eGEMs
Generating Evidence & Methods to improve patient outcomes

A Data Quality Assessment Guideline for Electronic Health Record Data Reuse

Nicole G. Weiskopf, PhD;[i] Suzanne Bakken, RN, PhD;[ii,iii] George Hripcsak, MD, MS;[ii] Chunhua Weng, PhD[ii]

ABSTRACT

Introduction: We describe the formulation, development, and initial expert review of 3x3 Data Quality Assessment (DQA), a dynamic, evidence-based guideline to enable electronic health record (EHR) data quality assessment and reporting for clinical research.

Methods: 3x3 DQA was developed through the triangulation results from three studies: a review of the literature on EHR data quality assessment, a quantitative study of EHR data completeness, and a set of interviews with clinical researchers. Following initial development, the guideline was reviewed by a panel of EHR data quality experts.

Results: The guideline embraces the task-dependent nature of data quality and data quality assessment. The core framework includes three constructs of data quality: complete, correct, and current data. These constructs are operationalized according to the three primary dimensions of EHR data: patients, variables, and time. Each of the nine operationalized constructs maps to a methodological recommendation for EHR data quality assessment. The initial expert response to the framework was positive, but improvements are required.

Discussion: The initial version of 3x3 DQA promises to enable explicit guideline-based best practices for EHR data quality assessment and reporting. Future work will focus on increasing clarity on how and when 3x3 DQA should be used during the research process, improving the feasibility and ease-of-use of recommendation execution, and clarifying the process for users to determine which operationalized constructs and recommendations are relevant for a given dataset and study.

[i]Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, [ii]Department of Biomedical Informatics, Columbia University, [iii]School of Nursing, Columbia University

Introduction

Previous research has highlighted the need for systematic methods for electronic health record (EHR) data quality assessment within the context of reuse. The quality of EHR data, which we define as all data potentially recorded within the patient record, including billing and administrative data, is variable,[1-3] and its fitness for use depends on context.[4,5] Attempts to create a systematic methodology for EHR data quality assessment must consider these issues. Data quality assessment is dependent upon and can be operationalized according to the dimensionality of the data available and the data quality requirements of the intended study.

EHR data quality affects the validity and reproducibility of research results. Data of poor and variable quality can lead to excessive noise in datasets, spurious outcomes, and erroneous cohort selection.[4,6-9] This is especially disappointing in light of the fact that a potential advantage of the reuse of EHR data is improved generalizability due to the fact that the subjects are more representative of real world clinical populations than those in traditional research.[10] There is a growing literature on EHR data quality assessment,[11-16] but the majority of published studies relying upon EHR data either do not mention data quality, or report ad hoc methods that are not supported by evidence nor expert knowledge.[17] While a number of data quality assessment methods have been described in the informatics literature, it may be difficult for researchers without expertise in this area to determine which assessments are appropriate, given the data available and the research questions of interest.

An ideal approach to EHR data quality assessment, therefore, must be both systematic and flexible enough to accommodate the needs of different datasets and study designs. Moreover, it should be usable by researchers. Substantial research on clinical guidelines, which are also intended to simplify and standardize complex processes, have been shown to improve the consistency and quality of care.[18] We believe that a data quality assessment guideline would do the same to promote best practices for the reuse of EHR data by improving the validity and comparability of datasets, and the transparency and consistency for data quality reporting.

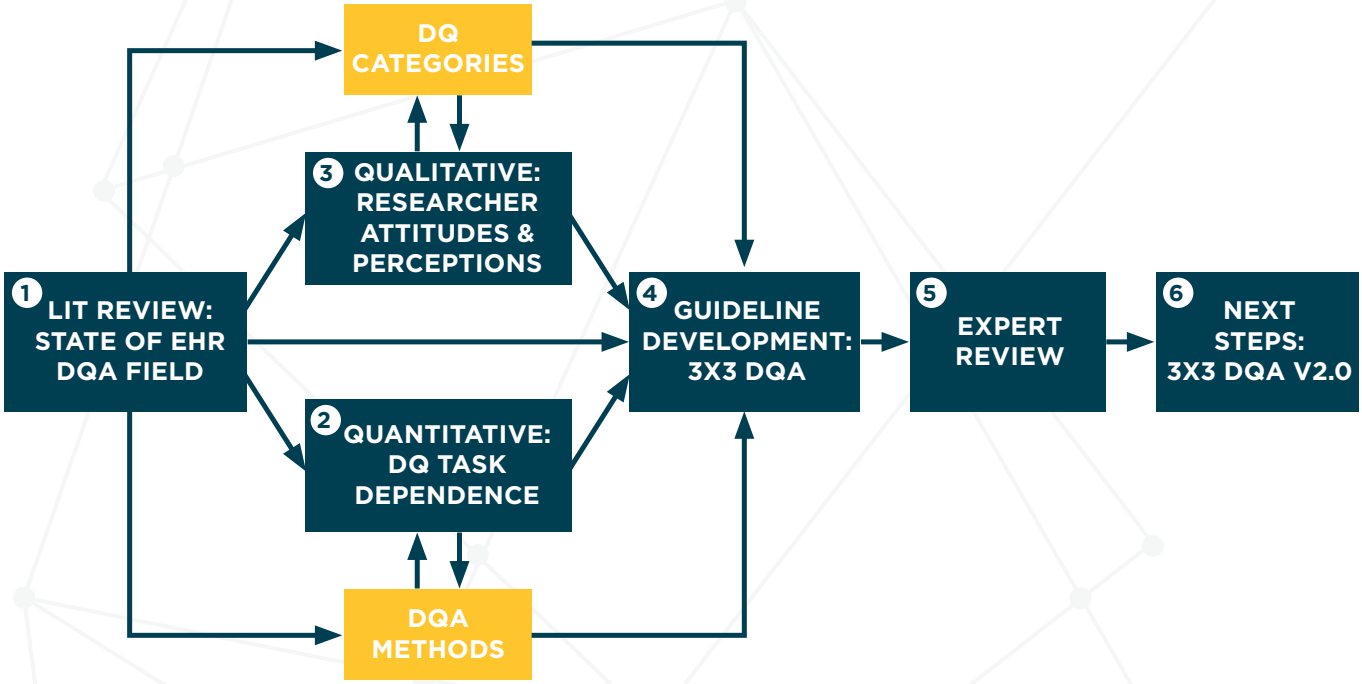
In this paper, we describe the formulation, development, and initial expert review of 3x3 Data Quality Assessment (DQA), a task-dependent, guideline-based approach to EHR data quality assessment. Our goal was to create a useful and usable guideline based upon a set of coherent and consistent guiding principles, featuring an accessible conceptual view of EHR data quality, with concrete and actionable recommendations for data quality assessment, in order to improve the overall validity and reproducibility of research conducted through the reuse of EHR data.

