within a certain time period, and is also known as phenotyping.[12,41] Due to the volume of data available in EMRs, in order for a record that meets the inclusion criteria to be correctly identified as eligible, the required elements must be searchable.[13,17] Billing codes, such as the 9th and 10th revisions of the *International Classification of Diseases*, have been widely used for this purpose, but inaccurate and missing billing codes result in missed records not identified by search criteria or records included when a diagnosis was not actually present.[17,20,22,29,36,42] Because of the difficulty of accurately identifying all patients who meet the specified inclusion criteria and excluding patients who do not, caution is required if the goal is to establish incidence and/or prevalence, due to the difficulty of establishing a denominator.[43]

A third idea prevalent in the literature with regard to data extraction is the lack of standardization in terminology used for diagnosis or clinical findings.[13,15,17,18,20–22,24,27,33–36,44–47] Free-text narrative notes present a particular challenge to this aspect of EMR-based research.[1,6,12,15–32] While they facilitate documentation and generation of a record that is considered complete from a clinical standpoint, there is wide variation in the human use of language encountered in narrative notes; there are many ways for different providers to identify the same diagnosis, clinical problem, or response to treatment with different terminology. This lack of standardization in narrative notes means that many eligible records may be missed if the exact terminology used in the record is not searched.[22,36] Circumventing the challenge posed by human use of language in narrative notes has become a focus of healthcare informatics research. Recent advances in the development of natural language processing applications have demonstrated reliable techniques to identify representative samples.[48–51]

Data Quality

The concept of data quality is complex and context-dependent.[2,6,7,22,30,52–54] The absence of data[2,6,7,11,13–17,19–23,25–29,31–33,36,40,42,47,53,55–67] and presence of inaccurate data create concern when using EMR data for research.[6–8,13,14,18–22,25,28,31–34,36,40,42,47,52,53,56–58,60–62,64,65,67,68] Many authors support the need to evaluate data quality prior to attempting to answer a research question when using EMR data.[2,6,7,13,22,28,30,53,60,65] There is a lack of standardized methods of data quality assessment in EMR-based research,[6,23,24,28,53,54,60,65] as well as a lack of a standardized reporting guidelines on data quality assessment.[23,24,28,53,54,60] Development of a standardized method of assessing data quality is an issue currently at the forefront of medical informatics research. While there is not a universally accepted data quality assessment framework, Kahn and colleagues'[60] harmonized data quality assessment terminology and framework are the most developed. The lack of a widely used method of data quality assessment and reporting in EMR-based research inhibits comparison among studies and valuation of the meaning of EMR-based study findings.[11,54,60,65] Although reviewing methods of data quality assessment is beyond the scope of this article, we have provided references to approaches in Table 1.

There is inconsistent use of terminology in the field of data quality assessment.[6,33] In addition to the terms "completeness" and "accuracy" previously discussed, some authors refer to "missing data" or "incomplete data"[16,17,20–23,26,27,29,36,55–58] and "inaccuracies."[20–22,26,33,36,57] Heinze et al[25] refer to "fragmentary information" and "imprecise measurements." "Discordance" and "(in)consistency" have been used to describe the concept of concordance.[13,41,57,58,61] Few authors define these terms, so it is possible similar terms are being used

