12-2-2014

Developing Electronic Data Methods Infrastructure to Participate in Collaborative Research Networks

Elisa L. Priest DrPH, MPH
Baylor Scott & White Health, elisapri@baylorhealth.edu

Christopher Klekar MPH
Baylor Scott & White Health

Gabriela Cantu MPH
Baylor Scott & White Health

Candice Berryman MBA
Baylor Scott & White Health

See next pages for additional authors

Recommended Citation

Priest, Elisa L. DrPH, MPH; Klekar, Christopher MPH; Cantu, Gabriela MPH; Berryman, Candice MBA; Garinger, Gina MBA; Hall, Lauren MPH; Kouznetsova, Maria PhD, MPH; Kudyakov, Rustam MD, MPH; and Masica, Andrew MD, MSCI (2014) "Developing Electronic Data Methods Infrastructure to Participate in Collaborative Research Networks," *eGEMs (Generating Evidence & Methods to improve patient outcomes)*: Vol. 2: Iss. 1, Article 18.
DOI: http://dx.doi.org/10.13063/2327-9214.1126
Available at: http://repository.academyhealth.org/egems/vol2/iss1/18

The Electronic Data Methods (EDM) Forum is supported by the Agency for Healthcare Research and Quality (AHRQ), Grant 1U18HS022789-01. eGEMs publications do not reflect the official views of AHRQ or the United States Department of Health and Human Services.

Developing Electronic Data Methods Infrastructure to Participate in Collaborative Research Networks

Abstract

Context: Collaborative networks support the goals of a learning health system by sharing, aggregating, and analyzing data to facilitate identification of best practices care across delivery organizations. This case study describes the infrastructure and process developed by an integrated health delivery system to successfully prepare and submit a complex data set to a large national collaborative network.

Case Description: We submitted four years of data for a diverse population of patients in specific clinical areas: diabetes, chronic heart failure, sepsis, and hip, knee, and spine. The most recent submission included 19 tables, more than 376,000 unique patients, and almost 5 million patient encounters. Data was extracted from multiple clinical and administrative systems.

Lessons Learned: We found that a structured process with documentation was key to maintaining communication, timelines, and quality in a large-scale data submission to a national collaborative network. The three key components of this process were the experienced project team, documentation, and communication. We used a formal QA and feedback process to track and review data. Overall, the data submission was resource intensive and required an incremental approach to data quality.

Conclusion: Participation in collaborative networks can be time and resource intense, however it can serve as a catalyst to increase the technical data available to the learning health system.

Acknowledgements

This submission is based on work presented at the 2014 EDM Forum Symposium. The findings and conclusions in this document are those of the authors, who are responsible for its content, and do not necessarily represent the views of the High Value Healthcare Collaborative.

Keywords

Methods, Informatics, Learning Health System

Disciplines

Health Services Research

Authors

Elisa L Priest, *Baylor Scott & White Health*; Christopher Klekar, *Baylor Scott & White Health*; Gabriela Cantu, *Baylor Scott & White Health*; Candice Berryman, *Baylor Scott & White Health*; Gina Garinger, *Baylor Scott & White Health*; Lauren Hall, *Baylor Scott & White Health*; Maria Kouznetsova, *Baylor Scott & White Health*; Rustam Kudyakov, *Baylor Scott & White Health*; Andrew Masica, *Baylor Scott & White Health*.

Developing Electronic Data Methods Infrastructure to Participate in Collaborative Research Networks

Elisa L. Priest, DrPH, MPH; Christopher Klekar, MPH; Gabriela Cantu, MPH; Candice Berryman, MBA; Gina Garinger, MBA; Lauren Hall, MPH; Maria Kouznetsova, PhD, MPH; Rustam Kudyakov, MD, MPH; Andrew Masica, MD, MSCI[i]

Abstract

Context: Collaborative networks support the goals of a learning health system by sharing, aggregating, and analyzing data to facilitate identification of best practices care across delivery organizations. This case study describes the infrastructure and process developed by an integrated health delivery system to successfully prepare and submit a complex data set to a large national collaborative network.

Case Description: We submitted four years of data for a diverse population of patients in specific clinical areas: diabetes, chronic heart failure, sepsis, and hip, knee, and spine. The most recent submission included 19 tables, more than 376,000 unique patients, and almost 5 million patient encounters. Data was extracted from multiple clinical and administrative systems.

