OPEN ACCESS

Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research

Nicole Gray Weiskopf, Chunhua Weng

Department of Biomedical Informatics, Columbia University, New York, New York, USA

Correspondence to
Nicole Gray Weiskopf, Department of Biomedical Informatics, Columbia University, 622 W 168th Street, VC-5, New York, NY 10032, USA; nicole.weiskopf@dbmi.columbia.edu

Received 3 November 2011
Accepted 3 May 2012
Published Online First
25 June 2012

ABSTRACT

Objective To review the methods and dimensions of data quality assessment in the context of electronic health record (EHR) data reuse for research.

Materials and methods A review of the clinical research literature discussing data quality assessment methodology for EHR data was performed. Using an iterative process, the aspects of data quality being measured were abstracted and categorized, as well as the methods of assessment used.

Results Five dimensions of data quality were identified, which are completeness, correctness, concordance, plausibility, and currency, and seven broad categories of data quality assessment methods: comparison with gold standards, data element agreement, data source agreement, distribution comparison, validity checks, log review, and element presence.

Discussion Examination of the methods by which clinical researchers have investigated the quality and suitability of EHR data for research shows that there are fundamental features of data quality, which may be difficult to measure, as well as proxy dimensions. Researchers interested in the reuse of EHR data for clinical research are recommended to consider the adoption of a consistent taxonomy of EHR data quality, to remain aware of the task-dependence of data quality, to integrate work on data quality assessment from other fields, and to adopt systematic, empirically driven, statistically based methods of data quality assessment.

Conclusion There is currently little consistency or potential generalizability in the methods used to assess EHR data quality. If the reuse of EHR data for clinical research is to become accepted, researchers should adopt validated, systematic methods of EHR data quality assessment.

As the adoption of electronic health records (EHRs) has made it easier to access and aggregate clinical data, there has been growing interest in conducting research with data collected during the course of clinical care.[1 2] The Natonal Institutes of Health has called for increasing the reuse of electronic records for research, and the clinical research community has been actively seeking methods to enable secondary use of clinical data.[3] EHRs surpass many existing registries and data repositories in volume, and the reuse of these data may diminish the costs and inefficiencies associated with clinical research. Like other forms of retrospective research, studies that make use of EHR data do not require patient recruitment or data collection, both of which are expensive and time-consuming processes. The data from EHRs also offer a window into the medical care, status, and outcomes of a diverse population that is representative of actual patients. The secondary use of data collected in EHRs is a promising step towards decreasing research costs, increasing patient-centered research, and speeding the rate of new medical discoveries.

Despite these benefits, reuse of EHR data has been limited by a number of factors, including concerns about the quality of the data and their suitability for research. It is generally accepted that, as a result of differences in priorities between clinical and research settings, clinical data are not recorded with the same care as research data.[4] Moreover, Burnum[5] stated that the introduction of health information technology like EHRs has led not to improvements in the quality of the data being recorded, but rather to the recording of a greater quantity of bad data. Due to such concerns about data quality, van der Lei[6] warned specifically against the reuse of clinical data for research and proposed what he called the first law of informatics: '[d]ata shall be used only for the purpose for which they were collected'.

Although such concerns about data quality have existed since EHRs were first introduced, there remains no consensus as to the quality of electronic clinical data or even agreement as to what 'data quality' actually means in the context of EHRs. One of the most broadly adopted conceptualizations of quality comes from Juran,[7] who said that quality is defined through 'fitness for use'. In the context of data quality, this means that data are of sufficient quality when they serve the needs of a given user pursuing specific goals.

Past study of EHR data quality has revealed highly variable results. Hogan and Wagner,[8] in their 1997 literature review, found that the correctness of data ranged between 44% and 100%, and completeness between 1.1% and 100%, depending on the clinical concepts being studied. Similarly, Thiru *et al*,[9] in calculating the sensitivity of different types of EHR data in the literature, found values ranging between 0.26 and 1.00. In a 2010 review, Chan *et al*[10] looked at the quality of the same clinical concepts across multiple institutions, and still found a great deal of variability. The completeness of blood pressure recordings, for example, fell anywhere between 0.1% and 51%. Due to differences in measurement, recording, information systems, and clinical focus, the quality of EHR data is highly variable. Therefore, it is generally inadvisable to make assumptions about one EHR-derived dataset based on another. We need systematic methods that will allow us to assess the

quality of an EHR-derived dataset for a given research task.

Our review primarily differs from those highlighted above in its focus. The previous reviews looked at data quality findings, while ours instead focuses on the methods that have been used to assess data quality. In fact, the earlier reviews were explicitly limited to studies that relied on the use of a reference standard, while we instead explore a range of data quality assessment methods. The contributions of this literature review are an empirically based conceptual model of the dimensions of EHR data quality studied by clinical researchers and a summary and critique of the methods that have been used to assess EHR data quality, specifically within the context of reusing clinical data for research. Our goal is to develop a systematic understanding of the approaches that may be used to determine the suitability of EHR data for a specific research goal.

METHODS

We identified articles in the literature by performing a search of the literature using standard electronic bibliographic tools. The literature search was performed by the first author on PubMed in February of 2012. As observed by Hogan and Wagner[8] in their literature review, there is no medical subheadings (MeSH) term for data quality, so a brief exploratory review was performed to identify relevant keywords. The final list included 'data quality', 'data accuracy', 'data reliability', 'data validity', 'data consistency', 'data completeness', and 'data error'. The MeSH heading for EHR was not introduced until 2010, so the older and more general MeSH heading 'medical record systems, computerized' was used instead. The phrases 'EHR', 'electronic medical record', and 'computerized medical record' were also included in order to capture articles that may not have been tagged correctly. We searched for articles including at least one of the quality terms and at least one of the EHR terms. Results were limited to English language articles. The full query is shown below.