Table 1. Challenge Domains Identified in Each Article

Author	Data Quality	Preprocessing	Generalizability	Privacy
Afzali et al[55]	✓			
Arterburn et al[56]	✓		✓	
Bagley and Altman[43]	✓		✓	
Baier et al[21]	✓		✓	
Botsis et al[57]	✓	✓		
Brookhart et al[70]	✓		✓	
Coleman et al[22]	✓	✓		
Coorevits et al[14]	✓	✓		✓
Damotte et al[69]	✓	✓		
Davis and Haines[23]	✓	✓		
de Lusignan et al[44]	✓	✓		
de Lusignan et al[45]	✓	✓		
de Lusignan et al[66]	✓	✓		
de Lusignan and van Weel[15]	✓	✓		✓
Dean et al[24]	✓	✓		
Embi and Payne[76]		✓		✓
Faulconer and de Lusignan[58]	✓			
Haneuse and Daniels[59]	✓		✓	
Heinze et al[25]	✓	✓	✓	
Hersh et al[26]	✓	✓	✓	
Hogan and Wagner[42]	✓			
Holve et al[34]	✓	✓	✓	
Johnson et al[77]	✓			
Johnson et al[67]	✓			
Kahn et al[60]	✓			
Kanas et al[17]	✓	✓	✓	✓
Kheterpal[13]	✓	✓		✓
Lau et al[61]	✓			
Leo et al[68]	✓			
Lin et al[27]	✓	✓	✓	
Lobach and Detmer[35]		✓		✓
Logan et al[52]	✓			
Majeed et al[28]	✓	✓		
Murphy et al[12]		✓		✓
Newton et al[41]	✓	✓		
Puttkammer et al[53]	✓			
Reimer et al[62]	✓			
Rosenthal[46]	✓	✓		
Rusanov et al[63]	✓		✓	
Safran et al[1]		✓		✓
Schwartz et al[29]	✓	✓		
Stewart et al[30]	✓	✓		
Sutherland et al[16]	✓	✓		
Tamersoy et al[39]				✓
Terry et al[20]	✓	✓		
Thiru et al[54]	✓			
van Velthoven et al[64]	✓			
Wagner and Hogan[31]	✓			
Wasserman[32]	✓	✓	✓	
Weiner et al[19]	✓	✓		

(continues)

Table 1. Challenge Domains Identified in Each Article, Continued

Author	Data Quality	Preprocessing	Generalizability	Privacy
Weiskopf et al[65]	✓			
Weiskopf et al[11]	✓			
Weiskopf et al[2]	✓		✓	
Weiskopf et al[6]	✓			
Wang and Strong[8]	✓			
Weng et al[40]	✓			✓
Yamamoto et al[18]	✓	✓		
Yim et al[47]	✓			
Young et al[33]	✓	✓		✓
Zampi et al[36]	✓	✓		
Zozus et al[7]	✓			

This table is intended to serve as a quick reference to identify additional articles for a more in-depth discussion on topics of interest.

to discuss different concepts. It is equally possible that conceptually dissimilar terms are being used interchangeably. Therefore, an exploration of the conceptual definitions of frequently used terms in data quality assessment is a necessary exercise. Weiskopf et al[65] provide such an exploration and have identified completeness, correctness, and currency as integral components of data quality assessment, with concordance and plausibility identified as methods of assessing correctness.

Completeness

Hogan and Wagner[42] define completeness as "the proportion of data observed that are actually recorded" Weiskopf et al[11] provide a more comprehensive approach to assessing data completeness. Completeness has four potential definitions, which are context-dependent: (1) if data are expected to be present in a record, they are present; (2) adequate breadth; (3) depth of data over time; and (4) enough data are present in the record to answer the question of interest about a clinical problem.[11,65] Complete data, from a clinical perspective, are data that represent everything that was observed in the clinical encounter; completeness from a research perspective is all data that are necessary to answer the research question at hand.[2] The latter definition introduces the concept of "fitness for purpose" or "fitness for use."[2,6–8] The concept of "fitness for purpose" asserts that whether data quality is sufficient depends on the presence of data required to answer the research question(s).[2,6,8]

An additional prominent theme in the literature regarding data completeness was selection bias resulting from worsening disease severity associated with more complete data.[2,17,19,24,36,43,59,61,63,69] For example, Weiskopf et al[2,63] demonstrate that availability of data that are complete from a research perspective depends on the severity of illness, resulting in potential for confounding effects.[70] Patients with more advanced illness will seek medical treatment and require follow-up visits more often than healthy individuals, and therefore more EMR data will be generated on their behalf. This is an especially important limitation for researchers using EMR data to be aware of, as it means that the study sample may not accurately represent the population of interest.

