Received: 7 September 2020 | Revised: 24 February 2021 | Accepted: 1 March 2021

DOI: 10.1002/lrh2.10264

TECHNICAL REPORT

Learning Health Systems

Developing a systematic approach to assessing data quality in secondary use of clinical data based on intended use

Hanieh Razzaghi[1,2] | Jane Greenberg[2] | L. Charles Bailey[1,3]

[1]Department of Pediatrics and Biomedical and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA

[2]Metadata Research Center, College of Computing and Informatics, Drexel University, Philadelphia, Pennsylvania, USA

[3]Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

Correspondence
Hanieh Razzaghi, Roberts Center for Pediatric Research 11361, 2716 South Street, Philadelphia, PA 19146, USA.
Email: razzaghih@chop.edu

Funding information
Patient-Centered Outcomes Research Institute, Grant/Award Number: RI-CRN-2020-007

Abstract

Introduction: Secondary use of electronic health record (EHR) data for research requires that the data are *fit for use*. Data quality (DQ) frameworks have traditionally focused on structural conformance and completeness of clinical data extracted from source systems. In this paper, we propose a framework for evaluating *semantic* DQ that will allow researchers to evaluate fitness for use prior to analyses.

Methods: We reviewed current DQ literature, as well as experience from recent multisite network studies, and identified gaps in the literature and current practice. Derived principles were used to construct the conceptual framework with attention to both analytic fitness and informatics practice.

Results: We developed a systematic framework that guides researchers in assessing whether a data source is *fit for use* for their intended study or project. It combines tools for evaluating clinical context with DQ principles, as well as factoring in the characteristics of the data source, in order to develop *semantic* DQ checks.

Conclusions: Our framework provides a systematic process for DQ development. Further work is needed to codify practices and metadata around both structural and semantic data quality.

KEYWORDS
data quality, EHR data, fit-for-use

1 | INTRODUCTION

The secondary use of electronic health record (EHR) data for research has rapidly accelerated in the past decade. Understanding the behavior of these data is important, given their complexity, the variation in processes driving collection, and limited ability in the secondary use conteDoCoxt to tailor data capture for a specific research purpose. As a result, data quality (DQ) assessment is increasingly recognized as a critical component of analysis planning and evaluation.[1-3] While complementary to traditional study design tasks, DQ assessment is distinct in its focus on the operating characteristics of *data*, such as structure, validity, reliability, and completeness, rather than threats to effectiveness of analytic *design*, such as bias, confounding, power, and methodologic assumptions.

Poor DQ in large datasets can lead to spurious cohort construction, misclassification of major variables, and misleading reporting of results.[4-8] Further, the risks to privacy present even in nominally deidentified datasets[9] mean that primary data will not usually be publicly available. Therefore, it is not possible for those who rely on the results to retrospectively examine aspects of DQ that impact whether those results apply to their situation. While ad hoc DQ and "data cleaning" procedures are often described alongside analytic results, a formal, reusable approach to DQ has not been adopted beyond a general concept of *fitness for use*.[10-12]

To address the problem of describing data quality, researchers and other stakeholders have developed standardized terminologies and consensus-driven frameworks to facilitate evaluation and communication[12-27] For example, the harmonized consensus-derived DQ terminology developed by Kahn et al is based on the terms *conformance*, *completeness*, and *plausibility* across the axes *verification* and *validation*. Weiskopf et al[27] propose a 3 x 3 DQA assessment to evaluate DQ along the context-specific constructs *time*, *variables*, and *patients*. Johnson et al[15] developed an ontology based on Weiskopf et al's[27] model to represent dimensions of data quality, namely, *completeness*, *correctness*, *concordance*, *plausibility*, and *currency*. These and other work can be leveraged when designing DQA for a specific dataset or data network. While terminologies provide a general means to describe data quality, they are agnostic about whether data align with clinical expectations.

Separately, numerous tools to evaluate specific aspects of DQ in a research network, data system, or healthcare organization have been published or made available.[21,28-42] For example, Pezoulas et al apply statistical approaches to produce curated datasets accompanied by reports summarizing and visualizing detected problematic fields. Estiri et al have done extensive work in implementing statistical tools in EHR-derived databases. Such tools facilitate characterization of data, but lack principles for integrating information or standardizing processes to determine whether the data are *fit for use*. They also make underlying assumptions about the data (eg, multi- or single- institutional, dataset structure, role of outliers in the data, etc), which limit their potential for adoption in new contexts. As a result of these limitations, important aspects of data accuracy and semantic/clinical heterogeneity can go undetected, leading to results that are less generalizable at best and incorrect at worst. For example, the presence, frequency, and distribution of serum creatinine, a measure of kidney function, in a cohort of patients with chronic kidney disease are expected to be significantly different than for the general ambulatory population. These differences are not due to missingness or plausibility of individual values, and will therefore not be detected by most general-purpose DQ analyses.

Given the limits in currently available tools, DQ assessment frameworks used by clinical research networks or learning health systems focus primarily on structural conformance,[16,34,38,43] and have limited assessment of *semantic* data quality. The PCORnet data network, for example, performs 314 DQ checks on every new refresh of the network data.[38] The PEDSnet DQ program consists of over 800 checks that are catalogued under 15 different check types.[34] These base DQ checks serve the general purpose of examining the structural consistency of the data. As a result, datasets derived from these research networks may contain DQ issues that require late changes to the clinical research study design or analytic plan,[44] thus delaying analyses or limiting power or generalizability. Identifying potential gaps in quality early will allow researchers to modify their study design or analyses to optimize the data available to them to answer their research questions.[24,45]

2 | RESEARCH INTEREST

The objective of this paper is to describe a conceptual framework that addresses gaps in current DQ theory and implementation. This framework provides strategies to evaluate DQ that extend beyond structural conformance and general terminologies. It attempts to tie clinical meaning to the data, and facilitate development of systematic DQ assessment that allows researchers and other stakeholders to understand DQ within their own clinical context.

3 | METHODS

The conceptual framework and accompanying check development process described here were constructed using principles drawn from two sources: current best practice in DQ assessment for clinical data and experience gained from studies spanning clinical research networks. For the former, we evaluated current knowledge by reviewing informatics literature, and for the latter, we compiled lessons learned and methods developed during the conduct of network research in which the PEDSnet Data Coordinating Center played an analytic role. The purpose of the framework is to address gaps in DQ testing specifically in cases where clinical or analytic requirements drive the evaluation. Therefore, we incorporated existing DQ approaches and methods into the framework to leverage existing advances in the field.