> ('data quality' OR 'data accuracy' OR 'data reliability' OR 'data validity' OR 'data consistency' OR 'data completeness' OR 'data errors' OR 'data error') AND (EHR OR electronic medical record OR computerized medical record OR medical records systems, computerized [mh]) AND English[lang]

This search produced 230 articles, all of which were manually reviewed by the first author to determine if they met the selection criteria. In particular, the articles retained for further review: (1) included original research using data quality assessment methods; (2) focused on data derived from an EHR or related system; and (3) were published in a peer-reviewed journal. Articles dealing with data from purely administrative systems (eg, claims databases) were not included. These inclusion criteria resulted in 44 relevant articles. Next, we performed an in-depth ancestor search, reviewing the references of all of the articles in the original pool of 44. This allowed us to identify an additional 51 articles, resulting in a final pool of 95 articles meeting our inclusion criteria that were then used to derive results in this study.

From each article we abstracted the features of data quality examined, the methods of assessment used, and basic descriptive information including about the article and the type of data being studied. Through iterative review of the abstracted data, we derived broad dimensions of data quality and general categories of assessment strategies commonly described in the literature. Finally, we reviewed the 95 articles again, categorizing every article based on the dimension or dimensions being assessed, as well as the assessment strategies used for each of those dimensions.

Before beginning this analysis, we searched for preexisting models of EHR data quality, but were unable to find any. We decided that the potential benefits of adapting a data quality model from another field were outweighed by the risks of approaching our analysis through the lens of a model that had not been validated in the area of EHR data quality. Furthermore, using an existing model to guide analysis is a deductive approach, which has the potential to obscure information contained in the data.[11] By imposing an existing model from a different discipline, we would have run the risk of missing important findings. Therefore, we decided to use an inductive, data-driven coding approach. This approach provides advantages over the deductive approach by allowing us better coverage of the dimensions and methods of data quality assessment.

RESULTS

The majority of papers reviewed (73%) looked at structured data only, or at a combination of structured and unstructured data (22%). For our purposes, unstructured data types include free-entry text, while structured data types include coded data, values from pre-populated lists, or data entered into fields requiring specific alphanumeric formats.

Ignoring variations due to lexical categories and negation, the articles contained 27 unique terms describing dimensions of data quality. Features of data quality that were mentioned or described but not assessed were not included in our analysis. We grouped the terms together based on shared definitions. A few features of good data described in the literature, including sufficient granularity and the use of standards, were not included in our analyses. This decision was made due to the limited discussion of these features, the fact that they could be considered traits of good data practice instead of data quality, and because no assessment methods were described. Overall, we empirically derived five substantively different dimensions of data quality from the literature. The dimensions are defined below.

- Completeness: Is a truth about a patient present in the EHR?
- Correctness: Is an element that is present in the EHR true?
- Concordance: Is there agreement between elements in the EHR, or between the EHR and another data source?
- Plausibility: Does an element in the EHR makes sense in light of other knowledge about what that element is measuring?
- Currency: Is an element in the EHR a relevant representation of the patient state at a given point in time?

The list of data quality terms and their mappings to the five dimensions described above are shown in table 1. The terms chosen to denote each of the dimensions were the clearest and

Table 1 Terms used in the literature to describe the five common dimensions of data quality

Completeness	Correctness	Concordance	Plausibility	Currency
Accessibility	Accuracy	Agreement	Accuracy	Recency
Accuracy	Corrections made	Consistency	Believability	Timeliness
Availability	Errors	Reliability	Trustworthiness	
Missingness	Misleading	Variation	Validity	
Omission	Positive predictive value			
Presence	Quality			
Quality	Validity			
Rate of recording				
Sensitivity				
Validity				

least ambiguous from each of the groups. There was a great deal of variability and overlap in the terms used to describe each of these dimensions. 'Accuracy', for example, was sometimes used as a synonym for correctness, but in other articles meant both correctness and completeness. The dimensions themselves, however, were abstracted in such a way as to be exhaustive and mutually exclusive based on their definitions. Every article identified could be matched to one or more of the dimensions.

A similar process was used to identify the most common methods of data quality assessment. The strategies used to assess the dimensions of data quality fell into seven broad categories of methods, many of which were used to assess multiple dimensions. These general methods are listed and defined below.

- Gold standard: A dataset drawn from another source or multiple sources, with or without information from the EHR, is used as a gold standard.
- Data element agreement: Two or more elements within an EHR are compared to see if they report the same or compatible information.
- Element presence: A determination is made as to whether or not desired or expected data elements are present.
- Data source agreement: Data from the EHR are compared with data from another source to determine if they are in agreement.
- Distribution comparison: Distributions or summary statistics of aggregated data from the EHR are compared with the expected distributions for the clinical concepts of interest.
- Validity check: Data in the EHR are assessed using various techniques that determine if values 'make sense'.
- Log review: Information on the actual data entry practices (eg, dates, times, edits) is examined.

A summary of which methods were used to assess which dimensions is shown in table 2. The graph in figure 1 shows the strength of the pairwise relationships between the dimensions and methods. Some of the methods were used to assess only certain dimensions of data quality, whereas other methods were applied more broadly. Element presence, for example, was used to assess completeness, but none of the other dimensions. Data element agreement and data source agreement, however, were applied more broadly. Most of the dimensions were assessed using an assortment of methods, but currency was only measured using a single approach.

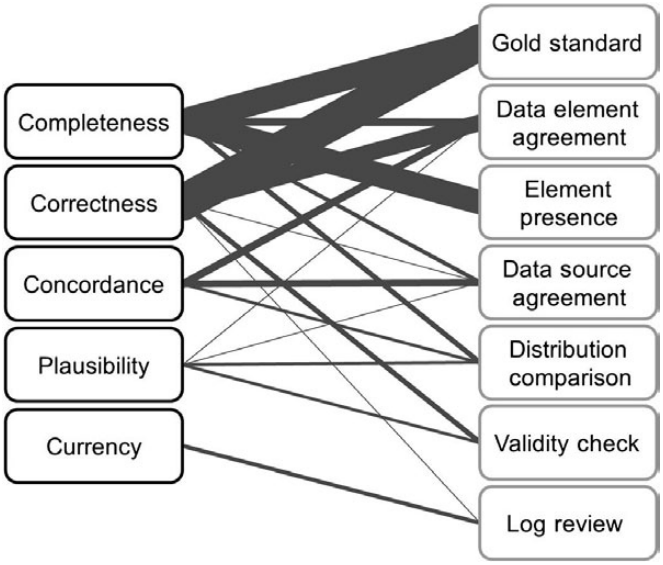

Figure 1 Mapping between dimensions of data quality and data quality assessment methods. Dimensions are listed on the left and methods of assessment on the right, both in decreasing order of frequency from top to bottom. The weight of the edge connecting a dimension and method indicates the relative frequency of that combination.