Methods

The initial version of 3x3 DQA is the cumulative product of three studies: a review of the literature on EHR data quality assessment,[17] a quantitative study of EHR data completeness,[4] and a set of interviews with ten clinical researchers. The overall development process is summarized in Figure 1.

The data quality constructs addressed, the methodology described, and the manner in which the entire process of EHR data quality assessment was approached, were determined through the synthesis of our previous research on EHR data quality and context of reuse.

Figure 1. Flowchart Summarizing Development of the EHR Data Quality Assessment Guideline



Literature Review: State of the EHR Data Quality Assessment Field

A systematic review of the informatics literature was used to identify the most common constructs of data quality and methods of data quality assessment. The details of this study can be found in Weiskopf and Weng, 2013.[17] Our inclusion criteria were that the papers must describe methods of data quality assessment, that they focus on EHR-derived data, and that they were peer-reviewed. A PubMed search was used to identify an initial pool of relevant papers, which were then used as the basis for an ancestor search of their references. From each paper we extracted: 1) the constructs of data quality assessed and 2) the methods of assessment used. An iterative process was used to consolidate the extracted data into fundamental constructs. We quantified the strength of the relationships between data quality constructs and assessment methods by the frequency, as indicated by the publication count, with which each method was used for each construct. These frequencies were used to select and prioritize methodological approaches to assessing each data quality construct during the guideline development.

Quantitative: Data Quality Task-Dependence

A quantitative approach was used to explore the concept of task-dependence (fitness for use), specifically focusing on the data quality construct of completeness as an exemplar. We hypothesized that EHR data completeness can be defined in multiple ways, depending upon intended use, and that, in turn, efforts to calculate rates of records completeness would vary based upon these different definitions and uses. The details of this study can be found in Weiskopf et al., 2013.[4] We proposed four definitions of completeness. A record could be complete if: all expected data are documented,

a sufficient breadth of data elements is available, a sufficient depth of data is available over time, or if the data present are sufficient to predict some clinical phenomena of interest. We adapted and developed methods to assess the completeness of the records stored in the Columbia University Medical Center's clinical data warehouse in accordance with each of these four definitions and then analyzed the differences and overlap in the results. This approach to the relationship between task dependence, data dimensionality, and data quality assessment methods became one of the core features of 3x3 DQA.

Qualitative Interviews: Researcher Attitudes and Perceptions

We used a series of semi-structured interviews to explore views on EHR data quality and EHR data reuse among clinical researchers in a single setting (Columbia University Medical Center). Participants were purposively selected[19] to represent various dimensions of experience and demographics: seniority, gender, clinical training, research experience, and experience using EHR data in research. We set an initial goal of ten semi-structured interviews, with the understanding that further interviews would be conducted if saturation was not achieved.[20,21] This research was guided by Phase 3 of the Precede-Proceed model, which focuses on potentially modifiable factors (predisposing, enabling, and reinforcing) that may influence a target behavior (reuse of EHR data, in our case).[22] The interview transcripts were analyzed using content analysis, with a combined top-down and bottom-up approach.[23,24]

Knowledge Synthesis and Guideline Development

The development of the data quality assessment guideline involved synthesizing the results from the three formative studies described above. The conceptual development process was broken into three primary steps. First, we selected a core set of data quality constructs to include based on findings from the literature review and interviews. Second, we mapped methods of data quality assessment to the data quality constructs for which they are appropriate. Finally, we mapped the data quality assessment methods to the dataset and study requirements under which they were feasible. Based on the constructs, methods, and context-specific mapping of methods to study types and data, we created both a guideline document and a tool to guide researchers with questions through the guideline based on their problem. That is, we described the data quality method and then provided self-reflection questions and metrics for the researcher to assess whether and how data quality issues are likely to affect the proposed study.