3.1 | Literature review

Our assessment of current state started with a survey of published literature. Our initial search strategy using disjoint terms "data" and "quality" in PubMed proved to have a very low specificity, identifying 138 848 reports in 2017 to 2019, many focusing on analytic or process quality. We therefore adopted an adjusted strategy using the compound terms "data quality" and "data characterization" to identify records in this time interval, resulting in 2701 records. We screened abstracts for relevance, with an a priori focus on applicability to clinical data and on scale to large databases or networks, iterating search terms (eg, "clinical data quality" or "clinical research networks and data quality") as necessary. This strategy yielded 160 reports with full text available. The search set was augmented by traversing references of published manuscripts. These were used as substrate for an inductive thematic analysis[46,47]; since our intent was to identify stable themes in DQ literature rather than assess conformance to a previously defined set of concepts, we adopted the reflexive approach described by Braun and Clarke.[48] An initial set of labeled codes was developed by the authors based on detailed examination of 10 publications describing broad-based DQ frameworks.[1,15,16,19,20,27,28,34,42,49,50] These codes were consolidated, and an additional 20 publications were reviewed to augment the initial list until saturation was reached. The two reviewers converged on four themes from the analysis: (a) *ontologies or classification*, which address labeling DQ checks or ideas (eg, *completeness*, *timeliness*, *conformance*, *validity*); (b) *methods and approaches*, which refer to measuring DQ terms with statistical analyses (eg, *frequency distributions*, *minimum covariance determinant*, *Spearman's correlation*) and visualizations (eg, *funnel plots*, *time trends*, *clustering*); (c) *implementation*, which consists of applying DQ terms to research network databases (eg,

PCORnet), or, more broadly, secondary use of EHR data; (d) *breadth or scope of application*, which refers to how broadly the DQ checks are applied. Review of additional publications added specific content within these areas, but did not result in elicitation of additional thematic areas.

3.2 | Network study review

Analysis for study-specific DQA was nucleated with review of analysis plans and results for the PCORnet Antibiotics and Growth Study[51] and the PEDSnet Quality Measurement Study.[52] This evaluation was done in the context of PEDSnet,[53] a learning health system and clinical data research network focused on improving the health of children and adolescents. Review was performed by two authors (H.R., C.B.) of both the structure of the analyses, and the DQ issues encountered during the analyses, as well as the methods developed to address them. Both the narrative themes and the artifacts produced by the studies (eg, visualizations, code for measures) were used to develop principles for DQ planning and measure design. Additional examples drawn from other studies in PEDSnet were used as thought experiments to refine the measure development process.

3.3 | Synthesis

Themes and practices derived from the literature review were harmonized by the authors with concrete examples extracted from case studies, in order to develop a conceptual framework incorporating major constructs driving design of well-founded DQ assessment. The synthesis was completed with the following objectives: (a) Identify gaps in current DQ approaches where clinical context is not integrated; (b) Represent design patterns used by existing tools in the framework; (c) Incorporate experience from current clinical research networks; (d) Develop a framework that allows researchers to systematically approach DQ for their research interests. Therefore, we did not comprehensively catalog existing DQ constructs and applications in the literature, but focused our synthesis on means to achieve our primary objectives.

4 | CONCEPTUAL FRAMEWORK

4.1 | Overview

Figure 1 provides an overview of the conceptual framework we propose as a foundation for addressing *semantic data quality*. The framework consists of two phases: (a) *Semantic DQ Design Principles*, which are theoretical considerations that drive DQ implementation and (b) *Semantic DQ Practice*, which operationalizes semantic DQ principles. Table 1 provides more detailed definitions of each DQ construct in the framework as well as clarifying examples. Essentially, semantic DQ extends both traditional models of DQ and statistical preprocessing by systematically accounting for clinical contexts in which data were collected, the role of specific variables in proposed analyses, and the science of previously published DQ approaches. It balances these three imperatives when examining the operating characteristics of a data network or study dataset. We contend that because the framework provides a systematic approach to DQ development, its use provides a level of assurance that important aspects of DQ assessment have not been overlooked.

The DQ Practice phase of the framework focuses on operationalizing DQ principles. Specifically, it applies methods that

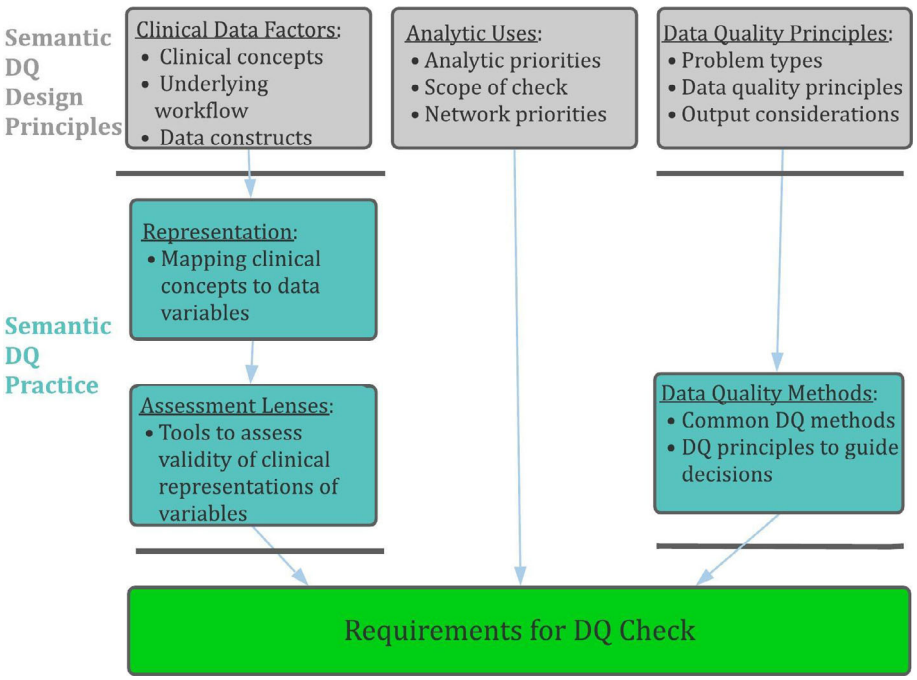

FIGURE 1 *Conceptual model for semantic DQA*, showing the major elements informing the development of semantic DQA checks