Completeness

Completeness was the most commonly assessed dimension of data quality and was an area of focus in 61 (64%) of the articles. Generally speaking, completeness referred to whether or not a truth about a patient was present in the EHR. Most of the articles used the term completeness to describe this dimension, but some also referred to data availability or missing data. In others, completeness was subsumed into more general concepts like accuracy or quality. Some articles cited the statistical definition of completeness suggested by Hogan and Wagner,[8] in which completeness is equivalent to sensitivity.

Many articles assessed EHR data completeness by using another source of data as a gold standard. The gold standards used included concurrently kept paper records,[12–18] information supplied by patients,[19–21] review of data by patients,[22–25] clinical encounters with patients,[26–28] information presented by trained standard patients,[29 30] information requested from the treating physician,[31] and alternative data sources from which EHR elements were abstracted.[32 33] A similar approach involved triangulating data from multiple sources within the EHR to create a gold standard.[34 35]

Other researchers simply looked at the presence or absence of elements in the EHR. In some cases, these were elements that were expected to be present, even if they were not needed for any specific task.[36–38 57 58 68 69 74–81] In other situations, the elements examined were dependent upon the task at hand, meaning that the researchers determined whether or not the EHR data were

Table 2 The dimensions of data quality and methods of data quality assessment

Dimension	Completeness	Correctness	Concordance	Plausibility	Currency	Total
Method						
Gold standard	12–35	12–23 25 26 28–48				37
Data element agreement	49–55 56	49 50 52–54 56–67	51 59 68–72	38 73		26
Element presence	36–39 49 57–59 68 69 74–86					23
Data source agreement	79 87–89	90	91–96			11
Distribution comparison	97–100		31 57 97	50 101–103		10
Validity checks		32 62 104–106		73 89		7
Log review		73			36 46 52 84	5
Total	61	57	16	7	4	

In decreasing order of frequency, the dimensions are listed from left to right, and the methods from top to bottom. The numbers in the cells correspond to the article references featuring each dimension–method pair.

complete enough for a specific purpose.[39 49 59 82–86] Other methods for assessing completeness included looking at agreement between elements from the same source,[49–55 56] agreement between the EHR and paper records,[79 87 88] agreement between the EHR and another electronic source of data,[79 89] and comparing distributions of occurrences of certain elements between practices[97] or with nationally recorded rates.[98–100]

Correctness

The second most commonly assessed dimension of data quality was correctness, which was included in 57 (60%) of the articles. EHR data were considered correct when the information they contained was true. Other terms that were commonly used to describe this concept included accuracy, error, and quality. Occasionally, correctness included completeness, due to the fact that some researchers consider missing data to be incorrect (ie, errors of omission). The definition of correctness suggested by Hogan and Wagner[8] states that data correctness is the proportion of data elements present that are correct, which is equivalent to positive predictive value.

Comparison of EHR data with a gold standard was by far the most frequently used method for assessing correctness. These gold standards included: paper records;[12–18 38 40 41] information supplied by patients through interviews,[19 20 36 42] questionnaires,[21 43] data review,[22 23 25 44] or direct data entry;[37] clinical encounters with patients;[26 28 45] information presented by trained standard patients;[29 30] automatically recorded data;[46] contact with the treating physician;[31 39 47] and alternative data sources from which information matching EHR elements were abstracted.[32 33] Some researchers developed gold standards by extracting and triangulating data from within the EHR.[34 35 43 48]

The second most common approach to assessing correctness was to look at agreement between elements within the EHR. Usually this involved verifying a diagnosis by looking at associated procedures, medications, or laboratory values.[49 50 52–54 57 58 60 61] Similarly, some articles reported on agreement between related elements[56 62] and errors identified through the examination of the use of copy and paste practices.[63 64] Other researchers looked specifically at agreement between structured elements and unstructured data within EHRs.[59 65] One of the more formal approaches described for assessing correctness was the data quality probe, proposed by Brown and Warmington,[66 67] which is a query that, when run against an EHR database, only returns cases with some disagreement between data elements.

A few articles described the use of validity checks to assess correctness. These included review of changes of sequential data over time,[105] identifying end digit preferences in blood pressure values,[104 106] and comparing elements with their expected value ranges.[32 62] Two other approaches to were using corrections seen in log files as a proxy for correctness,[73] and comparing data on the same patients from a registry and an EHR.[90]

Concordance

Sixteen (17%) of the articles reviewed assessed concordance. Data were considered concordant when there was agreement or compatibility between data elements. This may mean that two elements recording the same information for a single patient have the same value, or that elements recording different information have values that make sense when considered together (eg, biological sex is recorded as female, and procedure is recorded as gynecological examination). Measurement of concordance is generally based on elements contained within the EHR, but some researchers also included information from other data sources. Common terms used in the literature to describe data concordance include agreement and consistency.

The most common approach to assessing concordance was to look at agreement between elements within the EHR,[59 68–70] especially diagnoses and associated information such as medications or procedures.[51 71 72] The second most common method used to assess concordance was to look at the agreement of EHR data with data from other sources. These other sources included billing information,[91] paper records,[92–94] patient-reported data,[95] and physician-reported data.[96] Another approach was to compare distributions of data within the EHR with distributions of the same information from similar medical practices[57 97] or with national rates.[31]

Plausibility

Seven (7%) of the articles assessed the plausibility of EHR data. In this context, data were plausible if they were in agreement with general medical knowledge or information and were therefore feasible. In other words, assessments of plausibility were intended to determine whether or not data could be trusted or if they were of suspect quality. Other terms that were used to discuss and describe EHR data plausibility include data validity and integrity.