Completeness of data depends on the context in which the data are to be used,[2,11] how completeness is defined in that particular assessment of data quality,[6] and the influence of provider-related factors on the documentation of a complete record. These factors may include "enthusiasm,"[28] what is perceived as clinically necessary by the clinician,[20] and that severity of illness may produce a sample population that severely limits the generalizability of the study results.[2,36,43,61,63,69] Variability of documentation practices exists among institutions, among clinics within the same institution, and among clinicians.[46] Completeness may also be affected by billing methods.[32,47] Yim et al. describe differences in coding practices between a Veterans Affairs (VA) hospital and a community hospital.[47] The VA hospital had fewer diagnoses coded, which the authors attribute to a capitated billing system resulting in a reduced incentive to code.[47] These authors also note that longitudinal studies using EMR data are especially vulnerable to missing data as time progresses,[47] which threatens the internal validity of any study using this design. Data completeness has also been found to increase with the availability of free-text fields; however, this benefit was noted to come at the expense of the accuracy of decision support systems.[31] Haneuse and Daniels[59] suggest considering why data are present, rather than why data are not present, when assessing data completeness.

Correctness

The concept of correctness was the second most commonly addressed. However, the concept of correctness was more often referred to as "accuracy."[32,33,36,56,68] Hogan and Wagner[42] use the term "correctness," which they define as "the proportion of recorded observations that are

correct," as a measure of accuracy. According to Hogan and Wagner,[42] accuracy of data should be monitored and is highly variable in EMRs.[6,42] Weiskopf et al[65] define correct data as data that are free from error. Weiskopf et al[6,65] describe concordance and plausibility as falling under the scope of correctness. Concordance is the agreement among different elements of data from different sources or the longitudinal agreement of data over time.[6,62] Plausible data, as defined by Weiskopf et al,.[6,65] are data that "make sense" according to clinical knowledge or are feasible given other data that are present in the record. Kahn et al[60] add six additional types of correctness that include plausibility broken down into uniqueness, atemporal, and temporal, and concordance related to value, relational, and computational.

Currency

There is limited literature addressing the concept of currency. Puttkammer et al[53] refer to the concept as "timeliness" of documentation. Currency is defined at how well data represent the patient's state at a particular time.[6,62] Weiskopf et al[65] note that assessment of currency of EMR data can only be done when time-stamped metadata are available and that the determination of what constitutes adequate currency is context-dependent.

Limited Generalizability

That EMR data were not collected for research purposes is frequently expressed in the literature as a major limitation of the use of EMR data for research.[5,6,13,17,19,20,30,32] The fact that clinical data are not routinely collected by the same person in a standardized way undermines the scientific rigor achievable with this type of data.[6,17,19] Additionally, the equipment used to collect clinical data is rarely calibrated before each use or even at regular time points. This lack of rigor in the data collection process has been noted to contribute to a lack of meaning in any results obtained.[30,61] Conversely, the argument has also been made that this real-world method of data collection enhances internal validity.[63]

Electronic medical record data carry a risk of systematic error.[56] According to Sutherland et al,[16] some amount of missing data may not be problematic because of the very large sample size typical of studies using EMR data. However, findings may be less generalizable to other institutions or other patient populations due to the sampling bias inherent in the use of EMR data.[17,29,43] Handling of missing EMR data has many unique considerations, which are beyond the scope of this article. Bounthavong et al[71] and Wells et al[72] provide examples of several ways to handle missing data.

Additionally, the volume of EMR data sets is potentially very large. It is possible to achieve statistical significance simply because of the size of the sample or, in other words, commit a type 1 error. Researchers should be aware of the potential for a type 1 error during analysis and interpretation of results. This effect can be mitigated by setting α at .01 or stricter. Further, using a conceptual model that is based on sound reasoning and prior empirical evidence to form a priori definitions of the concepts and relationships being tested can provide support for the statistical testing results.

DISCUSSION

We completed a systematic review of articles discussing the challenges faced when using EMR data for the secondary purpose of research. Our analysis led to identification of four domains of potential limitations: privacy, data extraction and transformation, data quality, and potential for limited generalizability. Several authors provide reviews that include challenges belonging to these domains.[15,20,24,26] However, even the most recent of these articles were almost a decade old at the time of our search. The field of health informatics, and especially data quality assessment, is rapidly evolving. Thus, a more recent review was necessary. A recent review was published by Yim et al[47]; however, this review primarily focuses on the domain of data quality. Thiru et al[54] provide another review focused on data quality. Issues encountered in data extraction and privacy were reviewed by Coorevits et al.[14] A recent, comprehensive systematic review encompassing multiple domains of challenges of secondary EMR data use was not found in our search.

