Journal of the American Medical Informatics Association, 26(7), 2019, 637–645
doi: 10.1093/jamia/ocz014
Advance Access Publication Date: 29 March 2019
Research and Applications

Research and Applications

A federated EHR network data completeness tracking system

Hossein Estiri,[1,2,3] Jeffrey G Klann,[1,2,3] Sarah R Weiler,[3] Ernest Alema-Mensah,[4] R Joseph Applegate,[5] Galina Lozinski,[6] Nandan Patibandla,[7] Kun Wei,[8] William G Adams,[9] Marc D Natter,[10,11,12] Elizabeth O Ofili,[4] Brian Ostasiewski,[8] Alexander Quarshie,[4] Gary E Rosenthal,[13] Elmer V Bernstam,[5,14] Kenneth D Mandl,[10,11,15] and Shawn N Murphy[1,2,3,15,16]

[1]Laboratory of Computer Science, Massachusetts General Hospital, Boston, Massachusetts, USA, [2]Research Information Science and Computing, Partners HealthCare, Charlestown, Massachusetts, USA, [3]Harvard Medical School, Boston, Massachusetts, USA, [4]Morehouse School of Medicine, Atlanta, Georgia, USA, [5]School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, Texas, USA, [6]Boston University School of Medicine/Boston Medical Center, Boston, Massachusetts, USA, [7]Information Services Department, Boston Children's Hospital, Boston, Massachusetts, USA, [8]Wake Forest School of Medicine, Winston-Salem, North Carolina, USA, [9]Department of Pediatrics, Boston University School of Medicine/Boston Medical Center, Boston, Massachusetts, USA, [10]Computational Health Informatics Program, Boston Children's Hospital, Boston, Massachusetts, USA, [11]Department of Pediatrics, Harvard Medical School, Boston, Massachusetts, USA, [12]Program in Pediatric Rheumatology, Department of Pediatrics, Mass General Hospital for Children, Boston, Massachusetts, USA, [13]Department of Internal Medicine, Wake Forest School of Medicine, Winston-Salem, North Carolina, USA, [14]Division of General Internal Medicine, McGovern Medical School, University of Texas Health Science Center at Houston, Houston, Texas, USA, [15]Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA, and [16]Department of Neurology, Massachusetts General Hospital, Boston, Massachusetts, USA

Corresponding Author: Hossein Estiri, PhD, MGH Laboratory of Computer Science, 50 Staniford Street, Suite 750, Boston, MA 02114, USA (hestiri@mgh.harvard.edu)

Received 9 October 2018; Revised 4 January 2019; Editorial Decision 16 January 2019; Accepted 17 January 2019

ABSTRACT

Objective: The study sought to design, pilot, and evaluate a federated data completeness tracking system (CTX) for assessing completeness in research data extracted from electronic health record data across the Accessible Research Commons for Health (ARCH) Clinical Data Research Network.

Materials and Methods: The CTX applies a systems-based approach to design workflow and technology for assessing completeness across distributed electronic health record data repositories participating in a queryable, federated network. The CTX invokes 2 positive feedback loops that utilize open source tools (DQ^e-c and Vue) to integrate technology and human actors in a system geared for increasing capacity and taking action. A pilot implementation of the system involved 6 ARCH partner sites between January 2017 and May 2018.

Results: The ARCH CTX has enabled the network to monitor and, if needed, adjust its data management processes to maintain complete datasets for secondary use. The system allows the network and its partner sites to profile data completeness both at the network and partner site levels. Interactive visualizations presenting the current state of completeness in the context of the entire network as well as changes in completeness across time were valued among the CTX user base.

Discussion: Distributed clinical data networks are complex systems. Top-down approaches that solely rely on technology to report data completeness may be necessary but not sufficient for improving completeness (and quality) of data in large-scale clinical data networks. Improving and maintaining complete (high-quality) data in

such complex environments entails sociotechnical systems that exploit technology and empower human actors to engage in the process of high-quality data curating.

Conclusions: The CTX has increased the network's capacity to rapidly identify data completeness issues and empowered ARCH partner sites to get involved in improving the completeness of respective data in their repositories.

Key words: data completeness, data quality, electronic health records, systems thinking

INTRODUCTION

The adoption of electronic health record (EHR) systems[1–4] has sparked development of distributed clinical data research networks, such as the National Patient-Centered Clinical Research Network (PCORnet).[5] Evaluating data completeness—that is, the presence of "sensical" data[6]—in distributed clinical data research networks is critical to standardize and validate extraction, transformation, and loading (ETL) processes across the network sites.[7] Systematic data completeness assessment efforts may also lead to higher precision in identifying roots for plausibility and conformance issues in EHR data, where plausibility refers to "believability or truthfulness" of data values and conformance is a measure of data representation against internal or external standards.[7] Owing to the multiplicity of stakeholders, resources (eg, technical staff, computational), and regulatory environments in multisite data networks, as well as privacy and data-sharing concerns, designing and implementing a system for federated data completeness assessment remains a hurdle. In addition, existing top-down approaches commonly rely on technology to report data quality (ie, completeness, plausibility, and conformance) issues and expect an analyst to remotely address the reported issues. This approach can be insufficient in improving quality of data in large clinical data research networks. Improving and maintaining completeness in data repositories entails design and implementation of sociotechnical systems (as alternatives to the technology-driven systems in top-down approaches) that exploit technology and empower human actors (ie, staff, leadership, and stakeholders) to engage in the process of high-quality data curating.

