Research and Applications

A longitudinal analysis of data quality in a large pediatric data research network

Ritu Khare,1,2 Levon Utidjian,1,2 Byron J Ruth,1 Michael G Kahn,3 Evanette Burrows,1,2 Keith Marsolo,4 Nandan Patibandla,5 Hanieh Razzaghi,2 Ryan Colvin,6 Daksha Ranade,7 Melody Kitzmiller,8 Daniel Eckrich,9 and L Charles Bailey1,2,10

1Department of Biomedical and Health Informatics, Children’s Hospital of Philadelphia, Philadelphia, PA, USA, 2Department of Pediatrics, Children’s Hospital of Philadelphia, 3Department of Pediatrics, University of Colorado Denver Anschutz Medical Campus, Aurora, CO, USA, 4University of Cincinnati Department of Pediatrics, Cincinnati Children’s Hospital Medical Center, Cincinnati, OH, USA, 5Information Services Department, Children’s Hospital Boston, Boston, MA, USA, 6Department of Pediatrics, Washington University in St. Louis, St. Louis, MO, USA, 7Research Informatics, Seattle Children’s Research Institute, Seattle, WA, USA, 8Research Information Solutions and Innovation, Nationwide Children’s Hospital, Columbus, OH, USA, 9Center for Pediatric Auditory and Speech Sciences, Nemours Biomedical Research, Wilmington, DE, USA and 10Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

Corresponding Author: Ritu Khare, Department of Biomedical and Health Informatics, Children’s Hospital of Philadelphia, 3535 Market Street, Philadelphia, PA 19104, USA. Phone: +1-215-590-2592. E-mail: KHARER@email.chop.edu

Received 25 October 2016; Revised 18 January 2017; Accepted 16 March 2017

ABSTRACT

Objective: PEDSnet is a clinical data research network (CDRN) that aggregates electronic health record data from multiple children’s hospitals to enable large-scale research. Assessing data quality to ensure suitability for conducting research is a key requirement in PEDSnet. This study presents a range of data quality issues identified over a period of 18 months and interprets them to evaluate the research capacity of PEDSnet.

Materials and Methods: Results were generated by a semiautomated data quality assessment workflow. Two investigators reviewed programmatic data quality issues and conducted discussions with the data partners’ extract-transform-load analysts to determine the cause for each issue.

Results: The results include a longitudinal summary of 2182 data quality issues identified across 9 data submission cycles. The metadata from the most recent cycle includes annotations for 850 issues: most frequent types, including missing data (>300) and outliers (>100); most complex domains, including medications (>160) and lab measurements (>140); and primary causes, including source data characteristics (83%) and extract-transform-load errors (9%).

Discussion: The longitudinal findings demonstrate the network’s evolution from identifying difficulties with aligning the data to a common data model to learning norms in clinical pediatrics and determining research capability.

Conclusion: While data quality is recognized as a critical aspect in establishing and utilizing a CDRN, the findings from data quality assessments are largely unpublished. This paper presents a real-world account of studying and interpreting data quality findings in a pediatric CDRN, and the lessons learned could be used by other CDRNs.

Key words: CDRN, data quality, electronic health record, extract-transform-load, secondary use
BACKGROUND AND SIGNIFICANCE

The Patient-Centered Outcomes Research Institute supports the development of clinical data research networks (CDRNs) as a faster, easier, and less costly infrastructure for clinical research.1 CDRNs transform electronic health record (EHR) data from multiple institutions into common data models and make that data available, in either a centralized or distributed fashion, to conduct a wide range of scientific studies.2–4 Given that EHRs are designed for clinical operations rather than research use,5–7 the results of such scientific studies can be difficult to interpret.5 To this end, one of the most critical aspects in building a CDRN is to ensure that the aggregated clinical data are “high quality” or “ready for research use.”4,5,7 An analysis of the quality of network data serves many purposes. First, it highlights the types of data errors that could be resolved in the next iteration of data submissions, such as an incorrect mapping of patient diagnosis information into controlled vocabularies. Second, it helps in learning the particular characteristics of data, for instance, that “acute respiratory tract infection” and “attention deficit hyperactivity disorder” are likely to be among the most frequent diagnoses in a pediatric dataset, and that this is consistent with expected values. Third, and most important, it helps in mapping the data quality results onto study protocol descriptions, thereby assisting scientists and data consumers with conducting initial assessments of the suitability of the network data for specific research studies.5 For example, a research protocol that studies the effects of antibiotic prescription on hospital revisit rates could not be viably investigated on a network that does not capture prescription data or encounter data in a standardized manner.