TABLE 1 Central elements of semantic DQA conceptual model

Phase	Construct	Definition	Examples
Semantic DQ Design Principles	Clinical Data Factors	• Expresses clinical concept for which data quality (DQ) must be measured • Considers the ways in which underlying workflow affects potential variables • Connects clinical concepts and data provenance	• Hypertension can be measured through diagnoses, medications (prescriptions or administration of antihypertensives), or blood pressure measurements in EHR data
	Analytic Uses	• Weighs the impact of the clinical concept undergoing DQ assessment • Considers the scope: how widely the DQ check will be implemented	• Main exposure variables or outcomes may be more important than minor covariates
	DQ Principles	• Addresses the combination of established DQ theory with current needs • Develops roadmap to determine appropriate DQ method • Focuses the results of variable testing	• Benchmarking hypertension metrics across institutions for face validity requires a different set of tools than attempting to use external sources to test the plausibility of blood pressure values • Common DQ principles include outlier detection, completeness of records, variable concordance, and plausible distribution of facts
Semantic DQ Practice	Representation	• Translates clinical concepts to data-adapted variable definitions	• More precise clinical definitions should be considered—eg, hypertension defined as use of antihypertensives may be important to measure specificity and hypertension defined as a series of blood pressure measurements allows more flexibility in analytic modeling
	Assessment Lenses	• Supplies specific assessments to evaluate the validity of variables	• Common lenses to consider in clinical research are epidemiology, diagnoses, clinical care, and health care utilization.
	DQ Methods	• Applies statistical or descriptive methods to evaluate DQ principles	• Methods can range from simple (eg, proportions or frequency distributions) to complex (eg, PCA, clustering, or other machine learning) • Results can be categorical or can rely on visualization. • Thresholds for acceptable DQ can be pre-determined or part of the applied methodology.

Note: Green elements address development of clinical content for testing, while blue rows address application of DQA testing methods.

consider both clinical validity of variables and analytic approaches to implementing DQ tests. Traversing this phase consists of two steps—the first translates clinical concepts of interest to *representations* in the data. The second then examines their validity through *lenses* of clinical investigation, to define specific targets for assessment. DQ methods apply statistical or descriptive tools to assess these targets, and can range from simple frequency distributions or visualizations to complex machine learning algorithms. Since the framework addresses methods for testing and reporting DQ results in this way, it facilitates better communication of important aspects of DQ along with analytic results.

The remaining sections of this paper describe each of these topics in more detail. To further illustrate the conceptual framework, we will follow a hypothetical use-case throughout, as well as providing specific examples in which the framework has aided DQ development in other PEDSnet studies. We model use of PEDSnet for a descriptive study of infectious outcomes in children with cancer. The study focuses on empiric antibiotic usage in patients with probable or confirmed bacterial infection while undergoing cancer therapy.

4.2 | Clinical data factors, representation, and assessment lenses

The first phase of the framework addresses the clinical meaning that drives data use. The existing literature contains numerous examples of implementing DQ assessments for specific studies. Our proposed framework extends prior work by providing a systematic approach to considering not only standard study-related DQ constructs, but also clinical meaning and representation in the underlying data, yielding more targeted and clinically meaningful DQ evaluations.

At the top of the framework are *clinical data factors*, which include both intended use of the data (ie, analytic requirements) and its provenance. In effect, it reflects the motivation for using the data. When data are collected for a specific analytic purpose, its intended use is clear, and DQA focused on this single purpose is more straightforward. In large networks that serve as a data resource for a wide range of research and QI projects, the clinical needs may vary based on use-case or the priorities of the network. In our cancer study example, we wish to include a cohort of children with cancer actively receiving chemotherapy, which differs from patients in clinical trials, or those seeking second opinions or mid-course shifts in treatment. Additionally, the risk of infection may be higher in sub-cohorts of children. Further, if the study aims to focus on infection during active treatment, the terms *infection* and *active treatment* will need to be defined in relation to the data.

To operationalize *clinical data factors*, researchers must translate analytic constructs to data-adapted clinical definitions, which is reflected in the *representation* step in the conceptual framework. For example, infection can be represented in the data as a diagnosis, medications used to treat infection, procedures testing for infection, symptoms such as fever, evidence for hospitalization, or microbiology results indicating an infectious agent. Once these analytic constructs are defined, the next step is to map them to data domains and specific elements in the data. Pertinent issues to consider in our example study might include: How do custom chemotherapy orders appear in the EHR and how are they represented in the data source under study? What would be an expected gap in the date between diagnosis and treatment? What is the difference between a clinical diagnosis (eg, fever) in an outpatient (encounter diagnosis) or inpatient (discharge diagnosis) setting and how does that relate to timing of infection among children with cancer? In what clinical settings is temperature a more reliable indicator of infection than a fever diagnosis? What other elements of infection (eg, diagnosis, treatment, testing) should be present with fever to reliably represent infection? The mapping of *clinical concept* to *representation* can be complex, and strongly shapes how to measure data quality, which can in turn define which of several alternative DQ checks will prove most effective in the analysis.

Assessing the validity of clinical variable definitions and representation requires a set of tools in order to develop specific DQ checks. The *assessment lens* answers the question: *What specific tests can we use to evaluate the variables and ensure that scientific analyses conducted using the data are valid?* Four common lenses used to answer this question for clinical data are *epidemiology*, *diagnostics*, *clinical care*, and *utilization*, which are described in Table 2. For the current study, examples might include: Are the diagnoses for cancer types that are seen in the database the ones that are typical of pediatric cancer [*epidemiology*]? Are diagnoses of infection accompanied by clinical indicators such as fever or positive cultures [*diagnostics*]? Is the chemotherapy intensity for low- and high-risk patients what would be expected [*clinical care*]? Are patients with cancer undergoing treatment receiving the appropriate ongoing care at the institution [*utilization*]?