The Accessible Research Commons for Health (ARCH)—formerly known as Scalable Collaborative Infrastructure for a Learning Healthcare System[8]—has applied a systems thinking approach to design and implement a federated data completeness tracking system (CTX) to transparently evaluate EHR completeness across its distributed data repositories. The ARCH CTX involves a workflow consisting of 2 positive feedback loops. Each feedback loop is triggered by an open source tool, creating a transparent system that allows the network and its partner sites to profile data completeness both at the network and site levels and take the necessary actions to improve it. This system has enabled the network to monitor and adjust its data management processes to maintain high-quality data for secondary use. In this article, we describe our systems-based approach as well as the CTX workflow and its underlying technologies, and present results from the pilot implementation of the system across 6 ARCH partner sites between January 2017 and May 2018.

BACKGROUND

In light of the national supports for multisite clinical data research networks in recent years, evaluating quality of data extracted from multiple EHRs for secondary use in research has become increasingly important. Efforts in this area can be largely characterized as top-down approaches for data quality assessment in federated networks, in which network partner sites need to satisfy data quality targets defined by a coordinating center. Supported by the Food and Drug Administration, the Sentinel Initiative[9] has created a distributed quality assurance program, which involves an extensive set of data quality checks distributed by the network's coordinating center to the data partners. Sentinel's approach was later adapted by PCORnet. PCORnet's Coordinating Center aimed to evaluate foundational data quality across the Clinical Data Research Networks through the data curation/characterization process. Similar to Sentinel's distributed quality assurance program, PCORnet's data curation/characterization process involved distributing a set of analytic and querying activities to assess data quality across network partners.[10] This process produces a report that is intended to be used at the network partners' sites. The sites are instructed to send the report to the Coordinating Center. However, the report was not developed in collaboration with the sites and is heavily focused on making sure that sites meet PCORnet's guidelines. A number of the Clinical Data Research Networks have established their own data quality assessment processes. The Pediatric Learning Health System applies a similar approach in which a Data Coordinating Center develops and distributes queries that initiate a comprehensive list of data quality checks tailored to pediatric population.[11]

Distributed clinical data networks are complex systems, comprising diverse stakeholders who often have different human and technological resources and regulatory environments. Top-down approaches may not provide sustainable solutions for networks, as their effectiveness is highly dependent on top-down distribution of resources to keep the networks functioning. Systems thinking can offer sustainable solutions that are low cost and built on collaborations. A systems thinking approach, which has been applied to several public health problems, embodies notions from ecological models (often used in public health), system dynamics, and complexity theory[12] to understand and address problems in complex systems.[13] The systems thinking approach provides a holistic perspective to studying how systems function, emphasizing the interactions between its elements in the context of a "greater whole."[12] Therefore, improving the system's functionality (in this case, data completeness) would require improving those interactions within and between its elements.[14] Systems-based approaches can also contribute to harmonizing the language and methodology for conceptualizing and addressing issues within such complex systems.[15]

MATERIALS AND METHODS

We designed, implemented, and tested the CTX system across 6 ARCH partner sites, in a pilot phase. The ARCH CTX system implements a workflow that utilizes 2 open source software packages to assess, compare, and present data completeness. The process is initiated at each partner site. The input at each site is the local dataset, which is fed into DQ^e-c,[6] an open source tool for evaluating completeness in EHR data repositories that generates standard completeness assessment reports. DQ^e-c produces as output a set of com-

pleteness assessment reports. These reports are then used locally to improve the completeness of the site's data, as well as fed as input into Vue, which aggregates the reports from all partner sites. Vue aggregates DQ^e-c outputs and produces a feedback report for each site, which is fed back to the site for completeness improvement, as well as a single dashboard that can be used to present network-level completeness.

The ARCH CTX is different from other federated efforts in that the outputs of its workflows provide visual reports to its partner sites that are modified based on their needs. The reports are intended to encourage individual sites to initiate action(s) to improve data completeness—rather than satisfying a mandate in the top-down approaches. These reports enable sites to evaluate completeness in their respective data repositories, in the context of the network and over time, and make choices to improve their data. In our pilot implementation between January 2017 and May 2018, 6 ARCH partner sites installed and ran DQ^e-c and sent the outputs to the ARCH Data Quality Team (DQT). The process was repeated after each quarterly data refresh across the network.

Workflow and technology

To design the ARCH federated data CTX, we applied a systems thinking approach. The CTX pairs the 2 open source technologies with an interactive workflow to leverage and promote the network's already existing sociotechnical system for improving data completeness across the network. The CTX implements a standardized EHR data completeness assessment tool, collects site-level completeness results, and produces an aggregate report. The main function of the system is to operationalize a federated platform for in-context EHR data completeness assessment (ie, reflecting both cross-sectional and longitudinal states) at the site and network levels.

The workflow in CTX comprises 2 positive feedback loops that can lead to action initiation using outcomes from 2 corresponding tools (Figure 1). Two major actors play role in this workflow; (1) a central DQT that develops and maintains the technology and sustains workflow interactions, and (2) ARCH partner sites (sites) that participate in the system. The 2 feedback loops are identical in their function but differ in the information they use to initiate action. Each feedback loop begins with a completeness assessment tool that generates a completeness assessment report, based on which ARCH partner sites can act to improve completeness in their respective data repositories.