The most common approach to assessing the plausibility of EHR data was to perform some sort of validity check to determine if specific elements within the EHR were likely to be true or not. This included looking for elements with values that were outside biologically plausible ranges or that changed implausibly over time[89] or zero-valued elements.[73] Other researchers compared distributions of data values between practices[50 101] or with national rates,[102 103] or looked at agreement between related elements.[38 73]

Currency

The currency of EHR data was assessed in four (4%) of the 95 articles. Currency was often referred to in the literature as timeliness or recency. Data were considered current if they were recorded in the EHR within a reasonable period of time following measurement or, alternatively, if they were representative of the patient state at a desired time of interest. In all four articles, currency was assessed through the review of data entry logs. In three of the four, researchers reviewed whether desired data were entered into the EHR within a set time limit.[36 46 52] In the fourth, researchers considered whether each type of data element was measured recently enough to be considered medically relevant.[84]

DISCUSSION

We identified five dimensions of data quality and seven categories of data quality assessment methods. Examination of the types of methods used, as well as overlap of the methods between dimensions, reveals significant patterns and gaps in knowledge. Below, we explore the major findings of the literature review, specifically highlighting areas that require further attention, and make suggestions for future research.

Terminology and dimensions of data quality

One of the biggest difficulties in conducting this review resulted from the inconsistent terminology used to discuss data quality. We had not expected, for example, the overlap of terms between dimensions, or the fact that the language within a single article was sometimes inconsistent. The clinical research community has largely failed to develop or adopt a consistent taxonomy of data quality.

There is, however, overlap between the dimensions of data quality identified during this review and those described in preexisting taxonomies and models of data quality. Wang and Strong's[107] conceptual framework of data quality, for example, contains 15 dimensions, grouped into four categories: intrinsic, contextual, representational, and accessibility. Our review focused on intrinsic (inherent to the data) and contextual (task-dependent) data quality issues. The dimensions we identified overlapped with two of the intrinsic features (accuracy and believability, which are equivalent to correctness and plausibility) and two of the contextual features (timeliness, which is equivalent to currency, and completeness). The only dimension we identified that does not appear in Wang and Strong's[107] framework is concordance.

The Institute of Medicine identified four attributes of data quality relevant to patient records: completeness, accuracy, legibility, and meaning (related to comprehensibility).[108] As the Institute of Medicine points out, electronic records by their nature negate many of the concerns regarding legibility, so we are left with three relevant attributes, two of which we identified through our review. Meaning is a more abstract concept and is likely to be difficult to measure objectively, which may be why we did not observe assessments of this dimension in the literature.

Although the five dimensions of data quality derived during our review were treated as mutually exclusive within the literature, we feel that only three can be considered fundamental: correctness, completeness, and currency. By this we mean that these dimensions are non-reducible, and describe core concepts of data quality as it relates to EHR data reuse. Concordance and plausibility, on the other hand, while discussed as separate features of data quality, appear to serve as proxies for the fundamental dimensions when it is not possible to assess them directly. This supposition is supported by the overlap observed in the methods used to assess concordance and plausibility with those used to assess correctness and completeness. A lack of concordance between two data sources, for example, indicates error in one or both of those sources: an error of omission, resulting in a lack of completeness, or an error of commission, resulting in a lack of correctness. Similarly, data that do not appear to be plausible may be incorrect, as in the case of a measurement that fails a range check, or incomplete, such as aggregated diagnosis rates within a practice that do not match the expected population rates. It may be that correctness, completeness, and currency are properties of data quality, while plausibility and concordance are methodological approaches to assessing data quality. In addition, researchers may refer to plausibility or concordance when they believe that there are problems with completeness or correctness, but have no way to be certain that errors exist or which data elements might be wrong.

Data quality assessment methodology

We observed a number of noteworthy patterns within the literature in terms of the types of data quality assessments used and the manner in which data quality assessment was discussed. For example, 37 of the 95 articles in our sample relied on a gold standard to assess data quality. There are a few problems with this approach. First, the data sources used could rarely be considered true gold standards. Paper records, for example, may sometimes be more trusted than electronic records, but they should not be considered entirely correct or complete. Perhaps more importantly, a gold standard for EHR data is simply not available in most cases. This will become more problematic as the use of de-identified datasets for research becomes more common. A 'fitness for purpose' approach, which suggests that the quality of each dataset compiled for a specific task must be assessed, necessitates the adoption of alternatives to gold standard-based methods.

In addition to the overreliance on gold standards, the majority of the studies we identified relied upon an 'intuitive' understanding of data quality and used ad hoc methods to assess data quality. This tendency has also been observed in other fields.[107 109] Most of the studies included in this review presented assessment methodologies that were developed with a minimal empirical or theoretical basis. Only a few researchers made the effort to develop generalizable approaches that could be used as a step towards a standard methodology. Faulconer and de Lusignan,[52] for example, proposed a multistep, statistically driven approach to data quality assessment. Hogan and Wagner[8] suggested specific statistical measures of the correctness and completeness of EHR elements that have been adopted by other researchers. Certain methods, including comparing distributions of data from the EHR with expected distributions or looking for agreement between elements within the EHR, lend themselves more readily to generalization. Brown and Warmington's[66 67] data quality probes, for example, could be extended to various data elements, although they require detailed clinical knowledge to implement. Some researchers looking at the quality of research databases pulled from general practices in the UK have adopted relatively consistent approaches to comparing the distributions of data concerning specific clinical phenomena to information from registries and surveys.[31 98 99 102] In most cases, however, the specific assessment methods described in the literature would be difficult to apply to other datasets or research questions. If the reuse of EHR data for clinical research is to become common and feasible, development of standardized, systematic data quality assessment methods is vital.