Lessons Learned: We found that a structured process with documentation was key to maintaining communication, timelines, and quality in a large-scale data submission to a national collaborative network. The three key components of this process were the experienced project team, documentation, and communication. We used a formal QA and feedback process to track and review data. Overall, the data submission was resource intensive and required an incremental approach to data quality.

Conclusion: Participation in collaborative networks can be time and resource intense, however it can serve as a catalyst to increase the technical data available to the learning health system.

Context

The core of a learning health system is "continuous knowledge development, improvement, and application."[1] The first recommendation of the Institute of Medicine for developing a learning health system is to improve the digital infrastructure and the capacity to capture clinical, care delivery, and financial data.[1] However, "data alone are not sufficient for learning."[1,2] Collaborative networks support the goals of a learning health system by sharing, aggregating, and analyzing data to facilitate identification of best practice care models across delivery organizations.[3] Despite the benefits of participation, submission of required data elements is often complex and resource intensive.[4,5] Data may be stored across multiple systems, may require calculations and transformation, and may be unstructured.[6] Further, data may be incomplete, inconsistent across data sources, or otherwise inaccurate.[5,7]

This case study describes the infrastructure and process developed by an integrated health delivery system to successfully prepare and submit a complex data set to a large national collaborative network. First, we describe the scope of the data submission. Next, we examine the infrastructure and processes that we developed. Finally, we detail the lessons learned from completing two rounds of data submissions.

Case Description

Setting

Baylor Scott & White Health (BSWH) is the largest not-for-profit health care system in Texas and one of the largest systems in the United States. BSWH comprises Baylor Health Care System (BHCS) and Scott & White Healthcare (SWH) who joined together in 2013 to create a new model system to meet the demands of health care reform, the changing needs of patients, and the recent advances in clinical care.

The High Value Healthcare Collaborative (HVHC) consists of 19 geographically diverse health care delivery systems (encompassing markets of over 70 million individuals) and The Dartmouth Institute for Health Policy and Clinical Practice.[3] Participating members submit electronic data for inclusion in a comprehensive data warehouse designed to allow benchmarking, querying, reporting, and analysis. The goals of the HVHC are to improve care, improve health, and reduce costs by identifying and accelerating widespread adoption of best-practice care models and innovative, value-based payment models.[3]

[i]Baylor Scott & White Health

Both Baylor Health Care System and Scott & White Healthcare were members of the HVHC before the 2013 merger. The primary electronic health record (EHR) systems for the two systems are not integrated, and HVHC data submissions are performed by two independent teams. This case study focuses on the processes implemented by the legacy BHCS team.