The system's operation begins with development and maintenance of DQ^e-c. The first feedback loop is mostly operated at the ARCH partner sites, making it a federated operation across the network. ARCH partner sites access and install the latest version of the tool in their local environments and can participate in refining DQ^e-c through GitHub. Partner sites run the tool after each data reload that happens quarterly across the network. This process generates DQ^e-c completeness assessment reports. Ideally, sites use these reports to initiate any needed action to improve completeness issues highlighted in DQ^e-c reports. The first feedback loop closes by sharing 2 of the output files with the DQT as the input files for the second feedback loop. The site may also provide technical or usability feedback to improve tools and technologies used throughout the process. Outputs of the first feedback loop are routed back as inputs for the second feedback loop, which is operated centrally by the DQT. Through this process, the DQT collects aggregated individual site reports generated in the first feedback loop to feed a second tool that produces a comparative report for the sites to consider for taking further action on improving data completeness. The implemented workflow is presented in Figure 2.

Through the second feedback loop, the ARCH DQT uses a second tool, Vue, to produce an interactive network-level dashboard and a set of site-level feedback reports. Vue, like DQ^e-c, is an open source R-based tool that generates standard interactive visual reports using outputs from DQ^e-c (Figure 3). Every run of DQ^e-c generates a set of aggregated data, formatted as comma-separated value (CSV) files. These reports provide a visual snapshot of completeness at each site, but not the entire network as a whole. To compile the state of completeness in the network, the DQT designed Vue to aggregate DQ^e-c outputs from individual sites and generates reports that are also used to provide feedback to sites. After accumulating the DQ^e-c report files, Vue recalculates completeness (as represented by percent missingness) by aggregating the raw numbers and uses test dates (DQ^e-c run dates) available in the reports to tag rounds of DQ^e-c runs at each site. Vue uses the accumulated data to visualize the latest completeness status, as well as a longitudinal overview of completeness at both site and network levels, to show variations in completeness metrics over time. Vue generates 2 types of outputs: (1) a series of site-level reports that to be shared with each individual partner site and (2) an interactive dashboard intended to present network-level completeness. Examples of Vue's network- and site-level visualizations are provided in Supplementary Appendix.

The DQT utilizes cloud computing services to store multisite DQ^e-c reports. A local installation of Vue is utilized to generate feedback reports and distribute them to ARCH partner sites. The site-level reports enable ARCH partner sites to assess results of their completeness indices in the context of other partner sites, the entire network, and across time. Because the way DQ^e-c outputs are

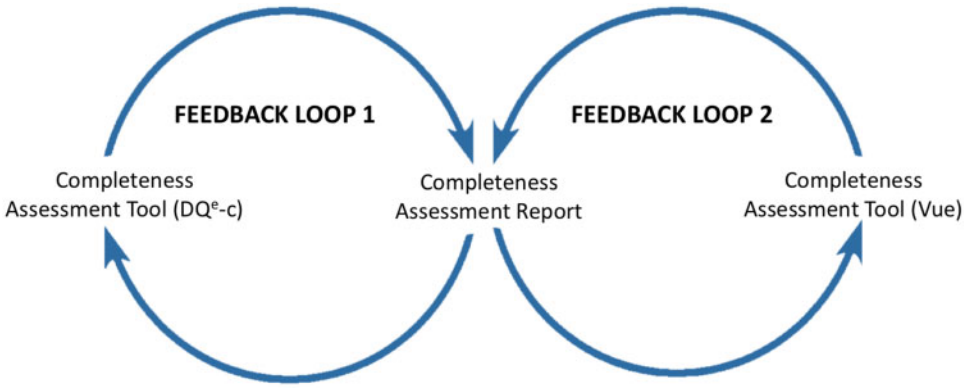

Figure 1. Feedback loops in the Accessible Research Commons for Health completeness tracking system.

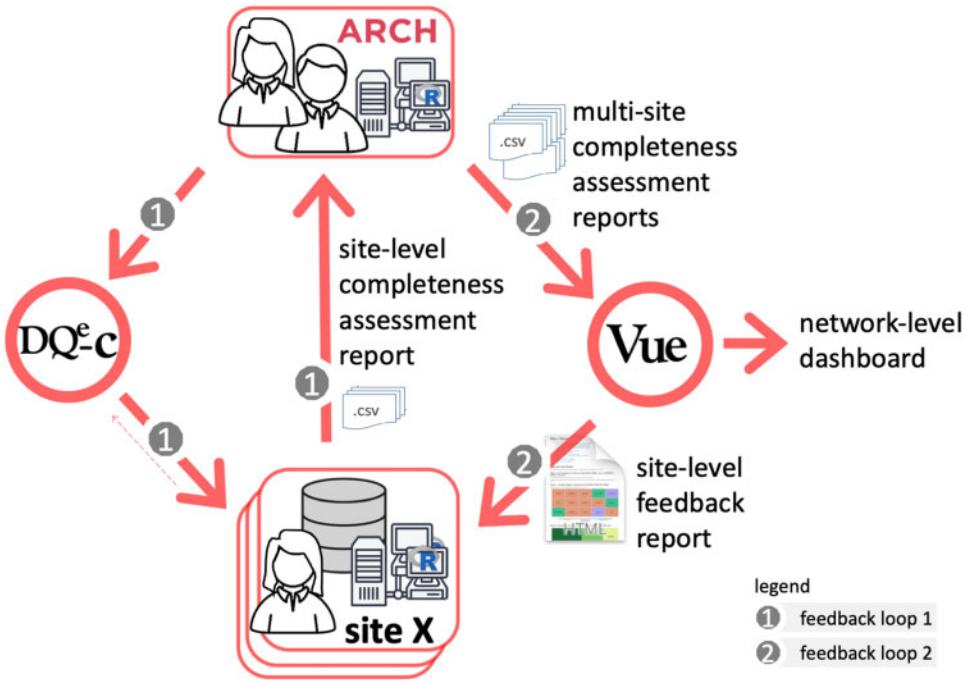

Figure 2. Implementation of the Accessible Research Commons for Health (ARCH) completeness tracking system.