TABLE 2 Assessment lens types for health care data

Assessment lens	Goal	Sample tests
Epidemiology	Examine incidence and distribution of analytic variables such as diagnoses, drug exposures, procedure events, etc, to check for validity or internal consistency	• Incidence and population characteristics • Subtype (eg, sub-diagnosis) breakdown and clinical setting characteristics • Variation of prevalence across participating institutions
Diagnostics	Identify and attempt to phenotype patients or medical events based on clinical criteria available in the data	• Major clinical facts available to cross-validate • Impact of variations on cohort definitions • Distributions of test results, vital signs, or other medical events used in diagnosis
Clinical Care	Examine treatments, evolution over time, or expected clinical pathways for patients to identify potential variations or outliers	• Common clinical co-occurrences • Treatment data available / plausible for indication • Sequence of events available / plausible over time
Utilization	Determine pattern of healthcare utilization given a set of clinical characteristics	• Visit type alignment with diagnoses/ comorbidity severity • Facts associated with visit types (eg, ICU with more frequent vital signs) • Variation of utilization across participating institutions

Note: Commonly used lenses for clinical data are described, with examples of specific types of tests each might produce.

Figure 2 illustrates considerations as the framework is applied at each step from *clinical data factors* to *representation* to *assessment lenses* for two aspects of the cancer study: the cohort definition (established cancer patient) and primary outcome (evidence of infection). In this figure, we assume that the cancer cohort will be limited to patients with leukemia, which is the most common type of pediatric cancer, for simplicity of illustration. By describing what data are available, the *clinical data factors* prescribe ways one might define these variables and cohorts. Examining *representations* in turn helps to make definitions precise, and narrows the range of potential questions that DQ checks should test. As a result, the middle boxes provide examples of how these variables are represented in the data. For example,

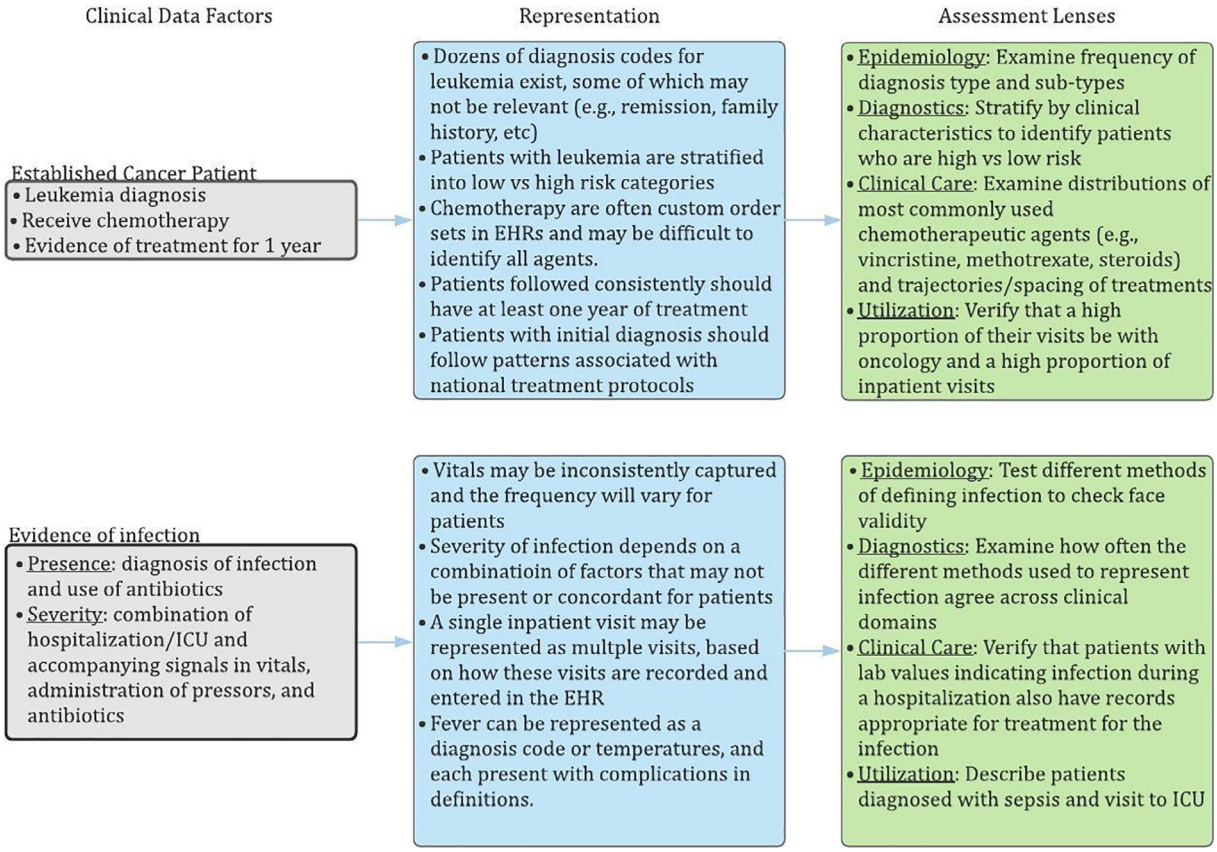

FIGURE 2 *Example semantic DQA specification process*, demonstrating the phases by which research hypotheses associated with the example chemotherapy/infection study are translated to requirements for semantic DQA checks

all chemotherapy orders may not be present, so focusing DQ testing on the most important chemotherapeutics may help to identify patterns for treatment. Similarly, infection severity can be defined in numerous ways (eg, test results confirming a pathogen, treatment with pressors, unstable vital signs, clinician-entered diagnosis), and each of these will need to be mapped to data domains, codesets, and definitions. Finally, the boxes on the right offer examples of *assessment lenses*, shaping the clinical question for DQ requirements. For example, multiple approaches for defining infection could be formally crosschecked to ensure that variable definitions are aligned.

4.3 | Analytic uses

Variable importance and research priorities often drive the focus of DQ in research networks as well as study-specific analyses. In learning health systems or clinical data networks, priorities may be set by governing bodies or by the network's focus in practice. For example, the PEDSnet learning health system has conducted several nephrology-related projects, and the network's DQ efforts extended to curating related variables, such as developing reliable urine protein test data, accurate eGFR trajectories, and a validated glomerular disease phenotype.[54] Similarly, study-specific priorities may drive focus on key variables such as those used in cohort definitions and outcomes, rather than on minor covariates. In the cancer study example, the *severe infection* variable may outweigh all others in the DQ testing plan, given that it is a primary outcome, has the greatest clinical impact, and is arguably the most complex to represent in the data.