A noteworthy observation is that an overwhelming majority of this work was conducted in the primary care setting. It remains unknown whether the effect of disease severity on completeness of data is as profound in the acute or intensive care settings as it is in primary care settings. The effect of disease severity could potentially be decreased in studies conducted with acute and intensive care patient populations due to all subjects being hospitalized, because of both stricter inclusion criteria and mandatory documentation requirements in the hospital setting. In the hospital setting, especially in intensive care, there is a wide range of acuity of illness. As in the primary care setting, higher acuity necessitates more frequent intervention and therefore more frequent documentation. The degree to which disease severity affects data quality likely depends on which data elements are required to answer the study questions and with what frequency documentation of those elements is dictated by standards of practice and hospital policies for each data element. This aspect of data quality may also depend on whether the study population consists of patients with a survivable acute illness or deterioration from a chronic condition, as these situations are approached differently in the hospital setting. Patients suffering from chronic diseases at end-of-life stages may not generate as much data simply because there are fewer medical interventions likely to improve their health, and care is typically deescalated. Their vital signs and test results, while

abnormal, may be stable or unamenable to intervention, thereby requiring less frequent monitoring than a patient with a rapidly changing clinical course. This highlights the need to assess data quality in terms of its fitness for use.

Interestingly, the breadth of data increases as disease severity increases, but only up to the point of terminal illness, at which point data completeness declines.[61,69] Goals of care at the terminal illness stage shift from curative treatment to managing symptoms and facilitating comfort, often resulting in fewer procedures, medications, and medical visits. Therefore, it makes sense that fewer records would be generated, and data would be less complete at this stage of illness.

Electronic medical record–based research is retrospective and observational.[8,12,13,55,56,63] For this reason, EMR-based research has been called hypothesis-generating, rather than hypothesis-confirming.[12,13,55] With this in mind, EMR data are still a valuable asset to the research community despite these limitations. Retrospective studies can offer a foundation on which to base prospective or confirmatory research, which does not entail the same sampling bias problem as retrospective EMR-based studies.

This review is limited by the lack of standardized terminology in health informatics research, and relevant articles using different terminology may have been missed. To address this problem, we used multiple search terms in multiple combinations. However, we used only "EMR" and not "EHR" (electronic health record) in our search terms, and this may have limited our results. To assess the potential impact of this limitation, we conducted a search using the same search terms previously described with "EHR" substituted for "EMR." While new articles were found, a review of these articles did not reveal any additional domains. However, an additional concept related to both privacy and extraction was identified: difficulty accessing records for research. Access may be limited by privacy laws or by the proprietary nature of software used to generate electronic records,[73–75] but in itself is not exclusive to the use of data for secondary research purposes as this is a primary limitation for access to medical data in general. Finally, this review was limited to only studies describing the use of EMR data for research purposes, not including work on other data quality assessment domains (eg, data quality checking related to data entry, clinical use, or clinical quality improvement); thus, there is potential for additional data quality concepts related to the use of EMR data in those specific contexts.

CONCLUSION

Privacy concerns, data preprocessing and data quality challenges, and the potential for limited generalizability are known challenges encountered when using EMR data for secondary research purposes. Researchers need to be aware of these challenges during the initial planning stages through to using EMR data for research. While this review provided only a cursory overview of the conceptual issues to consider when using EMR data, future work should include a review focused on clinical studies using EMR to identify common issues encountered and solutions employed that enhance clinical EMR data usefulness. Lastly, continued efforts in the data quality domain should focus on developing streamlined data extraction processes, ongoing development of data quality assessments with standardized applications and reporting practices, and continued use of standardized terms to contribute in developing and adopting a globally recognized standard for data quality assessment and reporting.