OBJECTIVE

This study focuses on the pediatric learning health system CDRN known as PEDSnet.10,11 PEDSnet has transformed EHR data from 8 of the nation’s largest children’s hospitals into a common data format and includes observational data on over 5 million children with at least 1 clinical encounter and at least 1 coded diagnosis during or after 2009. Given its ultimate goal of supporting a wide range of pediatric research and increasing interest by the academic pediatric community in large-scale research networks, a key challenge in PEDSnet is to study and quantify the capability of the data to conduct science and answer research questions. In the past, several studies recommended multiple techniques for conducting data quality assessments on EHR-derived datasets, such as expert judgment, heuristics, knowledge of impossibilities, gold standard benchmarking, code reviews, conformance to value set domains, and computation of derived values.3,5,12 However, the real-world outcomes of data quality assessments on multisite, or even single-site, registries continue to remain behind the scenes and largely undocumented.3–5,13

In this study, we present our empirical experience of studying data quality results over the course of building the digital infrastructure of PEDSnet. The data quality analyses report and interpret a range of data quality “issues,” where an issue is an indication that the data could be inaccurate or difficult to use for some research purposes.3,5,14 We report the longitudinal evolution of PEDSnet data quality assessment over a span of 18 months and a preliminary mapping of the data quality results to 3 different scientific study protocols to provide research usability estimates.

MATERIALS AND METHODS

The PEDSnet data resource has been built iteratively through periodic data cycles managed by the PEDSnet data coordinating center (DCC). Each data cycle is organized under the guidance of a predefined set of PEDSnet-specific network-wide extract-transform-load (ETL) conventions15 and comprises the following steps: (1) each site creates a local dataset by transforming data from clinical source systems into an extension of the Observational Medical Outcomes Partnership (OMOP) common data model (CDM; hereafter referred to as the PEDSnet CDM) according to the ETL conventions,3,16 and either submits the dataset to the DCC or executes network queries from the DCC using a local dataset; (2) the DCC conducts data quality workflow (available at https://github.com/PEDSnet/Data-Quality-Assessment) on these datasets, including application of checks, identification of issues, and holding discussions with the sites to address issues for the next cycle1–5,14; and (3) the aggregated data are transformed into the national Patient-Centered Clinical Research Network (PCORNet) CDM to facilitate Patient-Centered Outcomes Research Institute–based queries.17

Data collection

Figure 1 illustrates the conceptual schema for data quality assessment using Chen’s notation, further elaborated in Supplementary Appendix A.19 A data quality issue is observed in data from a particular site and is the result of applying a data quality check, drawn from a particular check type (see Table 1 for a list of check types) and implemented on a specific data element or field. A data quality issue has the following 4 attributes:

Description: A tailored description of the issue, including the percentage of domain records affected by the issue (or prevalence), as reported by the data quality workflow; eg, the MissData check type identifies 30% of missing data in the person.race_concept_id field.

Priority (high/medium/low): A derived attribute precomputed by the data quality workflow using heuristics involving attributes from other entities, such as check type, data element, likelihood that an element will be used in many studies, issue prevalence, etc.

GitHub issue link: The URL to the GitHub issue that captures the related narrative discussions. The data quality workflow tracks the conversations related to data quality issues using a shared private repository on GitHub.19 For each data quality issue, a corresponding GitHub issue is created to capture the DCC-site interactions. This is a derived attribute; the workflow maintains 1 GitHub issue for all data quality issues resulting from a given check for a given site, irrespective of the data cycle.