Decisions about analytic uses also drive the intended scope of testing. In the semantic DQ model, *scope* refers to the range of cases across which the DQ check must reach and its intended reuse. *Study-specific* scope shapes checks to a highly specific context, in which only the study's requirements and variables are considered. In our example, study-specific scope might yield checks limited to chemotherapeutic agents rather than broader medications, or of only infection types significant in this patient population. *Clinical-domain-specific* checks are more generalizable across analyses, but continue to focus on a defined set of clinical concepts. They typically contain a broader range of variables and more extended value sets than study-specific checks. Instead of limited DQ checks relating to infections in the context of chemotherapy, for example, clinical-domain-specific checks may examine infections across other conditions and settings. *General* DQ checks are usually applied to a large data resource where expected usage varies across studies. These checks may overlap with structural DQA, but remain based in broadly useful clinical concepts more than data model specifications. For example, a clinical-domain-restricted check may focus on chemotherapeutic agents, but overall drug utilization patterns may provide insight into the heterogeneity and

TABLE 3 Application of data quality practices

Data quality principle	Sample DQ checks
Numeric outlier detection	• Develop plots of blood pressure and temperature values to detect implausible values *(epidemiology)*
Temporal outlier detection	• Determine proportion of hospitalizations where antibiotic treatment is provided without evidence of abnormal vitals (ie, temperature or blood pressure), stratified by site. *(clinical care)* • Create time trend graphs to detect whether there are any abnormal spikes in hospitalizations for infection, in relation to chemotherapy intensity *(utilization)*
Completeness of records	• Determine proportion of antibiotic drugs appropriately mapped from source systems, stratified by institution and antibiotic name *(diagnostics)* • Create metrics for patients for whom complete chemotherapy data are available: eg, *number of hospitalizations*, *number of antibiotic drugs*, *number of blood pressure measurements*, *number of temperature checks*, *number of ICU visits*, *number of infection diagnoses*, etc, and perform cluster analysis by site, site*year, or site*risk category, to detect site outliers *(clinical care)*
Concordance of facts	• Create Venn Diagram of labs, vitals, diagnoses, and medications to understand how different definitions alter the overlap of the data domains to define infection *(clinical care)*
Plausible distribution of facts	• Proportion of patients with sepsis diagnosis and visit to the ICU in cancer cohort vs a cohort of healthy patients *(utilization)*

Note: Selected DQ targets from the example chemotherapy/infection study are shown, annotated by the assessment lenses that produced them.

characteristics of medication data as a whole. This scope may serve as the floor on which to build more meaningful semantic checks as analyses demand, or to develop broad benchmarks to follow over time.

4.4 | DQ principles and DQ methods

Data quality principles draw from DQ theory to develop a semantic DQ check. The range of possibilities includes often-needed DQ check types like: *outliers* (numeric or temporal), *completeness of records*, *variable concordance*, and *plausible distribution of facts*. Outlier detection is a common statistical approach for detecting questionable values in a dataset. In DQ assessment, this approach attempts to understand how analytic variables can be defined optimally within the parameters of available data, or to explore plausible explanations for variation. For example, extreme values in heights or weights can be explained by measurement or mechanical errors, and may be accounted for by examining where these errors are more likely to occur. In the context of semantic DQA, *completeness of records* is a common DQ problem in clinical data because data may be limited to care provided within a health care system. Therefore, information about care sought outside a health system or access to community resources is not available. For example, a child can receive ADHD medications or neuropsychologic testing within a particular health system, but primary care and social services provided in the community or school performance are not readily available in clinical data. *Variable concordance* measures the degree to which clinical concepts are consistently represented across different parts of the data. For example, hypertension can be expressed through a series of blood pressure measurements, antihypertensive medications, and diagnoses. Discordance can be explained by incomplete patient records, variable coding practices for health conditions, incomplete ETL processes, or sub-optimal codeset or variable development. Finally, *plausible distribution of facts* refers to the

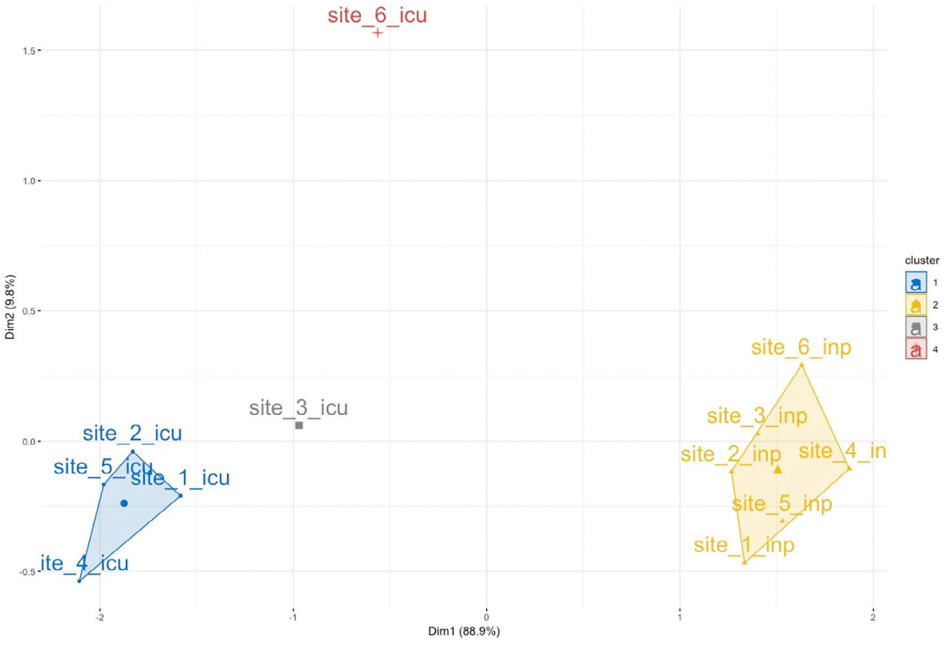

FIGURE 3 *DQ check for outliers in vital sign measurement*. The results of *k*-means clustering over the first two principal components evaluating measurement of four vital signs sensitive to severity of illness are shown. Measurement rates were assessed by site and within site by care setting (ICU or non-ICU inpatient unit)