Table 1 highlights the constructs identified, the sources, and the categories. Five constructs of data quality were originally identified in the literature review: completeness, correctness, concordance, plausibility, and currency. A partially overlapping set of seven constructs were derived from the interviews: completeness, correctness, concordance, granularity, fragmentation, signal-to-noise, and structuredness. Each of these nine constructs was mapped to the appropriate construct category. These categories are based on the data quality model proposed by Wang and Strong. They define *intrinsic* data quality as independent from the intended use, *contextual* data quality as dependent upon the task at hand, *representational* data quality as how the data are formatted and presented, and *accessibility* as how feasible it is for users to extract the data of interest.[25] We decided to limit the scope of the guidelines to intrinsic and contextual data quality, which focus on assessments of the data themselves. Issues relating to data representation and accessibility, while important considerations in EHR data reuse, are generally difficult to assess

Table 1. EHR Data Quality Constructs, Definitions, Counter Examples, Sources, Categories,[25] and Inclusion Status in Final Guideline

CONSTRUCT	DEFINITION (EXAMPLE VIOLATION)	SOURCE	PROXY FOR	CATEGORY	INCLUDED
Concordance	There is agreement between data elements. *(Diagnosis of diabetes, but all A1C results are normal.)*	literature, interviews	correctness	intrinsic	
Correctness	A value is true. *(Diagnosis of diabetes when patient does not have diabetes.)*	literature, interviews		intrinsic	X
Plausibility	A value "makes sense" based on external knowledge. *(A glucose value of 48.0 mmol/l.)*	literature	correctness	intrinsic	
Completeness	A truth about a patient is present. *(An A1C test was performed, but the result is not present.)*	literature, interviews		contextual	X
Currency	A value is representative of the clinically relevant time. *(The most recent A1C value is from more than 12 months prior.)*	literature		contextual	X
Granularity	A data value is neither too specific nor too broad. *(Diagnosis of diabetes mellitus with unspecified complications, which may not provide sufficient information.)*	interviews		contextual	
Fragmentation	A concept is recorded in one place in the record. *(Patient self-reported diabetes symptoms are in multiple narrative notes, scanned questionnaire, and in structured fields.)*	interviews		representational	
Signal-to-noise	Information of interest can be distinguished from irrelevant data in the record. *(Extent of narrative notes has resulted in substantial data that may obscure diabetes information.)*	interviews		contextual	
Structuredness	Data are recorded in a format that enables reliable extraction. *(Glucose values from external provider are in scanned document or narrative note.)*	interviews		representational	

following the process of extracting the data from the original system and generating the final dataset required by the user. We had also previously determined that two of the nine constructs, concordance and plausibility, were methodological proxies for correctness, and not in and of themselves essential constructs of data quality. Finally, we made the decision not to include granularity or signal-to-noise, due to sparse literature addressing these constructs. This left three core constructs of data quality: correctness, completeness, and currency.

In the quantitative study we demonstrated that EHR data completeness varied depending upon how we defined and operationalized complete data. Two of the operationalized constructs were especially relevant for clinical research: sufficient breadth of variables and sufficient depth of data over time. Time and variables make up two of the three major dimensions of EHR data, the third being patients. The majority of clinical studies include multiple patients and multiple variables. Many also include multiple time points and EHR data are, by nature, longitudinal. A full set of data quality assessment methods must address not only all three core constructs of data quality, as defined above, but all three dimensions of the data as well. Just as the construct of completeness can be operationalized across variables and across time, so too can it be operationalized across patients. Moreover, because data quality is task-dependent, the selection of assessment methods will be dependent upon the dataset and the intended study. The dimensionality of the data determines which assessments are feasible, and the design of the study determines which assessments are necessary.

The literature review identified seven categories of data quality assessment methods from the informatics literature: comparison to a gold standard, agreement between data elements, presence of data elements, agreement between different sources of data (e.g., administrative and clinical), comparisons of multiple data distributions (e.g., distributions of a specific laboratory result or rates of a diagnosis by age at two or more sites), checks of validity/plausibility, and review of log data. In the literature review we had mapped each assessment method to one or more data quality constructs and captured the strength of the relationship based on frequency of occurrence in the literature. To avoid reliance on alternative patient-level sources of information, we did not include gold standard comparisons or agreement between different sources of data. As mentioned above, it was also determined in the literature review that concordance and plausibility were proxies for correctness, so for the purposes of 3x3 DQA we combined the mappings of methods to assess plausibility, concordance, and correctness.

Following these steps, each of the five remaining categories of data quality assessment methods was mapped to one or two of the three core data quality constructs, as summarized in Table 2. We then determined the data dimensionality for which each methodological approach was possible. For example, one can check for data element agreement across different variables or across repeated measurements across time for the same variable, but not across patients. Distribution comparison, in contrast, requires compiling values for a single variable across multiple patients. In order to help users navigate the guideline and identify the appropriate data quality assessments, we also created a set of questions to determine the scope and requirements of their dataset and research question.

Expert Review

We invited ten experts on EHR data quality to evaluate the initial version of 3x3 DQA. The content experts were asked to consider the clarity, validity, comprehensiveness, feasibility, and usefulness of the different sections of the guideline.[26] A forced

Table 2. Mapping Method Categories to Data Quality Constructs and Data Dimensionality

METHOD CATEGORY	DEFINITION	PRIMARY CONSTRUCT	SECONDARY CONSTRUCT	PATIENTS	VARIABLES	TIME
Data element agreement	The values of two or more data elements are concordant	Correct	Complete		X	X
Element presence	Desired or expected data elements are present	Complete		X	X	X
Log review	Metadata (timestamps, edits, etc.) are used to determine quality	Current	Correct	X	X	X
Distribution comparison	Aggregated data are compared to external sources of information on the clinical concepts of interest	Correct	Complete	X		
Validity check	The face validity of values and changes in values is assessed	Correct	Complete	X		X

binary response option was used in order to enable rapid completion of the evaluation by experts, as well as more straightforward analysis of the results.[27] Each section of the evaluation was accompanied by space for open-ended responses and suggestions from the content experts. Content experts were identified through their participation in one of two collaborative efforts focused on EHR data quality definition and assessment, or through the publication of at least two peer-reviewed articles addressing EHR data quality assessment.