Scope of Data Submission

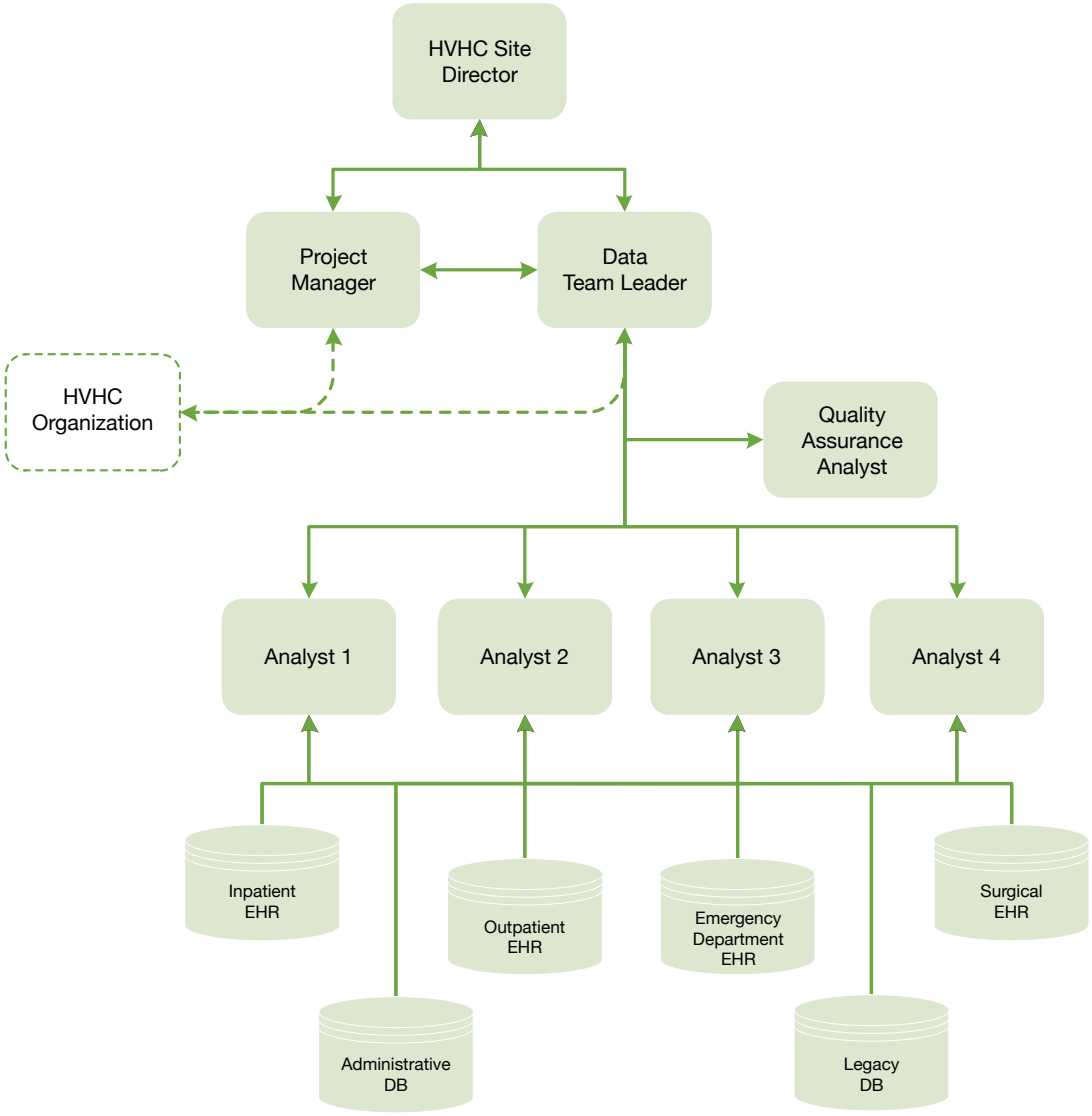
The High Value Healthcare Collaborative created a comprehensive specifications document for data submissions that included population selection and variable names, definitions and formats. Four years of data were required for a diverse population of patients in specific clinical areas: diabetes, chronic heart failure, sepsis, and total hip replacement, total knee replacement, and spine surgery.

The legacy BHCS team completed two large-scale data submissions in 2014. The most recent submission—in July 2014—included 19 tables, more than 376,000 unique patients, and almost 5 million patient encounters. Data was extracted from multiple clinical and administrative systems including the following: inpatient, outpatient, emergency department, and surgical EHRs; and administrative data. In addition, BHCS implemented new systems during the required time frame. Thus, legacy data from pharmacy and laboratory systems were also required. Over 80 raw data sets were produced, and then assembled into the final data submission.

Team

The large scope of these electronic data submissions justified the use of a structured project team. We created an interdisciplinary team with a data team leader, project manager, quality assurance analyst, and masters- and doctoral-level analysts and programmers (Figure 1). These personnel were existing resources provided from the departmental budget of the BHCS HVHC Site Director within the BHCS Center for Clinical Effectiveness because data management and analytic capability are viewed as essential infrastructure for health care delivery organization operations. The HVHC did not mandate the structure of the team.

Figure 1. Project Team Structure



Four full-time equivalent (FTE) analysts were assigned to coordinate specific cohorts of data (diabetes, chronic heart failure, sepsis, total hip replacement, total knee replacement, spine surgery) in order to develop clinical-area focused subject matter experts within the team. The quality assurance (QA) analyst (0.5 FTE) performed quality control checks on the data sets. The data team leader (0.5 FTE) oversaw the technical coordination of the project, developed and maintained the QA process, participated in quality control, and helped to manage priorities. The project manager (0.10 FTE) enforced timelines and facilitated communication with the HVHC coordinating center. The team met weekly in a structured format to ensure communication of goals and timelines, facilitate troubleshooting, and document progress.