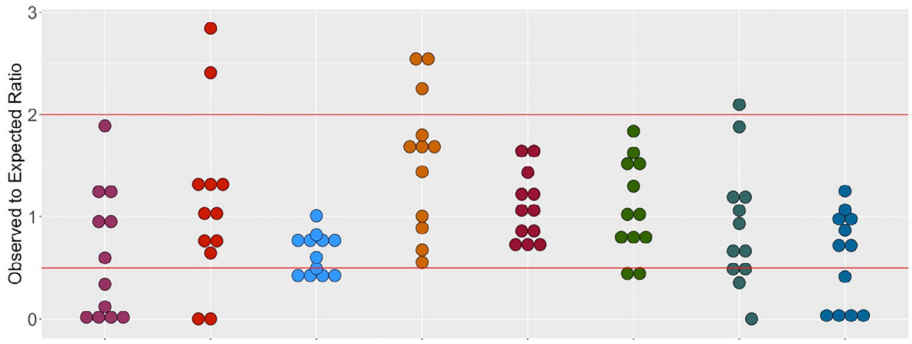

FIGURE 4 *Dotplot of study cohort contribution across PEDSnet institutions.* The observed to expected ratio for multiple PEDSnet studies is shown. Participating PEDSnet institutions are lined across the x-axis and dots represent individual network studies

believability of specific elements within the source data. For example, creatinine measurement frequency and values in a cohort of children diagnosed with chronic kidney disease should differ significantly from a cohort of patients diagnosed with anxiety or depression.

The *data quality methods* phase of our framework formalizes the implementation of *data quality principles* in the creation of DQ checks. The DQ literature includes approaches to visualizing DQ or applying statistical or descriptive methods measuring data quality. These methods vary in complexity and specificity. For example, setting predefined thresholds for acceptable DQ (eg, institutions must have mapped values for >80% of all prescription drugs) requires less intensive evaluation than detecting outliers within the data based on inductive methods (eg, clustering by institutions to track usage of diagnosis codes). DQ methods should align with any expected reporting requirements for the study or larger governing bodies for research networks, so that appropriate decisions (eg, whether data must be remediated or excluded, whether analyses must be adapted) can be made.

Importantly, selection of DQ methods includes consideration of what results are generated as well as what aspect of data is tested. The most effective checks may not produce a pass/fail result; often, summary visualizations of clinical facts are effective tools to quickly identify DQ problems without resource intensive modeling or check development. Similarly, reporting of detailed results using tools such as clustering or flow diagrams, or even databases of DQA test results, may be more valuable to future users of the data than a set of binary statements or an overall "passing grade" for DQA. For example, potential adopters of infection prevention measures resulting from our example study will need to assess whether the underlying antibiotic utilization and data capture align with their own clinical setting, in order to determine whether adopting the measures are likely valuable to them. The ability to assess DQA results in some detail may also help to reduce the ways that bias in any one person's weighting of DQ issues propagates into study design or interpretation of results.

4.5 | Requirements for DQ check construction

The product of the framework is a set of requirements that researchers, data scientists, clinicians, and informaticists can use to create DQ checks. The framework is open-ended in that it does not fully prescribe a checklist of concrete DQ tests to apply, as achieving this degree of specificity for all cases would result in a highly complex process. Nor, as a conceptual framework, does it create the executable code to evaluate the checks. Rather, the purpose of this framework is to systematically approach a known problem that is often addressed on an ad-hoc basis or as a partial preprocessing step for analyses.

Table 3 shows parts of the final output from the conceptual framework for the cancer study, beginning from the point shown in Figure 2. We applied *analytic uses* to focus the examples on cohort definition and the primary outcome variable, infections. The *scope* was limited to study-specific impact. *Data quality principles* and accompanying *methods* were then used as tools to examine the specific questions generated, with the purpose of validating infections in this cohort of patients. For example, the DQ principles "completeness of records" and "concordance of facts" provide means for testing the *clinical care* lens to examine how often the different activities that might be a response to infection agree across clinical domains. In this example, the selected methods are clustering a set of metrics as well as creating a visualization of the overlap.

4.6 | Additional applications of the semantic DQ framework

To further illustrate use of the framework to generate DQ checks beyond typical practice, we provide two examples drawn from work by the PEDSnet DQ team. For a study examining patient acuity, the analytic design compared physiologic state across inpatient and intensive care settings [*analytic use*], with the expectation that increased monitoring is a distinguishing characteristic of ICUs [*clinical data factor*]. While the analysis plan accounts for ascertainment bias due to differential sampling, it also relies on completeness [*data quality principle*] in each setting, to allow for reasonable comparison. To assess this, we selected a subset of vital signs that are sensitive to disease severity: blood pressure, oxygen saturation, temperature, and heart rate, each collected in PEDSnet data [*representation*]. Monitoring should be consistent across all institutions [*assessment lens: utilization*]. We therefore performed outlier detection using clustering [*data quality methods*], with results shown in Figure 3. The check reveals two institutions where the ICU data behave differently, which may point to a DQ problem in ICU vital sign measurement or unit classification. This finding has implications for overall completeness of ICU data at these two sites.

All PEDSnet members provide multispecialty tertiary care for children. On average, then, each institution contributes to cohorts selected by health care data proportionally to its share of the overall PEDSnet population [*clinical data factors*]. The pattern of contribution over time can therefore serve as a measure of overall site capability reflective in the network's research areas of focus [*analytic use*]. We assessed site activity against this expectation [*data quality principle*] by comparing an institution's observed cohort size [*representation*] across multiple studies to its expected size [*assessment lens: epidemiology*], reporting ratios per study as a dotplot [*data quality method*]. Figure 4 demonstrates that two sites consistently have fewer patients in study cohorts than expected, which indicates that patient records may be incomplete across multiple domains at these institutions [*data quality principle*]. This DQ problem points to systemic issues that warrant further investigation and discussion with ETL analysts and site investigators.

5 | CONCLUSION

This paper describes a semantic DQ model that extends beyond structural and general checks to assess *fitness for use*. We describe a systematic process for considering analytic requirements and data provenance, and applying DQA best practices to design checks that interrogate the ability of data to support a particular use. This allows for more transparent assessment of a data source's validity, as well as the generalizability of results obtained from analyzing it. The framework accounts for the need to incorporate clinical context in designing DQ assessment. It also can be applied at multiple levels to assess network-, clinical domain-, or study-specific aspects of DQ in a systematic and reproducible fashion.