Cause: The cause of the issue, determined based on the DCC-site interactions on GitHub. Two investigators reviewed the DCC-site interactions for over 200 issues in a pilot data cycle to determine a classification scheme for causes, including (1) ETL issues that could be resolved, in the next data cycle, by revising the site’s ETL source code; (2) characteristic issues that existed due to the nature of the source data and could not be fixed; (3) non-issues indicating false alarms by the DCC data quality workflow; and (4) i2b2 transformation issues due to bugs in an i2b2-to-PEDSnet CDM transformation process developed to accommodate the partner sites that submitted their datasets in the i2b2 format during early data cycles.20 In each subsequent cycle, the observed issues were coded, using GitHub labels, with the appropriate cause class after reviewing the related comments.

Status: The status of the identified issue in terms of its placement in the data cycle, eg, new, under review, solution proposed, persistent, or withdrawn.

Data quality interpretation

To study the evolution of data quality across data cycles in PEDSnet, we summarized the key attributes of data quality issues and
the number of ETL (or resolvable) issues for most frequently observed check types and data domains using standard descriptive statistics. To present the current state of data quality in PEDSnet, we computed the distribution of check types and domains (both grouped by issue priority [high, medium, low] as computed by the data quality workflow) and causes across issues. In addition, we chose 3 different types of studies as test cases to estimate the data quality impact on those studies. For each study protocol, we identified the PEDSnet elements that were relevant to identifying the study subjects, potential covariates, exposures/interventions, and outcomes. Estimates of data quality impact were computed by quantifying the issue scores for relevant PEDSnet data elements as shown in equation (a), where \( t \) is the study for which the estimates are being computed, \( e_t \) is a PEDSnet data element.

### Table 1. Data quality check types, harmonized terms, and example issues

<table>
<thead>
<tr>
<th>Check type (alias: Name)</th>
<th>Harmonized term</th>
<th>Example issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>InconCohort: Inconsistent cohort</td>
<td>Conformance</td>
<td>Patients in the database who do not conform to the PEDSnet inclusion criteria</td>
</tr>
<tr>
<td>InconSource: Inconsistency with source</td>
<td>Conformance</td>
<td>Distribution of NULL values in race_source_value does not match with the distribution of “No Information” concept in race_concept_id in Person table</td>
</tr>
<tr>
<td>InvalidValue: Value set violations</td>
<td>Conformance</td>
<td>A nonstandard concept used for populating the condition_concept_id field</td>
</tr>
<tr>
<td>UnexFact: Unexpected facts</td>
<td>Plausibility</td>
<td>A medication name entered into the location.zip field</td>
</tr>
<tr>
<td>ImplEvent: Implausible events</td>
<td>Plausibility</td>
<td>Found encounters with visit_start_date occurring after visit_end_date</td>
</tr>
<tr>
<td>NumOutlier: Numerical outlier</td>
<td>Plausibility</td>
<td>A height of 6 cm, or a body weight of 40 000 kg</td>
</tr>
<tr>
<td>ImplDate: Implausible date</td>
<td>Plausibility</td>
<td>Deaths in 1888, encounters in 1930, conditions recorded in 1800</td>
</tr>
<tr>
<td>CatOutlier: Categorical outlier</td>
<td>Plausibility</td>
<td>A patient with over 30 000 procedures</td>
</tr>
<tr>
<td>TempOutlier: Temporal outlier</td>
<td>Plausibility</td>
<td>Peak in the number of measurements on a day in 2012 several-fold higher than on any other day</td>
</tr>
<tr>
<td>ImplDistr: Implausible distributions</td>
<td>Plausibility</td>
<td>Over 10 organisms recorded for a single laboratory culture result</td>
</tr>
<tr>
<td>UnexTop: Unexpected most frequent values from the previous data cycle</td>
<td>Plausibility</td>
<td>“Injection for contraceptive” as the most frequent procedure at a site</td>
</tr>
<tr>
<td>UnexDiff: Unexpected difference from the previous data cycle</td>
<td>Plausibility</td>
<td>Decrease in the number of deaths or large increase (eg, 2 ( \times )) in the number of conditions</td>
</tr>
<tr>
<td>MissData: Missing data</td>
<td>Completeness</td>
<td>Gestational age is not available for 70% of patients</td>
</tr>
<tr>
<td>MissFact: Missing expected facts</td>
<td>Completeness</td>
<td>No serum creatinine record found in measurement table including many serum sodium values</td>
</tr>
<tr>
<td>MissStand: Frequent lack of matching standard concepts</td>
<td>Completeness</td>
<td>Over 50% drug_source_value values could not be mapped to RxNorm</td>
</tr>
</tbody>
</table>
identified to be relevant for a study \( t \), \( i \) is any issue observed for the PEDSnet element \( e \), and the weight parameters are as described in Table 2.