References

1. Safran C, Bloomrosen M, Hammond WE, 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*. 2007;14(1): 1–9.
2. Weiskopf NG, Rusanov A, Weng C. Sick patients have more data: the non-random completeness of electronic health records. *AMIA Annual Symposium Proceedings/AMIA Symposium*. 2013;2013: 1472–1477.
3. US Department of Health and Human Services Office of Human Subjects Research Protections. Revised common rule. https://www.hhs.gov/ohrp/regulations-and-policy/regulations/finalized-revisions-common-rule/index.html. Accessed December 3, 2018.
4. The European Parliament and the Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and the Council of the European Union: on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (general data protection regulation). *Official Journal of the European Union*. 2016;L(119): 1.
5. van der Lei J. Use and abuse of computer-stored medical records. *Methods of Information in Medicine*. 1991;30(2): 79–80.
6. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association*. 2013;20(1): 144–151.
7. Zozus MN, Hammond E, Green BB, et al. Assessing data quality for healthcare systems data used in clinical research. *National Institutes of Health: Health Care Systems Research Collaboratory*. National Institutes of Health; 2014. https://www.nihcollaboratory.org/Products/Assessing-data-quality_V1%200.pdf. Accessed December 3, 2018.
8. Wang RY, Strong DM. Beyond accuracy: what data quality means to data consumers. *Journal of Management Information Systems*. 1996;12(4): 5–33.
9. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS Medicine*. 2009;6(7): e1000100.
10. Moher D, Liberati A, Tetzlaff J, Altman DG. The PRISMA group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Medicine*. 2009;6(7): e100097.
11. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. *Journal of Biomedical Informatics*. 2013;46(5): 830–836.
12. Murphy S, Castro V, Mandl K. Grappling with the future use of big data for translational medicine and clinical care. *Yearbook of Medical Informatics*. 2017;26(1): 96–102.
13. Kheterpal S. Clinical research using an information system: the multicenter perioperative outcomes group. *Anesthesiology Clinics*. 2011;29(3): 377–388.