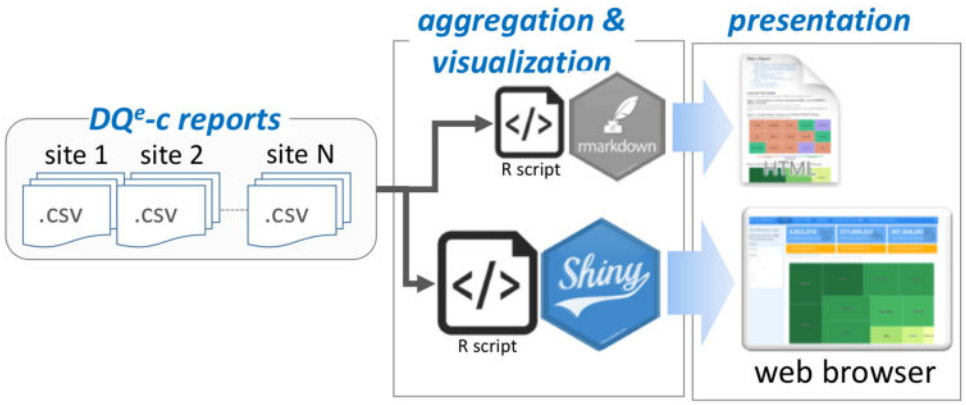

Figure 3. Vue pipeline.

designed, running Vue on site-level DQ^e-c outputs is equivalent to running DQ^e-c on a dataset comprising merging all data in the distributed repositories across the network.

Vue's site-level reports (in .html format) visualize the latest network-level completeness status, as well as variation in site-level completeness metrics over time (ie, by DQ^e-c run date). A standard Vue report can be categorized into 4 sections. First, it provides a preview of unique counts that tabulate the number of unique patients, encounters, diagnoses, and procedures in the network or the respective site. The preview is most useful in the network-level dashboard, where running DQ^e-c is virtually impossible, due to the federated nature of network. Changes in unique counts over data loads are also visualized in the preview. Second, Vue visualizes completeness across the network using tables, allowing individual sites to compare their missingness percentages with the network—using a leave-1-out method. The third section presents a column-level missingness for each table. Finally, Vue recompiles a network-level aggregation of missingness in key indicators, embracing a growing list of indicators such as percent of patients missing records on race or ethnicity, blood pressure, medication, weight, and height. Vue visualizes the latest status of missingness in key indicators and also provides their changes over time.

The information visualized in the 2 Vue outputs (the site feedback reports and the interactive dashboard) are similar in that they both present completeness indices across time and in the context of the entire network. All 4 groups of visualizations presented in Vue feedback reports are also embedded in the interactive network-level Vue dashboard. Deployed on the cloud, Vue's interactive network-level R Studio shiny dashboard (the ARCH-Vue dashboard), is primarily designed for the ARCH network's leadership. The ARCH-Vue dashboard provides an integrated tabular view of the 4 categories of visualizations produced by Vue. The shiny dashboard is password protected and can be accessed by designated users via a secured server. It allows the user(s) to interactively navigate through Vue visualizations for each site and for the entire network, and therefore enables a transparent data completeness evaluation across the entire network.

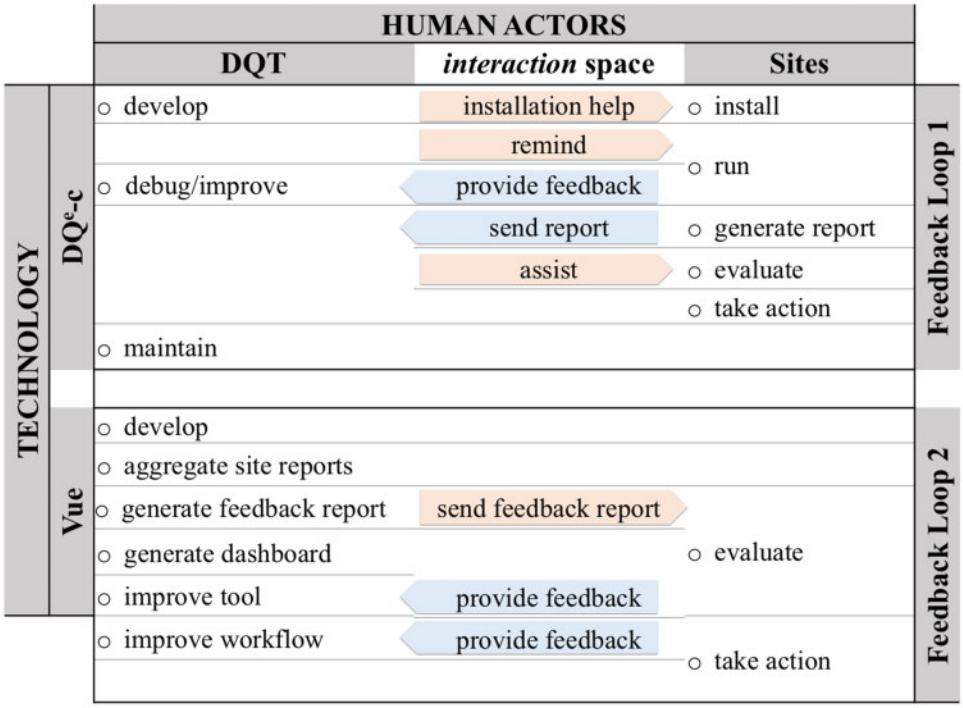

Figure 4. Bidirectional communications between Accessible Research Commons for Health completeness tracking system actors and technology through workflow. DQT: Data Quality Team.