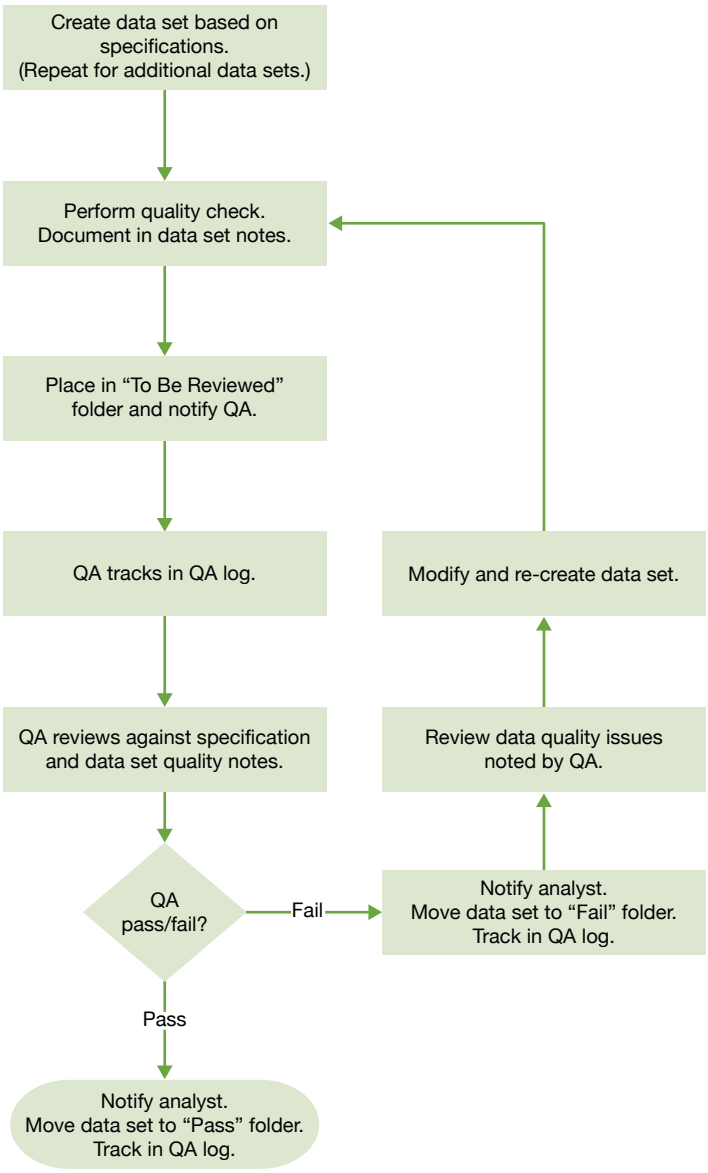
Process

We created a formal QA and feedback process to track and review data (Figure 2). This need was identified in the early team meetings as we reviewed the scope of the project. Before the first data set submission, a draft of the QA process was produced by the data team leader and then circulated to the entire team for review and feedback. The goal was to efficiently produce reliable, high-quality data sets. The QA process was modified as needed, based on feedback from the team.

We expanded the HVHC requirements document to map the origin of each variable within our local systems. In many cases, the same variable had more than one system origin (i.e., inpatient and outpatient). We added variables that were required for calculations or processing, and we also included formal quality-control checks in the documentation.

The analysts produced cohort-specific data sets, placed them in a staging folder called "To Be Reviewed," and notified the quality assurance (QA) team. This accelerated the review process because an available QA team member could quickly pick a file from the folder without waiting to be assigned. Each data set was accompanied by a standard set of notes outlining known data quality issues. Examples of data quality issues included missing data, data that was inconsistent across systems, or data that was coded after abstraction from text fields. The QA team compared data sets against the internal requirements documentation for correct naming, order, variable formats, and missing and unacceptable values, then placed them into a "Pass" or "Fail" folder. Data placed in the "Fail" folder were reviewed and updated by the assigned analyst, with repetition of this sequence until the data met specifications. The QA team tracked the preparation and validation steps in a log and provided feedback and updates on file status (Table 1). Once all cohort-specific data files were validated, final data set assembly was completed. This step required the concatenation of the more than 80 cohort-specific data sets into 19 Health Insurance Portability and Accountability Act (HIPAA) limited data sets. The last step in the process was to review and validate the final data sets and create the documentation for submission. This documentation included a summary of every table and the known data quality issues. In addition, the data team leader tracked every table submitted in a data transfer log.