14. Coorevits P, Sundgren M, Klein GO, et al. Electronic health records: new opportunities for clinical research. *Journal of Internal Medicine*. 2013; 274(6): 547–560.
15. de Lusignan S, van Weel C. The use of routinely collected computer data for research in primary care: opportunities and challenges. *Family Practice*. 2006;23(2): 253–263.
16. Sutherland SM, Kaelber DC, Downing NL, Goel VV, Longhurst CA. Electronic health record-enabled research in children using the electronic health record for clinical discovery. *Pediatric Clinics of North America*. 2016;63(2): 251–268.
17. Kanas G, Morimoto L, Mowat F, O'Malley C, Fryzek J, Nordyke R. Use of electronic medical records in oncology outcomes research. *ClinicoEconomics and Outcomes Research: CEOR*. 2010;2: 1–14.
18. Yamamoto K, Sumi E, Yamazaki T, et al. A pragmatic method for electronic medical record-based observational studies: developing an electronic medical records retrieval system for clinical research. *BMJ Open*. 2012;2(6).
19. Weiner MG, Lyman JA, Murphy S, Weiner M. Electronic health records: high-quality electronic data for higher-quality clinical research. *Informatics in Primary Care*. 2007;15(2): 121–127.
20. Terry AL, Chevendra V, Thind A, Stewart M, Marshall JN, Cejic S. Using your electronic medical record for research: a primer for avoiding pitfalls. *Family Practice*. 2010;27(1): 121–126.
21. Baier AW, Snyder DJ, Leahy IC, Patak LS, Brustowicz RM. A shared opportunity for improving electronic medical record data. *Anesthesia and Analgesia*. 2017;125(3): 952–957.
22. Coleman N, Halas G, Peeler W, Casaclang N, Williamson T, Katz A. From patient care to research: a validation study examining the factors contributing to data quality in a primary care electronic medical record database. *BMC Family Practice*. 2015;16: 11.
23. Davis MF, Haines JL. The intelligent use and clinical benefits of electronic medical records in multiple sclerosis. *Expert Review of Clinical Immunology*. 2015;11(2): 205–211.
24. Dean BB, Lam J, Natoli JL, Butler Q, Aguilar D, Nordyke RJ. Review: use of electronic medical records for health outcomes research: a literature review. *Medical Care Research and Review*. 2009;66(6): 611–638.
25. Heinze G, Wallisch C, Kainz A, et al. Chances and challenges of using routine data collections for renal health care research. *Nephrology, Dialysis, Transplantation*. 2015;30(suppl 4): iv68–iv75.
26. Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. *Medical Care*. 2013;51(8 suppl 3): S30–S37.
27. Lin J, Jiao T, Biskupiak JE, McAdam-Marx C. Application of electronic medical record data for health outcomes research: a review of recent literature. *Expert Review of Pharmacoeconomics & Outcomes Research*. 2013;13(2): 191–200.
28. Majeed A, Car J, Sheikh A. Accuracy and completeness of electronic patient records in primary care. *Family Practice*. 2008;25(4): 213–214.
29. Schwartz KL, Tu K, Wing L, et al. Validation of infant immunization billing codes in administrative data. *Human Vaccines & Immunotherapeutics*. 2015;11(7): 1840–1847.
30. Stewart M, Thind A, Terry AL, Chevendra V, Marshall JN. Implementing and maintaining a researchable database from electronic medical records: a perspective from an academic family medicine department. *Healthcare policy*. 2009;5(2): 26–39.
31. Wagner MM, Hogan WR. The accuracy of medication data in an outpatient electronic medical record. *Journal of the American Medical Informatics Association*. 1996;3(3): 234–244.
32. Wasserman RC. Electronic medical records (EMRs), epidemiology, and epistemology: reflections on EMRs and future pediatric clinical research. *Academic Pediatrics*. 2011;11(4): 280–287.
33. Young J, Eley D, Fahey P, Patterson E, Hegney D. Enabling research in general practice– increasing functionality of electronic medical records. *Australian Family Physician*. 2010;39(7): 506–509.
34. Holve E, Segal C, Hamilton Lopez M. Opportunities and challenges for comparative effectiveness research (CER) with electronic clinical data: a perspective from the EDM forum. *Medical Care*. 2012;50(suppl): S11–S18.
35. Lobach DF, Detmer DE. Research challenges for electronic health records. *American Journal of Preventive Medicine*. 2007;32(5 suppl): S104–S111.
36. Zampi JD, Donohue JE, Charpie JR, Yu S, Hanauer DA, Hirsch JC. Retrospective database research in pediatric cardiology and congenital heart surgery: an illustrative example of limitations and possible solutions. *World Journal for Pediatric and Congenital Heart Surgery*. 2012;3(3): 283–287.
37. Healthcare Information and Management Systems Society. HIMSS definition of interoperability. HIMSS Web site. 2013. https://www.himss.org/previous-himss-interoperability-definitions. Accessed November 19, 2018.
38. Institute of Electrical and Electronics Engineers. IEEE standard computer dictionary: a compilation of IEEE standard computer glossaries. *IEEE Standards 610*. 1991, 1-217. http://ieeexplore.ieee.org/servlet/opac?punumber=2267. Accessed November 19, 2018.
39. Tamersoy A, Loukides G, Denny JC, Malin B. Anonymization of administrative billing codes with repeated diagnoses through censoring. *American Medical Informatics Association Annual Symposium Proceedings*. 2010;2010: 782–786.
40. Weng C, Appelbaum P, Hripcsak G, et al. Using EHRs to integrate research with patient care: promises and challenges. *Journal of the American Medical Informatics Association: JAMIA*. 2012;19(5): 684–687.
41. Newton KM, Peissig PL, Kho AN, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. *Journal of the American Medical Informatics Association*. 2013;20(e1): e147–e154.
42. Hogan WR, Wagner MM. Accuracy of data in computer-based patient records. *Journal of the American Medical Informatics Association*. 1997;4(5): 342–355.
43. Bagley SC, Altman RB. Computing disease incidence, prevalence and comorbidity from electronic medical records. *Journal of Biomedical Informatics*. 2016;63: 108–111.
44. de Lusignan S, Metsemakers JF, Houwink P, Gunnarsdottir V, van der Lei J. Routinely collected general practice data: goldmines for research? A report of the European Federation for Medical Informatics Primary Care Informatics Working Group (EFMI PCIWG) from MIE2006, Maastricht, the Netherlands. *Informatics in Primary Care*. 2006;14(3): 203–209.
45. de Lusignan S, Valentin T, Chan T, et al. Problems with primary care data quality: osteoporosis as an exemplar. *Informatics in Primary Care*. 2004; 12(3): 147–156.
46. Rosenthal GE. The role of pragmatic clinical trials in the evolution of learning health systems. *Transactions of the American Clinical and Climatological Association*. 2014;125: 204–216; discussion 217-218.
47. Yim WW, Wheeler AJ, Curtin C, Wagner TH, Hernandez-Boussard T. Secondary use of electronic medical records for clinical research: challenges and opportunities. *Convergent Science Physical Oncology*. 2018;4(1).
48. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *Journal of the American Medical Informatics Association*. 2016;23(5): 1007–1015.
49. Huhdanpaa HT, Tan WK, Rundell SD, et al. Using natural language processing of free-text radiology reports to identify type 1 Modic endplate changes. *Journal of Digital Imaging*. 2018;31(1): 84–90.
50. Sippo DA, Warden GI, Andriole KP, et al. Automated extraction of BI-RADS final assessment categories from radiology reports with natural language processing. *Journal of Digital Imaging*. 2013;26(5): 989–994.
51. Zhou L, Friedman C, Parsons S, Hripcsak G. System architecture for temporal information extraction, representation and reasoning in clinical narrative reports. *AMIA Annual Symposium Proceedings/AMIA Symposium*. 2005; 869–873.
52. Logan JR, Gorman PN, Middleton B. Measuring the quality of medical records: a method for comparing completeness and correctness of clinical encounter data. *AMIA Annual Symposium Proceedings/AMIA Symposium*. 2001;408–412.
53. Puttkammer N, Baseman JG, Devine EB, et al. An assessment of data quality in a multi- site electronic medical record system in Haiti. *International Journal of Medical Informatics*. 2016;86: 104–116.
54. Thiru K, Hassey A, Sullivan F. Systematic review of scope and quality of electronic patient record data in primary care. *British Medical Association*. 2003;326(7398): 1070.