Because Vue directly operates on the data model in DQ^e-c reports, it can be considered an add-on to DQ^e-c. However, outputs from any other tool that can produce reports in a compatible data model can be integrated with Vue output. The tool is publicly available on GitHub—the link will be provided after the peer review process is complete (https://github.com/ARCH-commons/arch-vue).

Evaluating the system

After over a year in the pilot phase and several iterations of the 2 feedback loops, we evaluated the utility of the CTX by conducting (1) a set of semistructured interviews with the participating sites to collect their feedback and recommendations and (2) a test of usability with the network's data curation leadership focusing on the ARCH-Vue dashboard. The semistructured interviews were conducted through a moderated remote process in which the interviewer asked 9 open-ended questions from each site representative who was involved in installing and operating the CTX system. Each interview took between 45 minutes to 1 hour, using screen-sharing software or phone call. The test of usability was conducted through an in-person retrospective think-aloud protocol, in which the ARCH-Vue dashboard user reflected on the usability of each of the visualizations in the dashboard. We used qualitative coding to extract patterns and themes from the interview responses.

RESULTS

Results obtained from design, implementation, and evaluation of the CTX are twofold. First, we discuss what we learned from designing and implementing the system, focusing on bidirectional communications and informing action. Second, we describe the feedback we obtained from the system's users, including feedback from the sites and usability test of the ARCH-Vue dashboard.

Bidirectional communications

An important result from applying a systems thinking approach in design and implementation of CTX was enhancing the bidirectional communication between the human actors in the network (ie, staff, leadership, and stakeholders) through use of technology. Figure 4 presents integration of technology and social actors through CTX workflow to form bidirectional communications between the actors involved in the CTX. Sites provided feedback to improve the tool and the workflow during the early stages of the implementation that focused on tool installation and infrastructures for communications. The DQT developed DQ^e-c, assisted the participating sites in installing the tool, debugged or improved it based on the feedback received from the sites, and maintained DQ^e-c. E-mail communication is the primary means to remind pilot sites to run DQ^e-c, share the outputs, and distribute the feedback reports. After a DQ^e-c run, sites independently evaluated the results and decided whether action was needed to be taken. The DQT provided assistance (if needed) for evaluating DQ^e-c reports.

Through the second feedback loop, the DQT led the tool (Vue) development and generating the aggregate network- and site-level reports. Sites' responsibilities in this loop mainly included evaluating the aggregated reports and, if needed, taking actions. These interactions enhanced the CTX from a technical process into a sociotechnical system. The CTX has enabled the ARCH network to monitor, and where needed, adjust its data transformation to alleviate data missingness issues.

Informing action

To facilitate distribution of PCORnet analytic queries, ARCH has utilized methods and tools to directly transform data in an i2b2 (Informatics for Integrating Biology and the Bedside)[16] warehouse into physical database[17] structured with the PCORnet Common Data

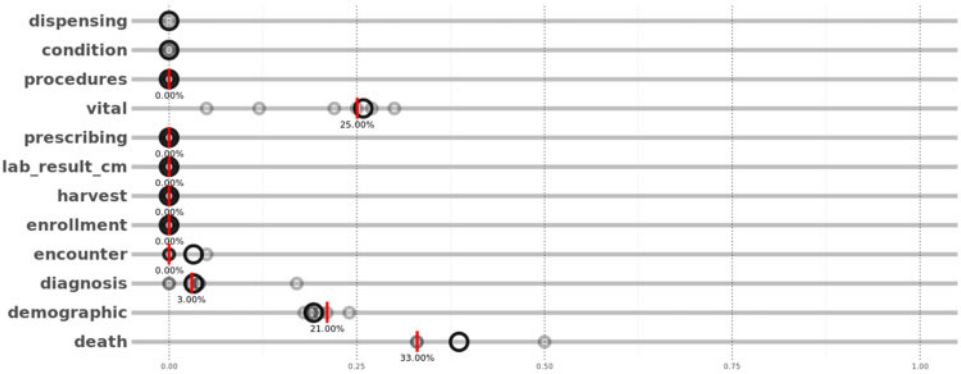

Figure 5. Vue's presentation of Common Data Model table missingness. Red bars represent a given site's missingness percentage and black rings represent missingness percentage in the network, excluding the site. Vue uses a leave-1-out approach to illustrates the network-level missingness.

Model (CDM). Here, we provide an example of how ARCH CTX has helped sites and the network to identify completeness issues in loading and refreshing PCORnet CDM tables.