Figure 2. Quality Assurance (QA) Process for each Data Set



After submission, the HVHC coordinating center produced a list of questions from its analysts about specific data fields. These questions arose from HVHC data quality checks designed to standardize and prepare the data for specific analyses. We discussed these questions during internal team meetings, and the analysts responsible for the data in question were assigned to research the issue and provide either an explanation or an updated data set. The data team leader tracked every question and resolution in a data issues log. Thus, one data submission actually consisted of one primary submission followed by incremental updates to improve the quality of the data. For example, we updated our documentation to note that we could not provide outpatient prescriptions filled due to a lack of outpatient pharmacies after the HVHC coordinating center noted that the relevant data points were missing.

Table 1. Quality Assurance (QA) Tracking Log

Demographics Table							
Cohort	**Original Name**	**Updated Name**	**Owner**	**Data Set Draft**	**Validation**	**Pass/Fail**	**Notes**
Hip/Knee	Demos_thr/thr_july	Demo_ip_thr	Analyst 1	Done	QA 1	Pass	Some last dates missing
	Tkth_op_demo_v1	Demo_op_thr	Analyst 2	Done	QA 2	Pass	
Diabetes	Dm_op_demo	Demo_op_dm	Analyst 3	Done	QA 1	Pass	
Sepsis	Demo_ip_sep	Demo_ip_sep	Analyst 4	Done	QA 1	Pass	
CHF	Demo_ip_chf	Demo_ip_chd	Analyst 4	Done	QA 1	Pass	
	CHf_op_demo_v2	Demo_op_chf	Analyst 3	Done	QA 2	Pass	
Spine	Demos_spine_july	Demo_ip_spine	Analyst 2	Done	QA 1	Pass	
	Spine_op_demo_v1	Demo_op_spine	Analyst 3	Done	QA 2	Pass	

Lessons Learned

Importance of a Coordinated, Well-Structured Process

We found that a structured process with documentation was key to maintaining communication, timelines, and quality in a large-scale data submission to a large national collaborative network. The three key components of this process were the project team, documentation, and communication (Table 2).

Table 2. Core Components

Core Components
Team • Experienced team • Defined roles – Data Team Lead – Project Manager – Analysts/Programmers ◦ Cohort specific ◦ Database specific – Quality Assurance Analyst
Documentation • Defined quality assurance (QA) process • Facilitated communication • Data management documentation – System-specific requirements – Status tracking log – Data set quality documentation – Data transfer log – Data issues log • Project management documentation – Timelines – Meeting notes – Follow-up items
Communication • Weekly meetings • Improved processes and documentation, as a result

Our first key component was the experienced team. Although the first team meetings were informal discussions to understand the scope of the project, it became clear that a structured team with clearly defined roles was necessary to complete a data submission of this scale. The data team leader had many years of experience developing and coordinating data management processes and the project manager had years of experience in coordinating technical projects. Similarly, the analysts were experienced and familiar with the data sources before the initiation of the project. This knowledge was critical for identification of the required variables across multiple complex data systems. In addition, it was necessary to begin to develop the analysts into subject matter experts who understood the nuances and complexities of the data. The Institute of Medicine states that the transformation of data into knowledge requires an understanding of who collects the data, how it is collected, why it is collected, and what is collected.[2] Each analyst was assigned to pull data for one or more clinical-area focused cohorts to develop this understanding. Health systems facing smaller data submissions can still apply our approach of clearly defining roles within a project team. For example, one team member may serve the role of both analyst and project manager. Health systems without experienced analysts should plan adequate time for training and development. Depending on the complexity of the systems, even basic familiarity with the data and systems can take months to develop.

Next, we relied on documentation. Our QA process was written and included specifications for the data and a QA tracking log to facilitate project management, organization, and communication. Each data set was produced with quality documentation so that the limitations and complexities of the data could be accurately communicated. All data transfers and data quality issues were documented. The project manager also kept written documentation of project meetings, follow-up items, and timelines. This additional documentation was important because the project generated hundreds of emails, and it was helpful to keep key decisions summarized in one location.

Likewise, communication was critical in these efforts. We had weekly meetings to discuss project status and challenges. For example, the internal specification document resulted from a discussion where we realized that analysts were creating and mapping variables using different methods. For example, the primary payer categories in our source systems were different, and in some instances were mapped to the HVHC payer categories inconsistently. Also, some extreme laboratory values were set to "missing" depending on the analyst's prior experience. The creation of the internal specifications required the analysts to review the sources and methods used to pull each data point and reach a consensus. This documentation helped to ensure that the methods used were

consistent across the cohorts. In addition, the documentation on the QA log allowed all team members to see the progress of the data set creation. Finally, we had team meetings after each major data submission to discuss lessons learned and ongoing challenges, and to identify targets for process improvement. These targets included technical processes such as database access issues and programming efficiencies, as well as improvements needed in documentation or communication.