Results

We present the results of this work in three sections below. First, we provide desiderata for EHR data quality assessment, which were derived from the literature review, qualitative study, and quantitative study described above. Next, we provide an overview of 3x3 DQA and its components, including the data quality constructs selected for inclusion in the guideline (the complete guideline can be found in Appendix B). Finally, we present the results from the expert review of the initial version of the guideline.

Desiderata for a Data Quality Assessment Guideline

Based on knowledge gaps identified in the literature review, lessons learned while assessing completeness in the quantitative study, and experiences and concerns relayed by clinical researchers in the interview study, we determined that an ideal EHR data quality assessment guideline must be:

- *Systematic and evidence- or expert-knowledge based.* This is in contrast to the frequently ad hoc methodology observed in the literature review, which limits validity and comparability of findings.
- *Flexible enough to accommodate the task-dependent nature of data quality.* As dictated by the findings that data quality methodology and results vary depending upon intended use.

- *Engaging of users in assessment and decision making, rather than a black-box process.* One of the primary findings from the interviews was that researchers believe that clinical expertise is necessary for understanding data limitations and interpreting data quality.
- *Independent from the availability of gold-standard data.* As found in the literature review, the most common approach to assessing EHR data quality is through comparison to a gold standard. Gold standards, however, are difficult to construct and rarely available, especially in the case of de-identified datasets.

3x3 DQA Guideline

Framework

The relationships between the three data quality constructs, the three data dimensions, and the categories of methodological approaches are summarized in the 3x3 DQA conceptual framework (Figure 2), which is the heart of the guideline. Each cell operationalizes a data quality dimension across one of the data dimensions. The three core constructs are defined as follows:

- ***Complete*** data are sufficient in quantity for the task at hand.
- ***Correct*** data are free from error.
- ***Current*** data were recorded at the desired relative and absolute time(s).

Scope Identification Questions

The scope identification questions are a series of yes/no questions to be answered by the user about the dimensionality of their data and their study design requirements, with the goal of incorporating the concept of task-dependence into the implementation and application of the guideline. The questions assess which data dimensions are present in the dataset and study as well as which, if any, operationalized constructs of

Figure 2. 3x3 DQA Framework

	A: COMPLETE	B: CORRECT	C: CURRENT
1: PATIENTS	**1A** There are sufficient data points for each patient.	**1B** The distribution of values is plausible across patients.	**1C** All data were recorded during the timeframe of interest.
2: VARIABLES	**2A** There are sufficient data points for each variable.	**2B** There is concordance between variables.	**2C** Variables were recorded in the desired order.
3: TIME	**3A** There are sufficient data points for each time.	**3B** The progression of data over time is plausible.	**3C** Data were recorded with the desired regularity over time.

Note: Data quality constructs are at the top, data dimensions along the side, and cells contain corresponding operationalized constructs.

currency are relevant. Upon completing the scope identification questions, the user should know which operationalized constructs and, by extension, which recommendations are necessary for the intended study.

Recommendations for Assessment and Reporting

Each operationalized construct within the framework has an associated recommendation for assessment and reporting, except for the assessment of correctness across patients (Cell 1B), which has two recommendations. Notably, the recommendations do not address implementation, which is highly dependent upon the data of interest, and often require the user to call upon relevant domain knowledge. The recommendations also do not include "cut-offs," or points at which a dataset would be deemed to be of sufficient quality. Such a determination must be made by the user depending upon their understanding of the data being used, the clinical phenomena being examined, and the methods of analysis utilized.

Complete Data Quality Assessments (1-3A)

Determine if there are sufficient data points available for each patient, for each variable, and for each measurement time. For each patient, calculate how many variables are present, how many measurement times they have data for, and how many overall data points are present (Cell 1A).[4] For each variable, calculate how many patients have that variable present, how many time points at which it is present, and how many data points overall are present (Cell 2A).[28,29] For each measurement time (e.g., before and after intervention), calculate the number of patients with data present, the number of variables with data present, and the overall number of data points present (Cell 3A).

Based on the results from the above calculations, users must determine if the data available are sufficient for their intended purpose. For example, are certain variables sparse enough that the power of a statistical test would be impacted? Are some patients missing variables that are necessary for analysis, or if a required variable is present, are there sufficient instances of that variable to resolve a temporal trend? If there are multiple measurement times, is there an effective "drop out" rate that makes it difficult to compare one time to another?

Correct Data Quality Assessments (1-3B)

Determine if the overall distribution of values for a variable of interest across patients is plausible. Calculate the proportion of patients with values for that variable that are likely to be the result of errors in measurement or recording (as opposed to legitimate clinical outliers). This can be done through hard cut-offs based on clinical knowledge or statistically (e.g., based on standard deviations).[30-32] The user can also calculate aggregate statistics across patients, including mean, median, skewness, etc., and compare these statistics to expert knowledge about the expected distribution or to external sources of data, like a research registry (Cell 1B).[33-35]

Also consider the concordance between different variables (Cell 2B).[36,37] Concordance checks can be conceptualized as if-then rules. E.g., if diagnosis is related to pregnancy, then sex should be "female." The variables and variable values most relevant for a given research question should be prioritized. If the dataset contains longitudinal data it is also possible to compare values of a single variable across time, by assessing the plausibility of value progressions over time (Cell 3B). For example, in a pediatric population height should not decrease from one measurement to the next.[38]

Caution must be taken when interpreting results from assessments of data correctness. The fact that a value is implausible or discordant does not necessarily imply error in measurement or

documentation. Values outside the clinical norm are to be expected in many medical settings. Therefore, the question that should be asked when interpreting correctness assessment results is not, "Do these values seem incorrect?" but rather, "Do these values seem incorrect given the setting and population?" Moreover, some degree of noise in a dataset can be ignored. Random errors have the potential to increase variance in a dataset, thereby obscuring clinical signals of interest, but appropriate statistical approaches can cope with outliers, whether correct or incorrect. Systematic error, on the other hand, like measurements from an improperly calibrated instrument or a tendency to choose the first item from a drop-down list, may introduce spurious findings into a study.