55. Afzali A, Ciorba MA, Schwartz DA, et al. Challenges in using real-world clinical practice records for validation of clinical trial data in inflammatory bowel disease: lessons learned. *Inflammatory Bowel Diseases*. 2017;24(1): 2–4.
56. Arterburn D, Ichikawa L, Ludman EJ, et al. Validity of clinical body weight measures as substitutes for missing data in a randomized trial. *Obesity Research & Clinical Practice*. 2008;2(4): 277–281.
57. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary use of EHR: data quality issues and informatics opportunities. *Summit on Translational Bioinformatics*. 2010;2010: 1–5.
58. Faulconer ER, de Lusignan S. An eight-step method for assessing diagnostic data quality in practice: chronic obstructive pulmonary disease as an exemplar. *Informatics in Primary Care*. 2004;12(4): 243–254.
59. Haneuse S, Daniels M. A general framework for considering selection bias in EHR-based studies: what data are observed and why? *eGEMs: Generating Evidence and Methods to Improve Patient Outcomes (Washington, DC)*. 2016;4(1): 1203.
60. Kahn MG, Callahan TJ, Barnard J, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *eGEMs: Generating Evidence and Methods to Improve Patient Outcomes (Washington, DC)*. 2016;4(1): 1244.
61. Lau EC, Mowat FS, Kelsh MA, et al. Use of electronic medical records (EMR) for oncology outcomes research: assessing the comparability of EMR information to patient registry and health claims data. *Clinical Epidemiology*. 2011;3: 259–272.
62. Reimer AP, Milinovich A, Madigan EA. Data quality assessment framework to assess electronic medical record data for use in research. *International Journal of Medical Informatics*. 2016;90: 40–47.
63. Rusanov A, Weiskopf NG, Wang S, Weng C. Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research. *BMC Medical Informatics and Decision Making*. 2014;14: 51.
64. van Velthoven MH, Mastellos N, Majeed A, O'Donoghue J, Car J. Feasibility of extracting data from electronic medical records for research: an international comparative study. *BMC Medical Informatics and Decision Making*. 2016;16: 90.
65. Weiskopf NG, Bakken S, Hripcsak G, Weng C. A data quality assessment guideline for electronic health record data reuse. *eGEMs: Generating Evidence and Methods to Improve Patient Outcomes (Washington, DC)*. 2017;5(1): 14.
66. de Lusignan S, Hague N, van Vlymen J, Kumarapeli P. Routinely-collected general practice data are complex, but with systematic processing can be used for quality improvement and research. *Informatics in Primary Care*. 2006;14(1): 59–66.
67. Johnson SG, Speedie S, Simon G, Kumar V, Westra BL. Application of an ontology for characterizing data quality for a secondary use of EHR data. *Applied Clinical Informatics*. 2016;7(1): 69–88.
68. Leo MC, Lindberg NM, Vesco KK, Stevens VJ. Validity of medical chart weights and heights for obese pregnant women. *eGEMs: Generating Evidence and Methods to Improve Patient Outcomes (Washington, DC)*. 2014;2(1): 1051.
69. Damotte V, Lizée A, Tremblay M, et al. Harnessing electronic medical records to advance research on multiple sclerosis. *Multiple Sclerosis*. 2019;25(3): 408–418.
70. Brookhart MA, Stürmer T, Glynn RJ, Rassen J, Schneeweiss S. Confounding control in healthcare database research: challenges and potential approaches. *Medical Care*. 2010;48(6 suppl): S114–S120.
71. Bounthavong M, Watanabe JH, Sullivan KM. Approach to addressing missing data for electronic medical records and pharmacy claims data research. *Pharmacotherapy*. 2015;35(4): 380–387.
72. Wells BJ, Chagin KM, Nowacki AS, Kattan MW. Strategies for handling missing data in electronic health record derived data. *eGEMs: Generating Evidence and Methods to Improve Patient Outcomes (Wash DC)*. 2013;1(3): 1035.
73. Adibuzzaman M, DeLaurentis P, Hill J, Benneyworth BD. Big data in healthcare—the promises, challenges and opportunities from a research perspective: a case study with a model database. *AMIA Annual Symposium Proceedings/AMIA Symposium*. 2018;2017: 384–392.
74. Liaw ST, Powell-Davies G, Pearce C, Britt H, McGlynn L, Harris MF. Optimising the use of observational electronic health record data: current issues, evolving opportunities, strategies and scope for collaboration. *Australian Family Physician*. 2016;45(3): 153–156.
75. Nordo AH, Levaux HP, Becnel LB, et al. Use of EHRs data for clinical research: historical progress and current applications. *Learning Health Systems*. 2019;3(1): e10076.
76. Embi PJ, Payne PR. Clinical research informatics: challenges, opportunities and definition for an emerging domain. *Journal of the American Medical Informatics Association*. 2009;16(3): 316–327.
77. Johnson SG, Speedie S, Simon G, Kumar V, Westra BL. A data quality ontology for the secondary use of EHR data. *American Medical Informatics Association Annual Symposium Proceedings*. 2015;2015: 1937–1946.