Using the feedback reports, one of the pilot sites noticed that the "procedures" table was not being loaded between November 2017 and April 2018, while the table was there in the 3 prior data refreshes. The site immediately diagnosed the data transformation issue, loaded the missing table, and performed a new DQ^e-c run to update the Vue reports. To convey table loading issues, Vue produces 2 sets of visualizations. First, treemaps are used to illustrate the average table PCORnet CDM loadings across the network and the site. Through the second feedback loop, the DQT uses the treemaps to highlight priority areas for generating missing tables for each site. ARCH partner sites can also compare percent of missingness in the "important" columns of each table (required columns as defined by CDM specifications) in their repository (red bars in Figure 5) against the entire network (black rings in Figure 5). The sites can also track these missingness percentages over time. In this particular example, the visualization presented in Figure 5 (though from a previous load) helped the site to identify and alleviate the table loading issue.

Vue also produces similar visualizations that have informed actions by ARCH partner sites to improve completeness issues in, for instance, key clinical indicators such as height, weight, and blood pressure.

Table 1. Summary of the CTX implementation outcomes based on the feedback from sites

CTX Component	Outcome
Technology (DQ^e-c and Vue)	**Increased capacity notice data completeness issues**
• **Report visualizations**	
○ Longitudinal	i. Enabled tracking ETL issues over time
○ In context	ii. Expedited prioritization of diagnostic efforts
○ Interactive (dashboard)	iii. Facilitated understanding visualizations
Workflow	**Empowered involvement in improving the quality of data**
• **Peer-to-peer interactions**	iv. Created sense of mutual understanding and trustv. Encouraged participation in the system
• **Tool run reminders**	vi. Use of email was convenient vii. Encouraged participation in the system
• **File sharing**	viii. Need to be streamlined with reminders into a single process ix. Secured file sharing was preferred

CTX: completeness tracking system; ETL: extraction, transformation, and loading.

Feedback from sites

Table 1 presents a summary of the CTX implementation outcomes based on the feedback from sites. Together, the technology and workflow created a sociotechnical system that allowed implementing a horizontal (vs vertical, or top-to-bottom) data CTX across the ARCH network.

Lack of resources is often an issue to assess EHR data quality. All of the participating sites in the pilot phase agreed that the CTX has increased their (or their respected institution's) capacity to notice data completeness issues. The ETL efforts needed to comply with CDM specifications consume most of the available resources ("our capacity is mostly limited by human resource"). As a result, without the CTX reports, sites do not have the necessary capacity to explore data completeness (and in a broader scale, data quality) issues in a systematic fashion. Some of the sites even mentioned that the findings that are presented through Vue are not "surprises." For instance, one site stated that "we know there are mapping problems," but the Vue reports and the 2 feedback loops facilitate identifying the roots and planning diagnostics efforts necessary to alleviate the issues (ie, "areas we need to look at" and "where we could improve").

Most pilot sites believed that by participating in the pilot they felt more empowered to get involved in improving the quality of data in their repositories. One of the reasons for the increased feeling of empowerment was that presenting the results in the context of the network and over time helps to evaluate issues easier. As the CTX highlights ETL and mapping issues caused through CDM data transformation processes, the visualizations provided by Vue allows the users to track ETL processes over time and identify issues that might have emerged from specific ETL process changes at specific times—"I can go back and look at the ETL procedures over time and see what was going on at that point in time."

Through the pilot phase, the DQT was primarily focused on implementation of the CTX and development of Vue. To increase the frequency of actions taken by sites and the network to improve data completeness, a recommendation was to incorporate the CTX into the data transformation and refresh processes. "Usually, decision

making about mapping issues or errors flows from the top [the network]. We don't make radical changes [on our own]," as one of the sites mentioned. Incorporating the CTX into the network's standard operating procedures for data management would justify resource allocation and facilitate taking actions to improve data completeness.

The DQT has encouraged participating sites to run DQ^e-c and send their outputs after each data refresh. This process is triggered by sending email reminders to the sites. It often takes between 1 and 3 emails to collect all DQ^e-c outputs from sites. Since the beginning of the pilot phase, some of the participating sites have run DQ^e-c and sent DQ^e-c outputs to the DQT more than the number of data refreshes that have happened in this time period. Two main factors contributed to the sites' continued participation in the pilot phase. The first factor was the commitment of sites' Principal Investigators to participate in the pilot, which has resulted in the participation being part of the job responsibilities of the data analyst or database administrators. The sense of mutual understanding and trust established through the communications between the DQT and the data analyst or database administrators was the second factor for the pilot sites to respond to email reminders from the DQT. Most of this communication development happened through providing help for installation of DQ^e-c and being responsive in debugging DQ^e-c runs during the early implementation times, through which the DQT spent several tailored individual web meetings and teleconferences with the sites.

Other suggestions concerned the use of email to transfer CTX files and visualizations on feedback reports. While most sites expressed satisfaction with email communications, streamlining the file transfers into a single process was recommended to improve security and size limitations. In addition, some of the graphics on the feedback report were hard to follow. We learned that the interactive visualizations on the Vue dashboard are easier to comprehend.

Usability test

The usability test with the network's data curation leadership evolved around an outstanding question about each tab or data visualization presented in the ARCH-Vue dashboard; "How can the information presented in the tab/visualization—as a snap-shot, or in combination with other tabs/visualizations—invoke action to improve data completeness in the network?"

We found that to invoke action, the dashboard needed to flag certain issues that demand action, rather than having exclusively exploratory approach. In addition, we found that when presenting percentages, conveying information on completeness is more intuitive. For example, it is more intuitive to flip the scales on Figure 5 to exhibit presence rather than missingness—this would specially make better sense when tracking data over time to highlight certain drops as negative patterns. Overall, we found again that presenting longitudinal data is very useful, as it is difficult to track changes over time with existing tools. The longitudinal presentation of data was particularly emphasized as more useful than the cross-sectional view for table-level missingness.