Current Data Quality Assessments (1-3C)

Assessments of currency are generally dependent upon the availability of timestamp metadata for data elements of interest. If a study focuses on a specific relative or absolute period of time (e.g., pediatric injury rates during the summer or hospitalization rates during a disease outbreak), a dataset should first be checked to ensure that all time-dependent variables fall within the window or windows of interest (Cell 1C). Studies that attempt to infer causality may require that certain clinical phenomena be measured or observed in a specific order (Cell 2C). For example, the presence of certain laboratory results prior to a diagnosis or following the prescription of a medication. It is important to note that these two assessments are not dependent upon the presence of a longitudinal dataset or study design. When the data ***are*** longitudinal it may be important to determine if the data were recorded with sufficient frequency and regularity to provide useful information (Cell 3C). If, for example, a patient has had blood pressure values recorded five times in five years, this may at first glance to provide sufficient information regarding that patient's blood pressure status. In a situation where all five of those recordings were from the same year, and the other four years have no current blood pressure recordings, then that information is actually insufficient. Regularity of data can be calculated using an equation proposed by Sperrin et al.[39]

Selections of appropriate currency assessment methods and determinations of sufficient currency are highly dependent upon the intended use case. Some studies will call for the use of all three assessment methods, while others will require none. Different variables and settings can also be considered to have different temporal resolutions. A study of clinical events in an intensive care unit, for example, will have a unique set of currency requirements. Applying and interpreting results related to data currency requires knowledge of the planned study design and relevant clinical area.

Expert review

Six of the ten content experts who were asked to review 3x3 DQA completed and returned the evaluations. The overall response was positive, though all six respondents had critiques and suggestions. The quantitative and qualitative results for the six responses are summarized below. More detailed respondent-level quantitative and qualitative responses are included in Appendix A.

Quantitative Results

Responses to the more theoretical components of the guideline, which include the framework (Figure 2), construct definitions, and operationalized constructs, were mixed. The strongest scores were for the correctness operationalized constructs. The recommendations, which are more concrete, were viewed more favorably. Correctness once again received the most approval, and the current x time recommendation, which focuses on the frequency and regularity of data over time, was viewed as

Table 3. Number of Experts (out of six) Who Agreed that a Given Component of the Guideline was Clear, Comprehensive, Feasible, or Valid

	CLEAR	COMPREHENSIVE	FEASIBLE	VALID
OVERALL FRAMEWORK	**3 (50%)**	**3 (50%)**		
DATA QUALITY CONSTRUCTS				
Complete	4 (67%)			4 (67%)
Correct	3 (50%)			3 (50%)
Current	2 (33%)			4 (67%)
FRAMEWORK OPERATIONALIZED CONSTRUCTS				
Complete x Patients	4 (67%)			4 (67%)
Complete x Variable	4 (67%)			4 (67%)
Complete x Time	3 (50%)			4 (67%)
Correct x Patient	6 (100%)			6 (100%)
Correct x Variable	4 (67%)			5 (83%)
Correct x Time	5 (83%)			5 (83%)
Current x Patients	3 (50%)			4 (67%)
Current x Variable	2 (33%)			4 (67%)
Current x Time	5 (83%)			5 (83%)
all Complete		4 (67%)		
all Correct		3 (50%)		
all Current		3 (50%)		
RECOMMENDATIONS				
Complete x Patients	4 (67%)		4 (67%)	
Complete x Variable	4 (67%)		5 (83%)	
Complete x Time	4 (67%)		6 (100%)	
Correct x Patient	5 (83%)		5 (83%)	
Correct x Variable	5 (83%)		4 (67%)	
Correct x Time	6 (100%)		6 (100%)	
Current x Patients	5 (83%)		5 (83%)	
Current x Variable	4 (67%)		5 (83%)	
Current x Time	4 (67%)		2 (33%)	

the least feasible to implement. Only three of the experts responded to the items related to the scope identification questions, so those results were not included.

Qualitative Results

Based on their comments, none of the respondents had problems with the clarity of the overall framework structure. Four out of the six, however, took issue with at least some of the wording used in the operationalized constructs within the framework cells, and three suggested additional categories or constructs of data quality for inclusion. One suggestion was that there needed to be more information on how the framework and guideline fit into the larger research process, with one expert stating, "Typically, a researcher will evaluate data quality/availability and develop a research design appropriate for the data. One question is whether the data are sufficient/appropriate given the research design. Without understanding how the data will be used it will be difficult to understand the questions or answers." There was concern amongst the experts that it would be difficult for a user to determine how exactly the framework should be applied, and that not all cells would map to all research studies.

Responses to the three construct definitions were mixed. A number of them disliked the emphasis on task-dependence (e.g., "sufficient for the task at hand"), while another said that task-dependence should be emphasized in all three definitions. One suggestion was that if we wanted to emphasize task-dependence, it might be better to use consistent phrasing across all operationalizes (e.g., "sufficient," as in the complete data operationalized constructs, instead of "desired" or "of interest"). Three of the experts thought the phrase "free from error," which defines correct data, was either unclear or not valid, while another picked this definition out as being especially strong: "This is simple—'free from error'—which is good."