In addition, if as a field we intend to adopt the concept of 'fitness for purpose', it is important to consider the intended research use of EHR data when determining if they are of sufficient quality.[110] Some dimensions may prove to be more task dependent, or subjective, while others are essentially task independent, or objective.[109] It will be important to develop a full understanding of the interrelationships of research tasks and data characteristics as they relate to data quality. For example, the completeness of a set of data elements required by one research protocol may differ from the completeness required for a different protocol. Many factors, including clinical focus, required resolution of clinical information, and desired effect size, can affect the suitability of a dataset for a specific research task.[26]

Future directions

We believe that efforts to reuse EHR data for clinical research would benefit most from work in a few specific areas: adopting a consistent taxonomy of EHR data quality; increasing awareness of task dependence; integrating work on data quality assessment from other fields; and adopting systematic, statistically based methods of data quality assessment. A taxonomy of data quality would enable a structured discourse and contextualize assessment methodologies. The findings in this review regarding the dimensions of data quality may serve as a stepping stone towards this goal. Task dependence is likely to become a growing issue as efforts to reuse EHR data for research increase, particularly as data quality assessment does not have a one-size-fits-all solution. One approach to addressing the problem of EHR data quality and suitability for reuse in research

would be to look at what has been done outside of clinical research, because data quality has been an area of study in fields ranging from finance to industrial engineering. Finally, it is important that the clinical researchers begin to move away from ad hoc approaches to data quality assessment. Validated methods that can be adapted for different research questions are the ideal goal.

Limitations

There were a number of limitations to this review. First, the search was limited. Due to the lack of MeSH terms for data quality and the variation in terminology used to discuss data quality, it is possible that our original search may have missed some relevant articles. We believe that our decision to review the references of each article improved the saturation of our sample.

It is also important to note that our classification process was largely subjective and was performed by only one of the authors. It is possible that the original researchers might disagree with our interpretations. We chose to use an iterative process to label and categorize the dimensions of data quality and methods of assessment described in each article in an effort to develop a consistent coding scheme.

Finally, it is likely that the dearth of literature discussing data quality in the reuse of EHR data for clinical research is due partly to underreporting. A common first step in analyzing any dataset is to review distributions, summary statistics, and histograms, but this process is rarely described in publications. Such methods are therefore likely to be underrepresented in this review. Greater transparency regarding data cleaning or checking steps would be advisable, as it could help to establish acceptable reporting standards for the reuse of EHR data in research.

CONCLUSION

The secondary use of EHR data is a promising area of research. However, the problems with EHR data quality necessitate the use of quality assessment methodologies to determine the suitability of these data for given research tasks. In this review of the literature we have identified the major dimensions of data quality that are of interest to researchers, as well as the general assessment techniques that have been utilized. Data quality is not a simple problem, and if the reuse of EHR data is to become an accepted approach to medical research, the clinical research community needs to develop validated, systematic methods of EHR data quality assessment. We encourage researchers to be consistent in their discussion of the dimensions of data quality, systematic in their approaches to measuring data quality, and to develop and share best practices for the assessment of EHR data quality in the context of reuse for clinical research.

Acknowledgments The authors would like to thank George Hripcsak, Adam Wilcox, and Suzanne Bakken for their assistance in the preparation of this manuscript.

Funding This research was supported by National Library of Medicine grants 5T15LM007079, R01LM009886, and R01LM010815.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

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. *J Am Med Inform Assoc* 2007;**14**:1–9.
2. **Hersh WR.** Adding value to the electronic health record through secondary use of data for quality assurance, research, and surveillance. *Am J Manag Care* 2007;**13**:277–8.
3. **National Center for Research Resources (US).** *Widening the Use of Electronic Health Record Data for Research*. 2009. http://videocast.nih.gov/summary.asp?live=8062 (accessed 9 Oct 2011).
4. **Weiner MG,** Embi PJ. Toward reuse of clinical data for research and quality improvement: the end of the beginning? *Ann Intern Med* 2009;**151**:359–60.
5. **Burnum JF.** The misinformation era: the fall of the medical record. *Ann Intern Med* 1989;**110**:482–4.
6. **van der Lei J.** Use and abuse of computer-stored medical records. *Methods Inf Med* 1991;**30**:79–80.
7. **Juran JM,** Gryna FM. *Juran's Quality Control Handbook*. 4th edn. New York: McGraw-Hill, 1988.
8. **Hogan WR,** Wagner MM. Accuracy of data in computer-based patient records. *J Am Med Inform Assoc* 1997;**4**:342–55.
9. **Thiru K,** Hassey A, Sullivan F. Systematic review of scope and quality of electronic patient record data in primary care. *BMJ* 2003;**326**:1070.
10. **Chan KS,** Fowles JB, Weiner JP. Review: electronic health records and the reliability and validity of quality measures: a review of the literature. *Med Care Res Rev* 2010;**67**:503–27.
11. **Thomas DR.** A general inductive approach for analyzing qualitative evaluation data. *Am J Eval* 2005;**27**:237–46.
12. **Ayoub L,** Fu L, Pena A, *et al*. Implementation of a data management program in a pediatric cancer unit in a low income country. *Pediatr Blood Cancer* 2007;**49**:23–7.
13. **Barrie JL,** Marsh DR. Quality of data in the Manchester orthopaedic database. *BMJ* 1992;**304**:159–62.
14. **Hohnloser JH,** Fischer MR, Konig A, *et al*. Data quality in computerized patient records. Analysis of a haematology biopsy report database. *Int J Clin Monit Comput* 1994;**11**:233–40.
15. **Mikkelsen G,** Aasly J. Consequences of impaired data quality on information retrieval in electronic patient records. *Int J Med Inform* 2005;**74**:387–94.
16. **Ricketts D,** Newey M, Patterson M, *et al*. Markers of data quality in computer audit: the Manchester Orthopaedic Database. *Ann R Coll Surg Engl* 1993;**75**:393–6.
17. **Roukema J,** Los RK, Bleeker SE, *et al*. Paper versus computer: feasibility of an electronic medical record in general pediatrics. *Pediatrics* 2006;**117**:15–21.
18. **Wallace CJ,** Stansfield D, Gibb Ellis KA, *et al*. Implementation of an electronic logbook for intensive care units. *Proc AMIA Symp* 2002:840–4.
19. **Kaboli PJ,** McClimon BJ, Hoth AB, *et al*. Assessing the accuracy of computerized medication histories. *Am J Manag Care* 2004;**10**:872–7.
20. **Porter SC,** Mandl KD. Data quality and the electronic medical record: a role for direct parental data entry. *Proc AMIA Symp* 1999:354–8.
21. **Whitelaw FG,** Nevin SL, Milne RM, *et al*. Completeness and accuracy of morbidity and repeat prescribing records held on general practice computers in Scotland. *Br J Gen Pract* 1996;**46**:181–6.
22. **Pyper C,** Amery J, Watson M, *et al*. Patients' experiences when accessing their on-line electronic patient records in primary care. *Br J Gen Pract* 2004;**54**:38–43.
23. **Staroselsky M,** Volk LA, Tsurikova R, *et al*. An effort to improve electronic health record medication list accuracy between visits: patients' and physicians' response. *Int J Med Inform* 2008;**77**:153–60.
24. **Staroselsky M,** Volk LA, Tsurikova R, *et al*. Improving electronic health record (EHR) accuracy and increasing compliance with health maintenance clinical guidelines through patient access and input. *Int J Med Inform* 2006;**75**:693–700.
25. **Weingart SN,** Cleary A, Seger A, *et al*. Medication reconciliation in ambulatory oncology. *Jt Comm J Qual Patient Saf* 2007;**33**:750–7.
26. **Logan JR,** Gorman PN, Middleton B. Measuring the quality of medical records: a method for comparing completeness and correctness of clinical encounter data. *Proc AMIA Symp* 2001:408–12.
27. **Smith PC,** Araya-Guerra R, Bublitz C, *et al*. Missing clinical information during primary care visits. *JAMA* 2005;**293**:565–71.
28. **Bentsen BG.** The accuracy of recording patient problems in family practice. *J Med Educ* 1976;**51**:311–16.
29. **Berner ES,** Kasiraman RK, Yu F, *et al*. Data quality in the outpatient setting: impact on clinical decision support systems. *AMIA Annu Symp Proc* 2005:41–5.
30. **Peabody JW,** Luck J, Jain S, *et al*. Assessing the accuracy of administrative data in health information systems. *Med Care* 2004;**42**:1066–72.
31. **Lewis JD,** Brensinger C. Agreement between GPRD smoking data: a survey of general practitioners and a population-based survey. *Pharmacoepidemiol Drug Saf* 2004;**13**:437–41.
32. **Staes CJ,** Bennett ST, Evans RS, *et al*. A case for manual entry of structured, coded laboratory data from multiple sources into an ambulatory electronic health record. *J Am Med Inform Assoc* 2006;**13**:12–15.
33. **Madsen KM,** Schonheyder HC, Kristensen B, *et al*. Can hospital discharge diagnosis be used for surveillance of bacteremia? A data quality study of a Danish hospital discharge registry. *Infect Control Hosp Epidemiol* 1998;**19**:175–80.
34. **Aronsky D,** Haug PJ. Assessing the quality of clinical data in a computer-based record for calculating the pneumonia severity index. *J Am Med Inform Assoc* 2000;**7**:55–65.
35. **Wagner MM,** Hogan WR. The accuracy of medication data in an outpatient electronic medical record. *J Am Med Inform Assoc* 1996;**3**:234–44.
36. **Ndira SP,** Rosenberger KD, Wetter T. Assessment of data quality of and staff satisfaction with an electronic health record system in a developing country