Resource Intensive

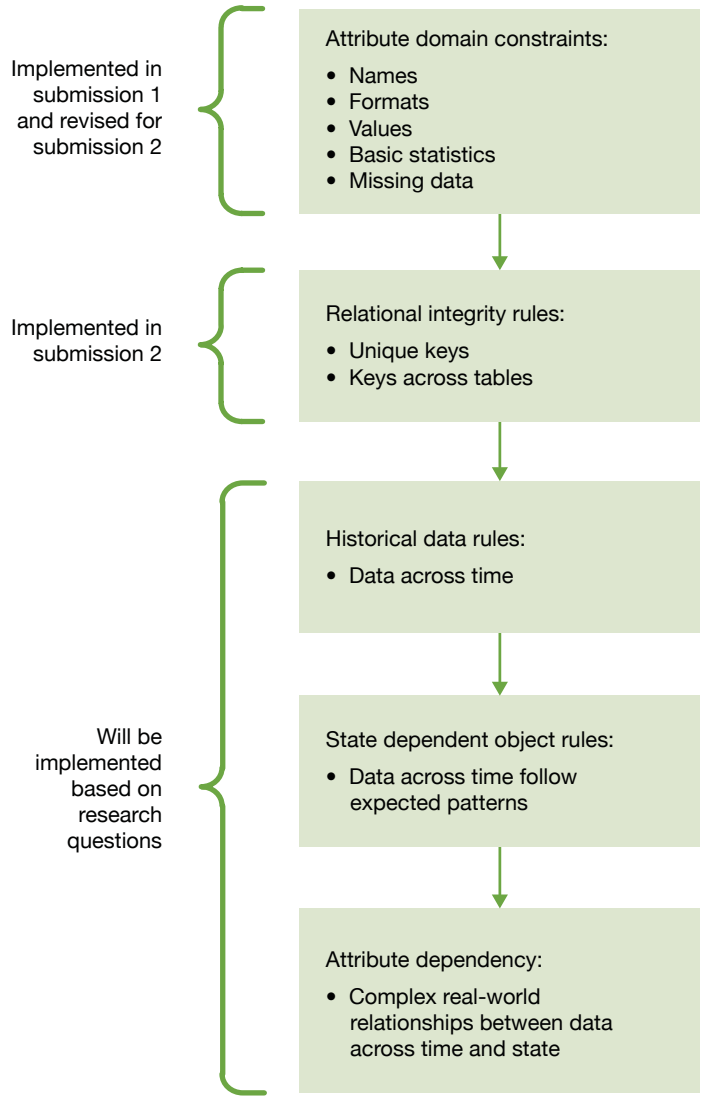
Even with our structured process, significant resources were required for the preparation and submission of the data sets. For both large submissions, we underestimated the time and resources needed. The first submission used over 1,800 hours of personnel time and included the bulk of the creation of the SAS and SQL programming. Much of this time was invested in exploratory analysis to better understand the data. The most recent data submission required just under 1,000 hours of time. This time included additional data exploration and programming refinements (due to changes in the specification document). We saw a large savings in time for the second data submission because the foundational programming and data exploration had been completed. Despite this, we still underestimated the resources needed because we faced new technical challenges. The data submission was larger and exposed the limits of our available technical infrastructure. The required processing time, memory, and space were extensive and exceeded our expectations. We will continue to refine our technical and QA processes, and we expect to see additional efficiencies in future submissions. However, we have learned to prepare for unexpected challenges and to overestimate the resource time needed for the submission.

Data Quality Improvement Is Continuous

Data in an integrated delivery system is expansive and complex. Every research and quality improvement project reveals different attributes of the data. This project taught our team to use an incremental approach to data quality. The scope of the project was too large to expect perfect data sets from the start. We used a data quality framework adapted by Khan et al. (2012) that describes quality assessment for single and multisite EHR-based research.[8,9] In a multisite research collaborative, two stages of quality assessment take place. First, each site must perform checks on its data before submission. Next, the coordinating center may perform quality checks and query the sites. These checks may occur as data is submitted and again as specific research questions are investigated.