Alongside the information presented in the dashboard, we found that organization of the presentation (ie, the order and linkage of the information) was critical for the usability of the dashboard to inform action. In several instances the information conveyed by certain visualizations could complement each other, through more efficient linkage and ordering, to express deeper information about the data. For example, adding the unique number of encounters to patient size would allow for better judgment on what may be causing drops in patient count (ie, whether the site was not getting enough data or whether they were improving the data).

DISCUSSION

Completeness is 1 of the 3 dimensions of EHR data quality.[7] Specially, in evaluating data quality CDMs, such as the Observational Medical Outcomes Partnership[18] or PCORnet,[5] completeness issues can accelerate diagnostic efforts to find important ETL and mapping problems that may have caused further data quality problems.[6] Over the past few years, multiple tool-based solutions have been introduced to assess different dimension of data quality. For example, PCORnet distributes SAS queries to characterize data repositories and provides resulting information back to individual sites.[10] The quality of data in federated data networks is paramount[19–22] and yet difficult to maintain and monitor, given a distributed and potentially diverse set of sites, stakeholders, information systems, and resources.[19] In addition, investigators using the network lack a transparent data quality reporting system that can help determine whether data are fit for secondary use.[23]

We argue that top-down approaches that merely rely on technology to report potential errors with the data are not enough to improve EHR data quality in distributed data networks. We have learned that it takes solid and stable commitment from many to improve data quality. Such a commitment requires all those who are involved in the process from data curation to data warehouse management to feel empowered that they can engage in improving the quality of data. The common impulse in solely tool-reliant systems is to obtain high data quality through top-down approaches (ie, to point to what is wrong with data and expect an analyst to resolve issues). We believe that high data quality is not a goal, but rather an outcome of a transparent and empowering system. A system that widens the gap between stimuli (data quality assessment reports) and responses to enable all participating actors to demonstrate that they are trying to improve data quality in their repositories and to communicate where they are struggling and what their needs are. By designing and establishing the ARCH CTX, we have aimed in creating such a system.

Through the CTX, the ARCH network now has a sociotechnical infrastructure (ie, workflows and technologies) for federated data completeness assessment, which can be further expanded to include more data quality metrics. Through the ARCH-Vue dashboard, the network now has technology for transparent data completeness (or data quality) and data profiling. Given any potential concerns with sharing the aggregate data from the network are resolved, the Vue dashboard may be repurposed as a public facing data-profiling application, enabling data consumers (eg, clinical investigators) to explore and evaluate network's data fitness for a particular use case.

Six of the ARCH network partner sites have been participating in the CTX since January 2017. As a result of establishing CTX, the network now has 6 engaged partner sites that are willing to participate in advancing the CTX from a data completeness assessment system into a complete data quality assessment system—by adding more data quality metrics to DQ^e-c. Limitations in technical resources (mainly technical staff) in installing and running DQ^e-c was the biggest challenge in implementing the system across the ARCH network. We learned that peer-to-peer communications are important in instilling a sense of trust needed for connecting technology with people to create an effective sociotechnical system. Most of this trust was built through initial implementation steps (eg, DQ^e-c installation

and debugging). In smaller settings, where fewer technical resources are often available, we found even more enthusiasm in participating in the system, for example, as a learning opportunity (learning R).

Despite the fact that the Vue is designed to operate on outputs reports from DQ^e-c, its generalizability is not limited to a specific tool. Any given tool or set of queries that produce similar outputs can use Vue to generate its site-level feedback reports and network-level dashboard. Vue uses 2 CSV files that can be easily reproduced—an example CSV file for each table is provided with DQ^e-c.[6] To implement this system at any distributed clinical data network, the only action that would need to be taken by partner sites is to run DQ^e-c and share the results with the network's governing board. A central DQT would use Vue to generate site-level feedback reports and the network-level Vue dashboard. Running Vue on aggregate site-level outputs is almost equivalent of running DQ^e-c on a dataset comprising merging all distributed data repositories across the network, which makes Vue extremely efficient from data governance and computation standpoints in multisite distributed clinical data networks. For the Vue dashboard to have the capacity to be used by a broader audience, it is designed to also provide additional competences that can extend its utility as a general data profiling[24] platform for the network. Further, utilization of the shiny dashboards provides a flexible platform for including experimental visualizations for user testing.

CONCLUSION

Systematic federated data completeness assessment is a critical first step toward monitoring and improving data quality in distributed data networks. ARCH has designed and implemented a federated data CTX to transparently evaluate EHR completeness across its distributed data repositories. The ARCH CTX, involving technologies and workflow, has enabled the network to monitor, and if needed, adjust its data management processes to maintain high-quality data for secondary use. The CTX workflow and technologies for federated data completeness assessment can be further expanded to include more data quality metrics and implemented in other networks. In addition, CTX can be adapted in other networks to improve both transparency and data quality assessment in large-scale distributed research data networks.

FUNDING

This work was funded through a Patient-Centered Outcomes Research Institute Award (CDRN-1306-04608) for development of the National Patient-Centered Clinical Research Network, known as PCORnet, National Human Genome Research Institute grants R01-HG009174 and U01-HG008685, National Institute on Minority Health and Health Disparities grants U54MD008149, 8U54MD007588, and U54MD008173, National Center for Advancing Translational Sciences (NCATS) of the National Institutes of Health grants UL1 TR000371, UL1TR002378, UL1 TR001105, UL1 TR001857, U01 TR002393, National Library of Medicine grant R01 LM011829, the Reynolds and Reynolds Professorship in Clinical Informatics, and the Cancer Prevention Research Institute of Texas (CPRIT) Data Science and Informatics Core for Cancer Research (RP170668).