(Uganda): a qualitative and quantitative comparative study. *Methods Inf Med* 2008;**47**:489—98.
37. **Olola CH,** Narus S, Poynton M, *et al*. Patient-perceived usefulness of an emergency medical card and a continuity-of-care report in enhancing the quality of care. *Int J Qual Health Care* 2011;**23**:60—7.
38. **Pearson N,** O'Brien J, Thomas H, *et al*. Collecting morbidity data in general practice: the Somerset morbidity project. *BMJ* 1996;**312**:1517—20.
39. **Lo Re V 3rd,** Haynes K, Forde KA, *et al*. Validity of The Health Improvement Network (THIN) for epidemiologic studies of hepatitis C virus infection. *Pharmacoepidemiol Drug Saf* 2009;**18**:807—14.
40. **Dambro MR,** Weiss BD. Assessing the quality of data entry in a computerized medical records system. *J Med Syst* 1988;**12**:181—7.
41. **Nazareth I,** King M, Haines A, *et al*. Accuracy of diagnosis of psychosis on general practice computer system. *BMJ* 1993;**307**:32—4.
42. **Dawson C,** Perkins M, Draper E, *et al*. Are outcome data regarding the survivors of neonatal care available from routine sources? *Arch Dis Child Fetal Neonatal Ed* 1997;**77**:F206—10.
43. **Van Weel C.** Validating long term morbidity recording. *J Epidemiol Community Health* 1995;**49**(Suppl. 1):29—32.
44. **Powell J,** Fitton R, Fitton C. Sharing electronic health records: the patient view. *Inform Prim Care* 2006;**14**:55—7.
45. **Meara J,** Bhowmick BK, Hobson P. Accuracy of diagnosis in patients with presumed Parkinson's disease. *Age Ageing* 1999;**28**:99—102.
46. **Vawdrey DK,** Gardner RM, Evans RS, *et al*. Assessing data quality in manual entry of ventilator settings. *J Am Med Inform Assoc* 2007;**14**:295—303.
47. **Van Staa TP,** Abenhaim L, Cooper C, *et al*. The use of a large pharmacoepidemiological database to study exposure to oral corticosteroids and risk of fractures: validation of study population and results. *Pharmacoepidemiol Drug Saf* 2000;**9**:359—66.
48. **Margulis AV,** Garcia Rodriguez LA, Hernandez-Diaz S. Positive predictive value of computerized medical records for uncomplicated and complicated upper gastrointestinal ulcer. *Pharmacoepidemiol Drug Saf* 2009;**18**:900—9.
49. **Linder JA,** Kaleba EO, Kmetik KS. Using electronic health records to measure physician performance for acute conditions in primary care: empirical evaluation of the community-acquired pneumonia clinical quality measure set. *Med Care* 2009;**47**:208—16.
50. **de Lusignan S,** Chan T, Wood O, *et al*. Quality and variability of osteoporosis data in general practice computer records: implications for disease registers. *Public Health* 2005;**119**:771—80.
51. **de Lusignan S,** Valentin T, Chan T, *et al*. Problems with primary care data quality: osteoporosis as an exemplar. *Inform Prim Care* 2004;**12**:147—56.
52. **Faulconer ER,** de Lusignan S. An eight-step method for assessing diagnostic data quality in practice: chronic obstructive pulmonary disease as an exemplar. *Inform Prim Care* 2004;**12**:243—54.
53. **Hassey A,** Gerrett D, Wilson A. A survey of validity and utility of electronic patient records in a general practice. *BMJ* 2001;**322**:1401—5.
54. **Tang PC,** Ralston M, Arrigotti MF, *et al*. Comparison of methodologies for calculating quality measures based on administrative data versus clinical data from an electronic health record system: implications for performance measures. *J Am Med Inform Assoc* 2007;**14**:10—15.
55. **Hohnloser JH,** Puerner F, Soltanian H. Improving coded data entry by an electronic patient record system. *Methods Inf Med* 1996;**35**:108—11.
56. **Horsfield P.** Trends in data recording by general practice teams: an analysis of data extracted from clinical computer systems by the PRIMIS project. *Informat Prim Care* 2002;**10**:227—34.
57. **Pringle M,** Ward P, Chilvers C. Assessment of the completeness and accuracy of computer medical records in four practices committed to recording data on computer. *Br J Gen Pract* 1995;**45**:537—41.
58. **Thiru K,** De Lusignan S, Hague N. Have the completeness and accuracy of computer medical records in general practice improved in the last five years? The report of a two-practice pilot study. *Health Informatics Journal* 1999;**5**:224—32.
59. **Botsis T,** Hartvigsen G, Chen F, *et al*. Secondary use of EHR: data quality issues and informatics opportunities. *AMIA Summits Transl Sci Proc* 2010;**2010**:1—5.
60. **de Lusignan S,** Khunti K, Belsey J, *et al*. A method of identifying and correcting miscoding, misclassification and misdiagnosis in diabetes: a pilot and validation study of routinely collected data. *Diabet Med* 2010;**27**:203—9.
61. **de Burgos-Lunar C,** Salinero-Fort MA, Cardenas-Valladolid J, *et al*. Validation of diabetes mellitus and hypertension diagnosis in computerized medical records in primary health care. *BMC Med Res Methodol* 2011;**11**:146.
62. **Basden A,** Clark EM. Data integrity in a general practice computer system (CLINICS). *Int J Biomed Comput* 1980;**11**:511—19.
63. **Hammond KW,** Helbig ST, Benson CC, *et al*. Are electronic medical records trustworthy? Observations on copying, pasting and duplication. *AMIA Annu Symp Proc* 2003:269—73.
64. **Weir CR,** Hurdle JF, Felgar MA, *et al*. Direct text entry in electronic progress notes. An evaluation of input errors. *Methods Inf Med* 2003;**42**:61—7.
65. **Hogan WR,** Wagner MM. Free-text fields change the meaning of coded data. *Proc AMIA Annu Fall Symp* 1996:517—21.
66. **Brown PJ,** Warmington V. Data quality probes-exploiting and improving the quality of electronic patient record data and patient care. *Int J Med Inform* 2002;**68**:91—8.