There was similar disagreement over the nine framework operationalized constructs. Again, there was concern about the usage of task-dependent terminology and requests for more concreteness. One expert did not like the approach of projecting the three constructs across the three data dimensions ("To me, these are all the same question. I want to know whether my cohort has the right data available for a study, during the study period. It is one question to me, not three"), and suggested the use of clear denominators for each construct instead. The correctness operationalized constructs were the most popular, though one expert criticized the reliance upon external knowledge.

The qualitative responses indicated that the recommendations were the strongest part of the guideline. One of the content experts stated, "These recommendations…are where this comes to life. I'd be hesitant to make it too formal…but there are key recommendations that everyone should do before using a data source." The majority of concerns centered on the feasibility of implementing the recommendation, and the clarity regarding how to interpret and take action upon the results. One expert pointed out that the data required for the correctness and currency assessments (external benchmarks/knowledge and metadata, respectively) might not be available. Other experts highlighted the potential difficulty of translating the recommendations into computerized processes that could be run against electronic datasets. In reference to the completeness recommendations, one expert requested greater clarity clearer regarding how judgments would be made regarding if a dataset met the necessary threshold for quality: "[The recommendations] are about availability. The recommendations say nothing about the judgment that is required to turn availability into sufficiency…"

Finally, only three of the experts directly addressed the questions to identify scope, which were developed to assist users in identifying the relevant data quality assessments based on study design and data availability. One thought the questions were unhelpful, since most research would have the same requirements. Another thought they were a good start, but required further clarity regarding the role of context (i.e., the intended research). The third expert thought the questions would be helpful: "I think the questions will help users understand the framework and the subsequently presented Recommendations."

Discussion

We succeeded in meeting three of our initial four requirements for an EHR data quality assessment guideline (Desiderata for a Data Quality Assessment Guideline). The majority of the data quality constructs and methods of assessment included in 3x3 DQA are all drawn from evidence-based literature, users are heavily involved in the selection of assessment methods and interpretation of data quality results, and none of the assessment methods are reliant upon the availability of gold standard data. The results from the expert review indicate the more concrete aspects of the guideline (i.e., the recommendations for assessment and reporting) were viewed most favorably. There remains disagreement over the precise definitions and verbiage of the construct definitions and operationalized constructs.

While the overall expert response to 3x3 DQA was positive, the quantitative and qualitative results both indicated that there is significant needed to improve the validity, clarity, comprehensiveness, and feasibility of the guideline. Many of the qualitative comments (see Appendix A) indicated that context is key in understanding and utilizing 3x3 DQA; that is, context within the broader research process, the context within which the data were originally collected, and the context of the intended study in determining data quality "sufficiency."

We identified three primary areas for improvement. First, there needs to be increased clarity on how and when 3x3 DQA should be used during the research process. Second, we must improve the feasibility and ease-of-use of implementing the recommendations in order to decrease user burden. And finally, the process by which users determine which operationalized constructs and recommendations are relevant for a given dataset and study must be improved and clarified (this last point was one the fourth original requirements for the guideline).

The lack of clarity on how and when to use 3x3 DQA, as well as confusion surrounding the selection and application of assessments, indicate that a linear document, either paper or electronic, is not the best approach for presenting the information and methodology contained therein. 3x3 DQA is intended to be comprehensive of many study designs and types of data, which means that the guideline may be too long to thoroughly digest in its entirety, and also that much of the guideline will not be relevant for a given research task. A desired approach would dynamically select and present data quality constructs and assessment methods that are relevant for the user's specific research task. This was the original goal of the scope identification questions, but the limited expert response to these items suggests that these questions were largely ignored. An interactive interface that centers these questions and does some of the work of identifying the appropriate portions of 3x3 DQA would be ideal.

The potential burden of implementing the methods described in the guideline recommendations also suggests that a linear guideline is not sufficient for truly improving the feasibility of data quality assessment. Moreover, it is difficult to explain algorithms or statistical approaches in human-

readable prose, so implementations may differ between users, which runs counter to the goals of transparent and comparable data quality reporting. Assessment methods will still need to be tailored to the data and study design, but at least partial implementation would be hugely helpful for users.

A final point is that there is always a trade-off between expressiveness and tractability.[40] One of the goals of 3x3 DQA was to impose a coherent conceptual framework onto the lengthy and difficult process of EHR data quality assessment. In improving the clarity and usability, however, there was a loss of scope and complexity. Going forward, it is important to determine if the losses in scope and complexity are outweighed by the gains in coherence and clarity, as well as if there are ways to improve the balance.

Limitations and Future Work

The work presented here is still in the early stages of development and evaluation. Further iteration and controlled testing for usability and usefulness are still needed. The next step in the development of 3x3 DQA will be another round of design and improvement, based upon the feedback of the six content experts, followed by a formal, scenario-based (i.e., based on one or more use cases) evaluation of usability and usefulness. While experts are important in establishing the underlying conceptual basis and validity of the guideline, user experiences and feedback are necessary for further development. The usability and usefulness of 3x3 DQA would also be greatly improved through partial or complete automation of the guideline, which is currently in progress. We intend to make the current version and any future computerized versions of 3x3 DQA freely available for use, dissemination, and improvement. Our hope is that real world use will lead to user-driven evaluation and enhancement of the guideline.

We also believe that 3x3 DQA can be extended to other forms of clinical data. Data from registries, claims databases, and health information exchanges, for example, could all benefit from many of the assessments contained in the guideline. Issues like standard utilization, concept mapping, and interoperability would be additional key considerations. Other EHR data use cases also require data quality assessment, like clinical quality measurement or clinical decision aids. These are all areas that we will study going forward.