ACKNOWLEDGMENTS

This work was funded by the Patient-Centered Outcomes Research Institute (RI-CRN-2020-007). Neither PCORI nor its representatives participated directly in any of the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

CONFLICT OF INTEREST

The authors declare there is no conflict to declare.

REFERENCES

1. Qualls LG et al. Evaluating foundational Data quality in the National Patient-Centered Clinical Research Network (PCORnet[R]). *EGEMS (Wash DC)*. 2018;6:3. https://doi.org/10.5334/egems.199.
2. NESTcc Data Quality Subcommittee. *NESTcc Data Quality Framework*; 2020. https://nestcc.org/nestcc-data-quality-framework/.
3. PCORI Methodology Committee. *PCORI Methodology Standards*; 2019. https://www.pcori.org/research-results/about-our-research/research-methodology/pcori-methodology-standards.
4. Bayley KB, Belnap T, Savitz L, Masica AL, Shah N, Fleming NS. Challenges in using electronic health record data for CER: experience of 4 learning organizations and solutions applied. *Med Care*. 2013;51:S80-S86. https://doi.org/10.1097/MLR.0b013e31829b1d48.
5. Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. *Med Care*. 2013;51:S30-S37. https://doi.org/10.1097/MLR.0b013e31829b1dbd.
6. Ladha KS, Eikermann M. Codifying healthcare – big data and the issue of misclassification. *BMC Anesthesiol*. 2015;15:179. https://doi.org/10.1186/s12871-015-0165-y.
7. Hripcsak G, Knirsch C, Zhou L, Wilcox A, Melton G. Bias associated with mining electronic health records. *J Biomed Discov Collab*. 2011;6:48-52. https://doi.org/10.5210/disco.v6i0.3581.
8. 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 Med Inform Decis Mak*. 2014;14:51. https://doi.org/10.1186/1472-6947-14-51.
9. Cimino JJ. The false security of blind dates: chrononymization's lack of impact on data privacy of laboratory data. *Appl Clin Inform*. 2012;3:392-403. https://doi.org/10.4338/ACI-2012-07-RA-0028.
10. Kerr KA, Norris T, Stockdale R. The strategic management of data quality in healthcare. *Health Informatics J*. 2008;14:259-266. https://doi.org/10.1177/1460458208096555.
11. Mashoufi M, Ayatollahi H, Khorasani-Zavareh D. A review of data quality assessment in emergency medical services. *Open Med Inform J*. 2018;12:19-32. https://doi.org/10.2174/1874431101812010019.
12. Houston L, Probst Y, Martin A. Assessing data quality and the variability of source data verification auditing methods in clinical research settings. *J Biomed Inform*. 2018;83:25-32. https://doi.org/10.1016/j.jbi.2018.05.010.
13. Feder S, Data L. Quality in electronic health records research: quality domains and assessment methods. *West J Nurs Res*. 2018;40:753-766. https://doi.org/10.1177/0193945916689084.
14. Gliklich RE, Leavy MB. Assessing real-world data quality: the application of patient registry quality criteria to real-world data and real-world evidence. *Ther Innov Regul Sci*. 2019;54(2):303-307. https://doi.org/10.1177/2168479019837520.
15. Johnson SG, Speedie S, Simon G, Kumar V, Westra BL. A Data quality ontology for the secondary use of EHR Data. *AMIA Annu Symp Proc*. 2015;2015:1937-1946.
16. Kahn MG et al. A harmonized Data quality assessment terminology and framework for the secondary use of electronic health record Data. *EGEMS (Wash DC)*. 2016;4:1244. https://doi.org/10.13063/2327-9214.1244.
17. Lapchak PA, Zhang JH. Data standardization and quality management. *Transl Stroke Res*. 2018;9:4-8. https://doi.org/10.1007/s12975-017-0531-9.
18. Lee K, Weiskopf N, Pathak J. A framework for data quality assessment in clinical research datasets. *AMIA Annu Symp Proc*. 2017;2017:1080-1089.
19. Pezoulas VC, Kourou KD, Kalatzis F, et al. Medical data quality assessment: on the development of an automated framework for medical data curation. *Comput Biol Med*. 2019;107:270-283. https://doi.org/10.1016/j.compbiomed.2019.03.001.
20. Rogers JR et al. A data element-function conceptual model for data quality checks. *EGEMS (Wash DC)*. 2019;7(1):1-14. https://doi.org/10.5334/egems.289.
21. Sengupta S et al. Data quality assessment and multi-organizational reporting: tools to enhance network knowledge. *EGEMS (Wash DC)*. 2019;7:8. https://doi.org/10.5334/egems.280.
22. Sentinel Operations Center. *Data Quality Review and Characterization*; 2019 https://www.sentinelinitiative.org/sentinel/data-quality-review-and-characterization
23. Shaheen NA, Manezhi B, Thomas A, AlKelya M. Reducing defects in the datasets of clinical research studies: conformance with data quality metrics. *BMC Med Res Methodol*. 2019;19:98. https://doi.org/10.1186/s12874-019-0735-7.