\[
\text{Score}(e_i) = w_{e_i} \times \sum_{i \in \mathcal{I}} w_{C_i} \times w_{P_i}
\]

The element-level weight, \( w_{e_i} \), is a subjective measure determined by 2 data scientists. The issue priority-related weight, \( w_{P_i} \), is automatically computed by DCC’s data quality workflow. The weight ranking for the check type, \( w_{C_i} \), is set using the following rationale. The completeness-related issues are ranked highest, as substantial missingness of elements would directly affect cohort selection impacting the study the most, followed by conformance-related issues that could lead to inaccuracy in the results of scientific studies, and finally plausibility-related issues that could lead to inaccuracy in results (or not). This ranking system also aligns with our process of building PEDSnet, where the initial focus was to get complete data on various elements and ensure that they conformed to the predefined ETL conventions. As the network data gets more complete and consistent, the focus will shift to ensuring and assessing clinical and pediatric plausibility.

**RESULTS**

A total of 2182 data quality issues were collected and processed from 9 data cycles in PEDSnet.

**Evolution of PEDSnet data quality**

Figure 2 shows the distributions of total numbers of data quality checks and issues across data cycles. The number of checks continues to increase because of continuous development of both the CDM and the data quality workflow; eg, a major increase was observed from cycles 3 to 4 when new domains (drug_exposure and measurement) and new types of checks were introduced. Conversely, the workflow did not change significantly from cycles 6 to 7 as the data model was stabilized. The number of issues increases across data cycles, due to both the addition of checks and the adoption of new versions of ETL conventions as data requirements evolve, thereby leading to new learning curves for each site’s ETL analysts. The number of issues decreases when the ETL conventions remain unchanged (eg, cycles 4 to 5, 6 to 7, and 8 to 9), with the exception of cycles 1 to 2, when new sites were added to the network, and 3 to 4, which represented a major development phase.

Figure 3 summarizes the causes of issues across data cycles, demonstrating a clear decrease in the proportion of ETL (fixable) issues over time as the network learned to conform to ETL conventions more effectively, and an increase in the number of characteristic (non-fixable) issues over time as the data scientists uncovered a wider variety of inherent data characteristics. In some cases, fixing an ETL issue (eg, missing data) in 1 cycle resulted in the discovery of new characteristic issues for that data element in the subsequent cycle. E2b2 transformation issues do not appear after the 7th data cycle, because all sites began to directly submit their datasets as PEDSnet CDM extracts.

Figure 4 shows the distribution of ETL issues in the network across key domains and check types. The MissData check type identified the highest number of ETL issues, and the overall trend is improving for each domain. Also, certain domains, such as drug_exposure and measurement, started out with relatively higher numbers of ETL issues due to the complexity of stabilizing conventions and performing ETL operations for these domains. For the InvalidValue check type, the person domain is an outlier because of changing network conventions on the representation of different flavors of NULL (null, others, unknown, unmapped) in PEDSnet during the cycle 4–7 timeline in order to more closely align with PCORnet semantics. Similarly, for the CatOutlier type (ie, abnormally large number of values or obvious peaks in frequency distribution graphs), the graph for the procedure domain reflects changes in the range of source systems from which ETL is being performed. For the UnexDiff check type, the surge at cycle 6 is due to a revision in the definition of the check, wherein in addition to creating issues for the decrease in number of records between cycles, the workflow created issues for an unexpected increase in the number of records.