The statements presented in this publication are solely the responsibility of the author(s) and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors or Methodology Committee or other participants in PCORnet.

AUTHOR CONTRIBUTIONS

Conceived study design: HE. Developed tools: HE. Contributed to server specification: JGK. Local tool implementation: EA, RJA, GG, GL, NP, KW. Contributed data to study: WGA, NRA, HM, MDN, EO, BO, AQ, GER. Contributed to data analysis and visualization: HE. Wrote manuscript: HE. Reviewed and edited manuscript: JGK, SRW, WGA, EVB, KDM, SNM. All authors approved the final manuscript.

SUPPLEMENTARY MATERIAL

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

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their helpful and constructive comments that greatly contributed to improving the final version of the article.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Hsiao C-J, Hing E, Socey TC, Cai B. Electronic health record systems and intent to apply for meaningful use incentives among office-based physician practices: United States, 2001-2011. *NCHS Data Brief* 2011; (79): 1–8.
2. Murdoch T, Detsky A. The inevitable application of big data to health care. *J Am Med Inform Assoc* 2013; 309 (13): 1351–2.
3. Liaw ST, Rahimi A, Ray P, *et al*. Towards an ontology for data quality in integrated chronic disease management: a realist review of the literature. *Int J Med Inform* 2013; 82 (1): 10–24.
4. Ross MK, Wei W, Ohno-Machado L. "Big data" and the electronic health record. *Yearb Med Inform* 2014; 9: 97–104.
5. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. *J Am Med Inform Assoc* 2014; 21 (4): 578–82.
6. Estiri H, Stephens KA, Klann JG, Murphy SN. Exploring completeness in clinical data research networks with DQe-c. *J Am Med Inform Assoc* 2018; 25 (1): 17–24.
7. Kahn MG, Callahan TJ, Barnard J, *et al*. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *EGEMS (Wash DC)* 2016; 4 (1): 1244.
8. Mandl KD, Kohane IS, McFadden D, *et al*. Scalable Collaborative Infrastructure for a Learning Healthcare System (SCILHS): architecture. *J Am Med Inform Assoc* 2014; 21 (4): 615–20.
9. Ball R, Robb M, Anderson SA, Dal Pan G. The FDA's sentinel initiative–a comprehensive approach to medical product surveillance. *Clin Pharmacol Ther* 2016; 99 (3): 265–8.
10. Qualls LG, Phillips TA, Hammill BG, *et al*. Evaluating foundational data quality in the National Patient-Centered Clinical Research Network (PCORnet®). *EGEMS (Wash DC)* 2018; 6 (1): 3.
11. Khare R, Utidjian L, Ruth BJ, *et al*. A longitudinal analysis of data quality in a large pediatric data research network. *J Am Med Informatics Assoc* 2017; 24 (6): 1072–9. doi: 10.1093/jamia/ocx033.
12. Trochim WM, Cabrera DA, Milstein B, Gallagher RS, Leischow SJ. Practical challenges of systems thinking and modelling in public health. *Am J Public Health* 2006; 96 (3): 538–46.
13. Leischow SJ, Best A, Trochim WM, *et al*. Systems thinking to improve the public's health. *Am J Prev Med* 2008; 35 (2): S196–203.
14. Peters DH. The application of systems thinking in health: why use systems thinking? *Health Res Policy Syst* 2014; 12: 51.

15. Carey G, Malbon E, Carey N, Joyce A, Crammond B, Carey A. Systems science and systems thinking for public health: a systematic review of the field. *BMJ Open* 2015; 5: e009002. doi: 10.1136/bmjopen-2015-009002.
16. Murphy SN, Weber G, Mendis M, *et al*. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc* 2010; 17 (2): 124–30.
17. Klann JG, Abend A, Raghavan VA, Mandl KD, Murphy SN. Data interchange using i2b2. *J Am Med Inform Assoc* 2016; 23 (5): 909–15.
18. Observational Medical Outcomes Partnership (OMOP) Initiative. What it is, its importance and results so far. *Basic Clin Pharmacol Toxicol* 2010; 107 (Suppl 1): 17.
19. Brown JS, Kahn M, Toh S. Data quality assessment for comparative effectiveness research in distributed data networks. *Med Care* 2013; 51: S22–9. [10.1097/MLR.0b013e31829b1e2c]
20. Kahn MG, Batson D, Schilling LM. Data model considerations for clinical effectiveness researchers. *Med Care* 2012; 50 (Suppl): S60–7.
21. 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.
22. Gregori D, Berchialla P. Quality of electronic medical records. In: Faltin FW, Kenett RS, Ruggeri F, eds. *Statistical Methods in Healthcare*. West Sussex, UK: Wiley; 2012: 456–76.
23. Kahn MG, Brown JS, Chun AT, *et al*. Transparent reporting of data quality in distributed data networks. *EGEMS (Wash DC)* 2015; 3 (1): 1052.
24. Estiri H, Lovins T, Afzalan N, Stephens KA. Applying a participatory design approach to define objectives and properties of a "data profiling" tool for electronic health data. *AMIA Summits Transl Sci Proc* 2016; 2016: 60–7.