For more than 202 additional continuing education articles related to research topics, go to NursingCenter.com.

Instructions for Taking the CE Test Online

Challenges Frequently Encountered in the Secondary Use of Electronic Medical Record Data for Research

- Read the article. The test for this CE activity can be taken online at www.NursingCenter.com. Tests can no longer be mailed or faxed.
- You will need to create a free login to your personal CE Planner account before taking online tests. Your planner will keep track of all your Lippincott Professional Development online CE activities for you.
- There is only one correct answer for each question. A passing score for this test is 13 correct answers. If you pass, you can print your certificate of earned contact hours and the answer key. If you fail, you have the option of taking the test again at no additional cost.
- For questions, contact Lippincott Professional Development: 1-800-787-8985.

Registration Deadline: September 2, 2022

Disclosure Statement:

The authors and planners have disclosed that they have no financial relationships related to this article.

Provider Accreditation:

Lippincott Professional Development will award 1.5 contact hours for this continuing nursing education activity.

Lippincott Professional Development is accredited as a provider of continuing nursing education by the American Nurses Credentialing Center's Commission on Accreditation.

This activity is also provider approved by the California Board of Registered Nursing, Provider Number CEP 11749. Lippincott Professional Development is also an approved provider of continuing nursing education by the District of Columbia, Florida, and Georgia, #50-1223.

Your certificate is valid in all states.

Payment:

- The registration fee for this test is $17.95