**Current state of data quality in PEDSnet**

In the most recent (ie, 9th) data cycle, 850 data quality issues were observed, including 119 new issues and 731 characteristic or under-review issues carried over from previous cycles. Table 3 shows the distribution of the prevalence of these issues (ie, the percentage of

<table>
<thead>
<tr>
<th>Equation parameter</th>
<th>Definition</th>
<th>Value (based on a 3-point ordinal scale)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( w_{e_i} )</td>
<td>The importance of the element ( e ) for the given study</td>
<td>Major = 3, Moderate = 2, Minimal = 1</td>
</tr>
<tr>
<td>( w_{C_i} )</td>
<td>The weight associated with the check type associated with issue ( i )</td>
<td>Completeness check = 3, Conformance check = 2, Plausibility check = 1</td>
</tr>
<tr>
<td>( w_{P_i} )</td>
<td>The weight associated with the priority of issue ( i )</td>
<td>High = 3, Medium = 2, Low = 1</td>
</tr>
</tbody>
</table>
records affected) across the associated OMOP domains. The last row represents the check types for which prevalence calculation was not applicable or was unknown, eg, MissFact, UnexDiff, InconSource, and ImplDistri.

Figure 5A shows the distribution of check types; MissData is the most common check type, with >300 occurrences. The other frequent check types (>50 occurrences) include CatOutlier, such as a patient with >2 million procedures recorded since 2009; ImplEvent, such as conditions diagnosed after a patient’s death; ImplDate, such as a measurement_date in 1800s; UnexDiff, such as a decrease in the number of encounters or a 2-fold increase in the number of procedures between cycles; and TempOutlier, such as a sudden increase in the number of facts around 2010.

Figure 5B shows the distribution of domains; drug_exposure and measurement (including vital signs and laboratory data) have the highest number of data issues, as both of these domains are particularly complex, given the variety of medication types and numerous laboratory measurements, requiring extraction from a range of source data records. In the drug_exposure domain, incomplete capture of the extent of exposure (days_supply, effective_drug_dose, stop_reason, and quantity fields) was much more common than drug identity. In the measurement domain, incompleteness was most common in test-related metadata, including range_high, range_low, and measurement_result_date (as distinct from the date the specimen was obtained). The CatOutlier type was very prominent for provider_id, person_id, and visit_occurrence_id fields in both domains, representing certain providers, patients, or visits with a large number of facts associated with them. The TempOutlier was also a frequently observed check type for measurement_date, representing significant peaks around certain dates.

Figure 5C shows the distribution of causes, with nearly 9% ETL issues and 83% characteristic issues. The ETL issues existed largely due to site-level programming errors and to a smaller extent to changing network conventions and other administrative reasons (eg, limited data access rights, dependency on other clinical units for access to data, etc.). Characteristic issues are largely due to limited data capture (eg, certain EHRs or site workflows do not capture the time of birth of a patient), point-of-care data-entry errors or administrative conventions for representing missing data (eg, unknown birthdates purposely documented as born in 1700), and true anomalies (eg, a patient with a >2-year hospital stay). Other reasons for characteristic issues include site-specific ETL decisions (eg, 1 site transiently had a significantly higher procedure-to-patient ratio because it was extracting level-of-service procedure codes for outpatient visits, which other sites had not yet implemented); introduction of a new clinical workflow, such as a sudden increase in documentation procedure (eg, when flowsheets were adopted at that site); and site-specific EHR configurations (eg, certain sites having a more granular range of procedure source values as compared to other sites). Nearly 7% of the issues were false alarms, due to bugs in the data quality workflow or ETL issues in the previous data cycle that were awaiting resolution, and 1% of the issues are still under review by the sites.

Preparedness for conducting science
We selected the first draft of 3 study designs from the set of PEDSnet research pilots as probes to assess the potential readiness of PEDSnet
anomalies, such as administrative workflows, future dates, or prenatal facts (procedure_date, measurement_date, measurement_result_date). It should be noted that a study may require additional elements (eg, radiation dose, intensive care unit admissions, intertransfer hospitals, conditions relevant to procedures) but they are not discretely represented in the CDM, and hence are not shown in the data quality heatmap.