Finally, the actual impact of 3x3 DQA will need to be evaluated. Does it improve the awareness and knowledge of researchers for appropriately selecting and applying pertinent data quality measures? Does it improve the transparency of research and also the interpretability of the research results? Does it improve the validity of research conducted with EHR data? This last question is especially hard to answer, since it requires the establishment of a difficult baseline: what is the validity of research conducted with EHR data without data quality assessment, or with ad hoc approaches to EHR data quality assessment?

Conclusion

3x3 DQA is a guideline targeted at clinical researchers engaged in the reuse of EHR data. It is meant to embrace a fitness-for-use approach to data quality that is flexible enough to accommodate different study designs and data requirements. Rather than relying upon the availability of a reference standard, the 3x3 DQA guides users in utilizing external sources of medical information and knowledge to evaluate data quality. The guideline results from qualitative, data-driven, and literature-based investigation to understand and assess EHR data quality issues.

Based on a review of the validity, clarity, comprehensiveness, and feasibility of 3x3 DQA by

EHR data quality content experts, the guideline appears to be a promising start, though it requires continual development. Specifically, the constructs and operationalized constructs of EHR data quality need to be improved, and primary goals and principles of the guideline to be explicitly stated and explained to intended users. Automated execution of the guideline should also be explored to reduce cognitive overhead for potential users to interpret the complex guideline logic. Further iterations of 3x3 DQA will require extensive testing and evaluation to demonstrate real world usefulness and impact.

Acknowledgements

The authors would like to thank Michael Kahn, Adam Wilcox, and Meredith Nahm Zozus for their guidance throughout the research process; David Dorr for providing invaluable assistance in the editing and formulation of this manuscript; and the EHR data quality experts who provided their time and feedback to review and evaluate the initial version of 3x3 DQA.

Funding

This work was supported by funding from National Library of Medicine grants 5T15LM007079, R01 LM009886, R01 LM010815, and R01 LM006910, as well as National Center for Advancing Translational Sciences grants UL1 TR000040 and UL1 TR000128.

References

1. 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(5):503-27.
2. Hogan WR, Wagner MM. Accuracy of data in computer-based patient records. J Am Med Inform Assoc. 1997;4(5):342-55.
3. Thiru K, Hassey A, Sullivan F. Systematic review of scope and quality of electronic patient record data in primary care. BMJ. 2003;326(7398):1070.
4. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. J Biomed Inform. 2013;46(5):830-6.
5. Juran JM, Gryna FM. Juran's quality control handbook. 4th ed. New York: McGraw-Hill; 1988.
6. Hripcsak G, Knirsch C, Zhou L, Wilcox A, Melton G. Bias associated with mining electronic health records. Journal of biomedical discovery and collaboration. 2011;6:48-52.
7. 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.
8. Hasan S, Padman R. Analyzing the effect of data quality on the accuracy of clinical decision support systems: a computer simulation approach. AMIA Annu Symp Proc. 2006:324-8.
9. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PR, Bernstam EV, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Medical care. 2013;51(8 Suppl 3):S30-7.
10. Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". Lancet. 2005;365(9453):82-93.
11. Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Steiner JF. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. Medical care. 2012;50 Suppl:S21-9.
12. Zozus M, Hammond W, Green B, Kahn M, Richesson R, Rusinkovich S, et al. Assessing Data Quality for Healthcare Systems Data Used in Clinical Research (Version 1.0). NIH Collaboratory. 2014.
13. 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. Appl Clin Inform. 2016;7(1):69-88.
14. Reimer AP, Milinovich A, Madigan EA. Data quality assessment framework to assess electronic medical record data for use in research. Int J Med Inform. 2016;90:40-7.
15. Huser V, DeFalco FJ, Schuemie MJ, Ryan PB, Shang N, Velez M, et al. Multi-site Evaluation of a Data Quality Tool for Patient-Level Clinical Datasets. eGEMs (Generating Evidence & Methods to improve patient outcomes). 2016.
16. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. EGEMS (Wash DC). 2016;4(1):1244.
17. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc. 2013;20(1):144-51.
18. Grimshaw JM, Russell IT. Effect of clinical guidelines on medical practice: a systematic review of rigorous evaluations. Lancet. 1993;342(8883):1317-22.
19. Teddlie C, Yu F. Mixed methods sampling: a typology with examples. Journal of Mixed Methods Research. 2007;1(1):77-100.
20. Guest G, Bunce A, Johnson L. How many interviews are enough? An experiment with data saturation and variability. Field Methods. 2006;18(1):59-82.
21. Francis JJ, Johnston M, Robertson C, Glidewell L, Entwistle V, Eccles MP, et al. What is an adequate sample size? Operationalising data saturation for theory-based interview studies. Psychol Health. 2010;25(10):1229-45.