67. **Brown PJ,** Warmington V. Info-tsunami: surviving the storm with data quality probes. *Inform Prim Care* 2003;**11**:229—33; discussion 234—7.
68. **Goulet JL,** Erdos J, Kancir S, *et al*. Measuring performance directly using the veterans health administration electronic medical record: a comparison with external peer review. *Med Care* 2007;**45**:73—9.
69. **Jelovsek F,** Hammond W. Formal error rate in a computerized obstetric medical record. *Methods Inf Med* 1978;**17**:151—7.
70. **Owen RR,** Thrush CR, Cannon D, *et al*. Use of electronic medical record data for quality improvement in schizophrenia treatment. *J Am Med Inform Assoc* 2004;**11**:351—7.
71. **Stein HD,** Nadkarni P, Erdos J, *et al*. Exploring the degree of concordance of coded and textual data in answering clinical queries from a clinical data repository. *J Am Med Inform Assoc* 2000;**7**:42—54.
72. **Terry AL,** Chevendra V, Thind A, *et al*. Using your electronic medical record for research: a primer for avoiding pitfalls. *Fam Pract* 2010;**27**:121—6.
73. **Benson M,** Junger A, Quinzio L, *et al*. Influence of the method of data collection on the documentation of blood-pressure readings with an Anesthesia Information Management System (AIMS). *Methods Inf Med* 2001;**40**:190—5.
74. **Einbinder JS,** Rury C, Safran C. Outcomes research using the electronic patient record: beth Israel Hospital's experience with anticoagulation. *Proceedings of the Annual Symposium on Computer Application in Medical Care*. 1995:819—23. This conference was held in New Orleans, Louisiana from October 28th to November 1st in 1995. http://www.ncbi.nlm.nih.gov/pmc/issues/173480/
75. **Forster M,** Bailey C, Brinkhof MW, *et al*; ART-LINC collaboration of International Epidemiological Databases to Evaluate AIDS. Electronic medical record systems, data quality and loss to follow-up: survey of antiretroviral therapy programmes in resource-limited settings. *Bull WHO* 2008;**86**:939—47.
76. **Hahn KA,** Ohman-Strickland PA, Cohen DJ, *et al*. Electronic medical records are not associated with improved documentation in community primary care practices. *Am J Med Qual* 2011;**26**:272—7.
77. **Jones RB,** Hedley AJ. A computer in the diabetic clinic. Completeness of data in a clinical information system for diabetes. *Practical Diabetes International* 1986;**3**:295—6.
78. **Porcheret M,** Hughes R, Evans D, *et al*. Data quality of general practice electronic health records: the impact of a program of assessments, feedback, and training. *J Am Med Inform Assoc* 2004;**11**:78—86.
79. **Scobie S,** Basnett I, McCartney P. Can general practice data be used for needs assessment and health care planning in an inner-London district? *J Public Health Med* 1995;**17**:475—83.
80. **Soto CM,** Kleinman KP, Simon SR. Quality and correlates of medical record documentation in the ambulatory care setting. *BMC Health Serv Res* 2002;**2**:22.
81. **Tang PC,** LaRosa MP, Gorden SM. Use of computer-based records, completeness of documentation, and appropriateness of documented clinical decisions. *J Am Med Inform Assoc* 1999;**6**:245—51.
82. **de Lusignan S,** Hague N, Brown A, *et al*. An educational intervention to improve data recording in the management of ischaemic heart disease in primary care. *J Public Health (Oxf)* 2004;**26**:34—7.
83. **Jensen RE,** Chan KS, Weiner JP, *et al*. Implementing electronic health record-based quality measures for developmental screening. *Pediatrics* 2009;**124**: e648—54.
84. **Williams JG.** Measuring the completeness and currency of codified clinical information. *Methods Inf Med* 2003;**42**:482—8.
85. **Agnew-Blais JC,** Coblyn JS, Katz JN, *et al*. Measuring quality of care for rheumatic diseases using an electronic medical record. *Ann Rheum Dis* 2009;**68**:680—4.
86. **Asche C,** Said Q, Joish V, *et al*. Assessment of COPD-related outcomes via a national electronic medical record database. *Int J Chron Obstruct Pulmon Dis* 2008;**3**:323—6.
87. **Jick H,** Terris BZ, Derby LE, *et al*. Further validation of information recorded on a general practitioner based computerized data resource in the United Kingdom. *Pharmacoepidemiology and Drug Safety* 1992;**1**:347—9.
88. **Neal RD,** Heywood PL, Morley S. Real world data — retrieval and validation of consultation data from four general practices. *Family Practice* 1996; **13**:455—61.
89. **Noel PH,** Copeland LA, Perrin RA, *et al*. VHA Corporate Data Warehouse height and weight data: opportunities and challenges for health services research. *J Rehabil Res Dev* 2010;**47**:739—50.
90. **Lawrenson R,** Todd JC, Leydon GM, *et al*. Validation of the diagnosis of venous thromboembolism in general practice database studies. *Br J Clin Pharmacol* 2000;**49**:591—6.
91. **Roos LL,** Sharp SM, Wajda A. Assessing data quality: a computerized approach. *Soc Sci Med* 1989;**28**:175—82.
92. **Mikkelsen G,** Aasly J. Concordance of information in parallel electronic and paper based patient records. *Int J Med Inform* 2001;**63**:123—31.
93. **Jick H,** Jick SS, Derby LE. Validation of information recorded on general practitioner based computerised data resource in the United Kingdom. *BMJ* 1991;**302**:766—8.
94. **Stausberg J,** Koch D, Ingenerf J, *et al*. Comparing paper-based with electronic patient records: lessons learned during a study on diagnosis and procedure codes. *J Am Med Inform Assoc* 2003;**10**:470—7.
95. **Pakhomov SV,** Jacobsen SJ, Chute CG, *et al*. Agreement between patient-reported symptoms and their documentation in the medical record. *Am J Manag Care* 2008;**14**:530—9.