**DISCUSSION**

We conducted an analysis of data quality issues identified during the 18-month startup phase of a large-scale research network comprising 8 large children’s hospitals. Through this work, we identified various potential causes of data quality issues in a CDRN. Prior authors have described 2 major causes of data issues: systematic data errors caused by programming errors, and random data errors (eg, inaccurate data transcription or typing errors).\(^{24}\) The classification of causes provided in this work is more granular (10 precise causes) and motivated by real-world scenarios. As of May 2016, the network contained 850 element-level data quality annotations, with 16% pending or ETL issues and 83% characteristic issues, of which 35% were issues that existed due to limited data capture in EHRs, consistent with prior studies.\(^{25}\)

The longitudinal analysis of causes shows a prominent trend of decreasing ETL (or fixable) issues, even in the face of steadily increasing numbers of data quality checks and an increasing awareness of inherent data issues. This trend strongly illustrates the evolution of the network from expending the most effort on aligning with CDM conventions to increasing focus on norms in clinical pediatrics and research readiness. Although the trend of ETL issues is downward, even at the end of the 9th data cycle, nearly 60 ETL issues were observed in the network, representing the collective responsibility of sites (programming errors) and the PEDSnet DCC (ambiguity in ETL conventions). Brown et al.\(^{2}\) noted that a previously validated dataset does not necessarily guarantee that there will be no new data quality problems in the next revision. A longitudinal analysis of ETL issues helps to identify potential areas to focus on in the next data cycle. For example, incompleteness and drug code standardization continue to drive issues in the drug_exposure domain, and the trend of ETL issues in the procedure_occurrence domain continues to fluctuate.

This work re-emphasizes the complexity of using EHR data for research and the need for thoughtful derivation of study variables rather than simplistic mapping. When planning for a particular study, techniques such as the fitness heatmap could help in identifying specific elements that may not be readily usable, thereby requiring improvements to the overall ETL process (eg, improved mapping of drug concepts from source vocabularies to RxNorm), study-specific data collection, or using statistical imputation methods to overcome those issues.\(^{12, 2}\) Even lower-risk (light-colored) elements require study-specific evaluation as a routine part of analysis, as low risk is not equivalent to perfection in data quality.

The data quality findings of this study are interpreted contemporoaneously with other CDRN-based data quality efforts. For example, the PCORnet Distributed Research Network Operations Center developed a Statistical Analysis System–based data characterization program based on checks developed for the claims-based mini-Sentinel drug safety network. These checks were run against the PCORnet CDM for PEDSnet at the end of the 6th data cycle. A total of 19 data quality issues were identified, including 5 clarification questions. The PEDSnet data quality results included all of these issues, except 1 that occurred due to a programming bug in the DCC’s PEDSnet-to-PCORnet transformation code. Achilles, developed by the Observational Health Data Sciences and Informatics collaborative, is another prominent and widely used data quality

---

**Table 3. Distribution of issue prevalence in the most recent data cycle**

<table>
<thead>
<tr>
<th>Prevalence range (%)</th>
<th>Percentage of issues within that prevalence range (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>15.29</td>
</tr>
<tr>
<td>30–99.99</td>
<td>30.23</td>
</tr>
<tr>
<td>1–29.99</td>
<td>17.18</td>
</tr>
<tr>
<td>&lt;1</td>
<td>30.94</td>
</tr>
<tr>
<td>Not applicable or unknown</td>
<td>6.35</td>
</tr>
</tbody>
</table>

**Figure 5.** Summary of issues for most recent data cycle: (A) distribution of check types, (B) distribution of domains, and (C) distribution of causes. The stacks in Figures 5A and B refer to issue priority as determined by the data quality workflow.