22. Green L, Kreuter M. Health promotion planning: an educational and ecological approach. 4th ed. Boston, MA: McGraw Hill; 2005.
23. Thomas DR. A general inductive approach for analyzing qualitative evaluation data. American Journal of Evaluation. 2005;27(2):237-46.
24. Fereday J, Muir-Cochrane E. Demonstrating rigor using thematic analysis: a hybrid approach of inductive and deductive coding and theme development. International journal of Qualitative Methods. 2006;5(1):80-92.
25. Wang RY, Strong DM. Beyond accuracy: what data quality means to data consumers. Journal of Management Information Systems. 1996;12(4):5-34.
26. Grant J, Davis L. Selection and use of content experts for instrument development. Research in Nursing & Health. 1997;20(3):269-74.
27. Dolnicar S, Grun B, Leisch F. Quick, simple and reliable: forced binary survey questions. International Journal of Market Research. 2011;53(2):231-52.
28. Pipino LL, Lee YW, Wang RY. Data quality assessment. Communications of the ACM. 2002;45(4):211-8.
29. Soto CM, Kleinman KP, Simon SR. Quality and correlates of medical record documentation in the ambulatory care setting. BMC Health Services Research. 2002;2(1):22-.
30. Basden A, Clark EM. Data integrity in a general practice computer system (Clinics). International Journal of Bio-Medical Computing. 1980;11(6):511-9.
31. Staes CJ, Bennett ST, Evans RS, Narus SP, Huff SM, Sorensen JB. 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(1):12-5.
32. Genes N, Chandra D, Ellis S, Baumlin K. Validating emergency department vital signs using a data quality engine for data warehouse. The open medical informatics journal. 2013;7:34-9.
33. 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(4):243-54.
34. Johnson N, Mant D, Jones L, Randall T. Use of computerised general practice data for population surveillance: comparative study of influenza data. BMJ (Clinical Research Ed). 1991;302(6779):763-5.
35. Iyen-Omofoman B, Hubbard RB, Smith CJ, Sparks E, Bradley E, Bourke A, 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(1):857.
36. 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(1-3):91-8.
37. Brown PJ, Warmington V. Info-tsunami: surviving the storm with data quality probes. Inform Prim Care. 2003;11(4):229-33; discussion 34-7.
38. Maydanchik A. Data Quality Assessment: Technics Publications, LLC; 2007.
39. Sperrin M, Thew S, Weatherall J, Dixon W, Buchan I. Quantifying the longitudinal value of healthcare record collections for pharmacoepidemiology. AMIA Annu Symp Proc. 2011;2011:1318-25.
40. Levesque HJ, Brachman RJ. Expressiveness and tractability in knowledge representation and reasoning. Computational Intelligence. 1987;3(1):78-93.

Appendix A. Quantitative and Qualitative Responses

	CLEAR	COMP.	VALID	FEASIBLE	SAMPLE STATEMENTS FROM EXPERT
Framework	Y	Y			Typically, a researcher will evaluate data quality/availability and develop a research design appropriate for the data. Recording a diagnosis code for "diabetes" in an EHR during a visit in which the clinician orders a test for diabetes is not necessarily an error, it might be a local policy that all diabetes tests ordered get coded with that diagnosis. I want to know whether my cohort has the right data available for a study, during the study period. It is one question for me, not three. These recommendations…are where this comes to life.
Constructs	0/3		0/3		
Cells	4/9	1/3	3/9		
Recs.	8/9			8/9	
Framework	N	N			I think missing from the framework is actually the frame—when in the research process are we supposed to use this? It seems to be aimed at the analysis of a data set—after the data-collection process has been specified. [T]he concept missing for me is my data-quality workflow, as a quality assessor or researcher. [Regarding] the "pathway" of data from physical event to recording in the dataset. These three operationalizations don't cover all of them, so I presume you are making a choice based on some sort of tradeoff, having to do with ease of checking.
Constructs	0/3		0/3		
Cells	1/9	0/3	6/9		
Recs.	5/9			6/9	
Framework	Y	N			Again it is contextual—fitness for purpose definition. But overall the logic of self-assessment and self-determination of what "sufficient" is makes sense. Progression on data over time reflects clinical course and will vary depending on a number of diagnostic, management and prognostic factors. So need constraints in framing the research question(s). I think the realist approach should be emphasized, i.e. the importance of context. The issue of "actors" is another important scope question as EHR-based research can be used for research about the care provider and interventions as well as impact on patients.
Constructs	3/3		2/3		
Cells	6/9	0/3	6/9		
Recs.	7/9			9/9	
Framework	Y	N			I don't see anything to address the quality issue of, "Are the right patients included in the data?" Perhaps this is more of a research question…but it seems to cross into the data quality boundary when someone attempts to use the data for something that's not fit for purpose. [I]t feels as if [completeness] depends on the 'task at hand.' If the goal is to estimate an effect, then completeness requires that the estimate can be generated without bias due to confounding. [Using an] external reference is a good idea, but practically is quite difficult, both in terms of logistics and methodologically ensuring that the external reference should be comparable to the source population.
Constructs	0/3		2/3		
Cells	9/9	3/3	9/9		
Recs.	4/9			3/9	

	CLEAR	COMP.	VALID	FEASIBLE	SAMPLE STATEMENTS FROM EXPERT
Framework	N	Y			Each construct seems like it should be followed by the term "for the task at hand." [The completeness across patients recommendation] may be difficult for larger datasets, composite variables, and deciding when to do this… where does this get represented? [For the current across patients recommendation,] this is clear—I'm not sure how feasible it is. [For the current across time recommendation,] without metadata for recording data I'm not sure how feasible this is.
Constructs	2/3		3/3		
Cells	7/9	3/3	8/9		
Recs.	8/9			7/9	
Framework	Y	Y			I found the questions to be a nice way to frame the framework! I think the questions will help users understand the framework and the subsequently presented Recommendations. [Y]our framework fits into the larger picture of data quality, and into the frameworks created by others. These Guidelines beg the question of how these fit together. What about single-site versus multiple-site EHRs? What is the bigger picture? Does presentation of the final version of your Guidelines call for a short description of context?
Constructs	3/3		3/3		
Cells	9/9	3/3	9/9		
Recs.	9/9			9/9	

Appendix B. 3X3 DQA

Please use the following link to download a PDF of the complete guideline:
http://doi.org/10.13063/egems.1280.s1