96. **Conroy MB,** Majchrzak NE, Silverman CB, *et al*. Measuring provider adherence to tobacco treatment guidelines: a comparison of electronic medical record review, patient survey, and provider survey. *Nicotine Tob Res* 2005;**7**(Suppl. 1): S35—43.
97. **de Lusignan S,** Chan T, Wells S, *et al*. Can patients with osteoporosis, who should benefit from implementation of the national service framework for older people, be identified from general practice computer records? A pilot study that illustrates the variability of computerized medical records and problems with searching them. *Public Health* 2003;**117**:438—45.
98. **Haynes K,** Forde KA, Schinnar R, *et al*. Cancer incidence in The Health Improvement Network. *Pharmacoepidemiol Drug Saf* 2009;**18**:730—6.
99. **Iyen-Omofoman B,** Hubbard RB, Smith CJ, *et al*. The distribution of lung cancer across sectors of society in the United Kingdom: a study using national primary care data. *BMC Public Health* 2011;**11**:857.
100. **Johnson N,** Mant D, Jones L, *et al*. Use of computerised general practice data for population surveillance: comparative study of influenza data. *BMJ* 1991;**302**:763—5.
101. **Haynes K,** Bilker WB, Tenhave TR, *et al*. Temporal and within practice variability in the health improvement network. *Pharmacoepidemiol Drug Saf* 2011;**20**:948—55.
102. **Kaye JA,** Derby LE, del Mar Melero-Montes M, *et al*. The incidence of breast cancer in the General Practice Research Database compared with national cancer registration data. *Br J Cancer* 2000;**83**:1556—8.
103. **Hansell A,** Hollowell J, Nichols T, *et al*. Use of the General Practice Research Database (GPRD) for respiratory epidemiology: a comparison with the 4th Morbidity Survey in General Practice (MSGP4). *Thorax* 1999;**54**(1):413—9.
104. **de Lusignan S,** Belsey J, Hague N, *et al*. End-digit preference in blood pressure recordings of patients with ischaemic heart disease in primary care. *J Hum Hypertens* 2004;**18**:261—5.
105. **Haerian K,** McKeeby J, Dipatrizio G, *et al*. Use of clinical alerting to improve the collection of clinical research data. *AMIA Annu Symp Proc* 2009;**2009**:218—22.
106. **Alsanjari ON,** de Lusignan S, van Vlymen J, *et al*. Trends and transient change in end-digit preference in blood pressure recording: studies of sequential and longitudinal collected primary care data. *Int J Clin Pract* 2012;**66**:37—43.
107. **Wang RY,** Strong DM. Beyond accuracy: what data quality means to data consumers. *J Manag Inform Syst* 1996;**12**:5—34.
108. **Committee on Improving the Patient Record Institute of Medicine.** *The Computer-Based Patient Record: An Essential Technology for Health Care, Revised Edition*. Institute of Medicine. The National Academies Press, 1997. http://iom.edu/Reports/1997/The-Computer-Based-Patient-Record-An-Essential-Technology-for-Health-Care-Revised-Edition.aspx
109. **Pipino LL,** Lee YW, Wang RY. Data quality assessment. *Comm ACM* 2002;**45**:211—18.
110. **de Lusignan S.** The optimum granularity for coding diagnostic data in primary care: report of a workshop of the EFMI Primary Care Informatics Working Group at MIE 2005. *Inform Prim Care* 2006;**14**:133—7.