Data for scientific use. The first (S1-abx) is a comparative effectiveness study of narrow- vs broad-spectrum antibiotic use in pediatric pneumonia, the second (S2-rad) is a descriptive study assessing pediatric radiation exposure from computed tomography, and the third (S3-glomer) is a computable phenotype development project for glomerular disease. Figure 6 presents a heatmap denoting the scores of data quality issues observed at various PEDSnet elements for these 3 studies. In general, darker shades denote the elements that are poorly captured at the PEDSnet member sites, such as pn_gestational_age for children first seen years after infancy, while intermediate shades denote date...
After the 9th data cycle in PEDSnet, we catalogued >100 issues for each site, whereas Achilles identified nearly 35 issues per site, including some false alarms against prenatal facts representing real in utero care. One limitation of this study is that it does not include data quality issues arising from other methods downstream of the PEDSnet CDM (eg, during transformation to PCORnet CDM or design of other network queries for science or operations purposes). In addition, heuristics for cause categorization and prioritization continue to be refined as more experience with the data is gained. The results of this study are also incomplete due to limitations of the data quality workflow: it does not yet include validation against external established gold standard datasets, validation of automated phenotype definitions for accuracy, assessment of clinical guidelines (eg, recommended immunization schedules for patients), or assessment of unstructured clinical data.

While this study focuses on network-level or intrinsic data quality assessment, the ultimate goal of data quality assessment in PEDSnet is to help draw conclusions on the quality of aggregated data for research purposes. Although defined from an end user’s perspective, data quality is often accounted for from the perspective of the data producer. Through this work, we provide a broad summary of the current state of data quality of PEDSnet, which could be consumed by any investigator interested in using this network to conduct research. Since PEDSnet has extended the OMOP CDM as an internal data model and the data quality checks are designed accordingly, the lessons learned through this study could be leveraged by other users of OMOP-based CDMs.

**CONCLUSION**

A key challenge in building a CDRN is to understand and interpret the quality of aggregated data for research purposes. Although defined from an end user’s perspective, data quality is often accounted for from the perspective of the data producer. Through this work, we provide a broad summary of the current state of data quality of PEDSnet, which could be consumed by any investigator interested in using this network to conduct research. Since PEDSnet has extended the OMOP CDM as an internal data model and the data quality checks are designed accordingly, the lessons learned through this study could be leveraged by other users of OMOP-based CDMs.

**FUNDING**

Research reported in this work was funded through a Patient-Centered Outcomes Research Institute Award (CDRN-1306-01556).

**COMPETING INTERESTS**

There are no competing interests.
CONTRIBUTORS
RK conducted the computational analysis of data quality issues and prepared the results to be presented in this study. RK and LU conducted the data quality discussions with various sites on GitHub and developed and assigned the category of causes for various issues. BR developed the tools used for managing the warehouse of data quality issues. RK, MGK, HR, and LCB helped in developing the methodology for the study. EB, KM, NP, MGK, RC, DR, MK, and DE are site representatives who actively participated in data quality discussions. LCB is the senior author, who supervised and led the work presented in this study. All co-authors contributed to writing, proofreading, and editing the manuscript.

DISCLAIMER
The statements presented in this work are solely the responsibility of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute or its Board of Governors or Methodology Committee.

SUPPLEMENTARY MATERIAL
Supplementary material is available at Journal of the American Medical Informatics Association online.

ACKNOWLEDGMENTS
The authors would like to thank members of the PEDSnet informatics team, including Eileen Antico, Eric Thum, Gregory Schulte, Janet Zahner, John Newland, Jonathan Bickel, Parth Divatekar, Rajesh Ganta, Rob Brown, Sherry Lassa-Claxton, Steven Griggs, and Victoria Soucek, and the principal investigators for the 3 scientific studies assessed in this work: Dr Katherine Deans, Mark Neuman, and Michelle Denburg. The authors would also like to thank the anonymous reviewers for their insightful comments and suggestions for strengthening the manuscript. Most importantly, the authors wish to thank the patients and families at PEDSnet member institutions (Boston Children's Hospital, Children's Hospital Colorado, Children's Hospital of Philadelphia, Cincinnati Children's Hospital Medical Center, Nationwide Children's Hospital, Nemours Children's Health System, St. Louis Children's Hospital, and Seattle Children's Hospital) who contribute to the PEDSnet data network.

REFERENCES