24. Wang Z, Dagtas S, Talburt J, Baghal A, Zozus M. Rule-based data quality assessment and monitoring system in healthcare facilities. *Stud Health Technol Inform*. 2019;257:460-467.
25. Tute E, Wulff A, Marschollek M, Gietzelt M. Clinical information model based data quality checks: theory and example. *Stud Health Technol Inform*. 2019;258:80-84.
26. Charnock V. Electronic healthcare records and data quality. *Health Info Libr J*. 2019;36:91-95. https://doi.org/10.1111/hir.12249.
27. Weiskopf NG, Bakken S, Hripcsak G, Weng C. A data quality assessment guideline for electronic health record data reuse. *EGEMS (Wash DC)*. 2017;5:14. https://doi.org/10.5334/egems.218.
28. Alvarez Sanchez R, Beristain Iraola A, Epelde Unanue G, Carlin P. TAQIH, a tool for tabular data quality assessment and improvement in the context of health data. *Comput Methods Programs Biomed*. 2018;181:104824. https://doi.org/10.1016/j.cmpb.2018.12.029.
29. Browne AN, Pennington JW, Bailey LC. *Promoting Data Quality in a Clinical Data Research Network Using GitHub*; 2015. https://amiajointsummits2015.abstractcentral.com/abstract?FILE_DOWNLOAD_RESPONSE_ID=89967&TAG_ACTION=DOWNLOAD_DETAIL_RESPONSE_FILE&XIK_ABSTR_ID=2091652
30. Davis LM, Zalisk K, Herrera S, Prosnitz D, Coelho H, Yourkavitch J. iCCM data quality: an approach to assessing iCCM reporting systems and data quality in 5 African countries. *J Glob Health*. 2019;9:010805. https://doi.org/10.7189/jogh.09.010805.
31. Fowler J, San Lucas FA, Scheet P. System for quality-assured data analysis: flexible, reproducible scientific workflows. *Genet Epidemiol*. 2019;43:227-237. https://doi.org/10.1002/gepi.22178.
32. Huser V et al. Multisite evaluation of a data quality tool for patient-level clinical data sets. *EGEMS (Wash DC)*. 2016;4:1239. https://doi.org/10.13063/2327-9214.1239.
33. Johnson SG, Pruinelli L, Hoff A, et al. A framework for visualizing Data quality for predictive models and clinical quality measures. *AMIA Jt Summits Transl Sci Proc*. 2019;2019:630-638.
34. Khare R, Utidjian L, Ruth BJ, et al. A longitudinal analysis of data quality in a large pediatric data research network. *J Am Med Inform Assoc*. 2017;24:1072-1079. https://doi.org/10.1093/jamia/ocx033.
35. Pezoulas VC et al. Enhancing medical data quality through data curation: a case study in primary Sjogren's syndrome. *Clin Exp Rheumatol*. 2019;118(3):90-96.
36. Taggart J, Liaw ST, Yu H. Structured data quality reports to improve EHR data quality. *Int J Med Inform*. 2015;84:1094-1098. https://doi.org/10.1016/j.ijmedinf.2015.09.008.
37. Wirsching J, Graßmann S, Eichelmann F, et al. Development and reliability assessment of a new quality appraisal tool for cross-sectional studies using biomarker data (BIOCROSS). *BMC Med Res Methodol*. 2018;18:122. https://doi.org/10.1186/s12874-018-0583-x.
38. Qualls LG, Phillips TA, Hammill BG, et al. Evaluating foundational data quality in the National Patient-Centered Clinical Research Network (PCORnet). *eGEMS*. 2018;6:3.
39. Estiri H, Klann JG, Murphy SN. A clustering approach for detecting implausible observation values in electronic health records data. *BMC Med Inform Decis Mak*. 2019;19:142. https://doi.org/10.1186/s12911-019-0852-6.
40. Estiri H, Klann JG, Weiler SR, et al. A federated EHR network data completeness tracking system. *J Am Med Inform Assoc*. 2019;26:637-645. https://doi.org/10.1093/jamia/ocz014.
41. Estiri H, Murphy SN. Semi-supervised encoding for outlier detection in clinical observation data. *Comput Methods Programs Biomed*. 2019;181:104830. https://doi.org/10.1016/j.cmpb.2019.01.002.
42. Estiri H, Stephens K. DQ(e)-v: a database-agnostic framework for exploring variability in electronic health record Data across time and site location. *EGEMS (Wash DC)*. 2017;5(1):1-16. https://doi.org/10.13063/2327-9214.1277.
43. Jantzen R, Rance B, Katsahian S, Burgun A, Looten V. The need of an open Data quality policy: the case of the "transparency - health" database in the prevention of conflict of interest. *Stud Health Technol Inform*. 2018;247:611-615.
44. Block JP, Bailey LC, Gillman MW, et al. PCORnet antibiotics and childhood growth study: process for cohort creation and cohort description. *Acad Pediatr*. 2018;18:569-576. https://doi.org/10.1016/j.acap.2018.02.008.
45. Terry AL, Stewart M, Cejic S, et al. A basic model for assessing primary health care electronic medical record data quality. *BMC Med Inform Decis Mak*. 2019;19:30. https://doi.org/10.1186/s12911-019-0740-0.
46. Castleberry A, Nolen A. Thematic analysis of qualitative research data: is it as easy as it sounds? *Curr Pharm Teach Learn*. 2018;10:807-815. https://doi.org/10.1016/j.cptl.2018.03.019.
47. Sundler AJ, Lindberg E, Nilsson C, Palmer L. Qualitative thematic analysis based on descriptive phenomenology. *Nurs Open*. 2019;6:733-739. https://doi.org/10.1002/nop2.275.
48. Braun V, Clarke V. Using thematic analysis in psychology. *Qual Res Psychol*. 2006;3:77-101. https://doi.org/10.1191/1478088706qp063oa.
49. Houston ML, Yu AP, Martin DA, Probst DY. Defining and developing a generic framework for monitoring data quality in clinical research. *AMIA Annu Symp Proc*. 2018;2018:1300-1309.
50. Kahn MG et al. Transparent reporting of data quality in distributed data networks. *EGEMS (Wash DC)*. 2015;3:1052. https://doi.org/10.13063/2327-9214.1052.
51. Block JP et al. Early antibiotic exposure and weight outcomes in young children. *Pediatrics*. 2018;142(6):1-12. https://doi.org/10.1542/peds.2018-0290.
52. Hartley DM et al. Use of EHR-based pediatric quality measures: views of health system leaders and parents. *Am J Med Qual*. 2019;35(2):177-185. https://doi.org/10.1177/1062860619850322.
53. Forrest CB, Margolis PA, Bailey LC, et al. PEDSnet: a National Pediatric Learning Health System. *J Am Med Inform Assoc*. 2014;21:602-606. https://doi.org/10.1136/amiajnl-2014-002743.
54. Denburg MR, Razzaghi H, Bailey LC, et al. Using electronic health record data to rapidly identify children with glomerular disease for clinical research. *Journal of the American Society of Nephrology*. 2019;30(12):2427-2435. http://dx.doi.org/10.1681/asn.2019040365.

How to cite this article: Razzaghi H, Greenberg J, Bailey LC. Developing a systematic approach to assessing data quality in secondary use of clinical data based on intended use. *Learn Health Sys*. 2022;6:e10264. https://doi.org/10.1002/lrh2.10264