At the individual site, there are five domains of data quality checks of increasing complexity that can be applied to data (Figure 3).[8] Because of the resources required to complete the data submissions, we focused first on basic attribute domain constraints. These checks included variable names, formats, and value distributions. For the second data submission, we expanded the attribute domain constraints and added relational integrity rules to the checks performed. Completing these data checks was necessary, but challenging, due to time and resource constraints. Frequently, the checks revealed data quality issues that needed to be addressed. For example, we found duplicate records in our tables created from inconsistent matching and identifiers across systems. A single data quality issue could result in hours of research time, modifications to programming, and reproduction of data sets, and could expose other previously unknown issues. We discussed these issues in the team meetings, and ad hoc meetings were called when necessary. Further, the data team leader tracked and prioritized the issues and decided which issues to resolve immediately and which issues to document for future resolution. We plan to expand and refine our quality checks for future submissions.

Figure 3. Data Quality Assessment[8] (QA) Implementation



Value of Participation in Collaborative Networks

Finally, we found that participation in a data collaborative drives learning system status by forcing the examination of data that has not been extensively used at the local system level. For example, BHCS used benchmarking data from partners in data collabora-

tives (obtained through analogous data extraction processes) to help guide initiatives related to the improvement of sepsis care. In addition, for this project we identified data elements from the EHR used exclusively in surgical suites. Specific fields such as device serial number and manufacturer were not available in other systems and we worked with the necessary technical teams to create a data mart to directly access the data. An extensive amount of time was required to understand the tables and variables and to begin to understand the strengths and limitations of the data. This process was repeated multiple times for each of the cohorts and data systems.

Thus, we are now able to access additional data systems for internal quality improvement initiatives. Our next step is to work with the data governance workgroup to identify the application owners for the systems and develop a process for providing formal feedback to improve data quality if we see data inconsistencies.

Conclusion

This case study described the infrastructure and process developed by an integrated health delivery system to successfully prepare and submit a complex data set to a large national collaborative network. This process required an experienced technical team and sufficient infrastructure to assure success. Our team of analysts accessed multiple electronic systems, identified and interpreted required data, and assembled the data sets. We used a structured quality assessment process to ensure the data met the specifications. Other key team members focused on project management and QA and were essential to help the technical team to communicate, remain organized, and document processes. The lessons that we learned and the framework we developed can be applied by any health delivery system preparing data submissions.

Participation in networks can be time and resource intense, however it can serve as a catalyst to increase the technical data available to the learning health system. This aligns with the Institute of Medicine's assertion that the digital infrastructure is one of the foundations of a learning health system.[1] Further, collaborative networks support the goals of a learning health system by sharing, aggregating, and analyzing data to facilitate identification of best practices care across delivery organizations.

Acknowledgements

This submission is based on work presented at the 2014 EDM Forum Symposium. The findings and conclusions in this document are those of the authors, who are responsible for its content, and do not necessarily represent the views of the High Value Healthcare Collaborative.

References

1. Institute of Medicine. *Best Care at Lower Cost: The Path to Continuously Learning Health Care in America*. Washington, D.C.: The National Academies Press;2013.
2. Institute of Medicine. *Digital Data Improvement Priorities for Continuous Learning in Health and Health Care: Workshop Summary*. Washington, D.C.: The National Academies Press;2013.
3. High Value Healthcare Collaborative. http://highvaluehealthcare.org. Accessed November 4, 2014.
4. McGraw D, Leiter A. Pathways to Success for Multisite Clinical Data Research. *eGEMS*. 2013;1(1).
5. 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. *Medical care*. Aug 2013;51(8 Suppl 3):S80-86.
6. Capurro D, Yetisgen M, van Eaton E, Black R, Tarczy-Hornoch P. Availability of Structured and unstructured Clinical Data for Comparative Effectiveness Research and Quality Improvement: A Multisite Assessment. *eGEMS*. 2014;2(1).
7. Hersh WR, Cimino J, Payne PRO, et al. Recommendations for the Use of Operational EHR Data in Comparative Effectiveness Research. *eGEMS*. 2013;1(1).
8. Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Steiner JF. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. *Medical care*. Jul 2012;50 Suppl:S21-29.
9. Brown JS, Kahn M, Toh S. Data quality assessment for comparative effectiveness research in distributed data networks. *Medical care*. Aug 2013;51(8 Suppl 3):S22-29.