LC Data QUEST: A Technical Architecture for Community Federated Clinical Data Sharing

Kari A. Stephens PhD, Ching-Ping Lin PhD, Laura-Mae Baldwin MD, Abigail Echo-Hawk MPS, Gina A. Keppel MPH, Dedra Buchwald MD, Ron J. Whitener JD, Diane M. Korngiebel DPhil, Alfred O. Berg MD, Robert A. Black MA, Peter Tarczy-Hornoch, MD
Institute of Translational Health Sciences
University of Washington, Seattle, WA

Abstract

The University of Washington Institute of Translational Health Sciences is engaged in a project, LC Data QUEST, building data sharing capacity in primary care practices serving rural and tribal populations in the Washington, Wyoming, Alaska, Montana, Idaho region to build research infrastructure. We report on the iterative process of developing the technical architecture for semantically aligning electronic health data in primary care settings across our pilot sites and tools that will facilitate linkages between the research and practice communities. Our architecture emphasizes sustainable technical solutions for addressing data extraction, alignment, quality, and metadata management. The architecture provides immediate benefits to participating partners via a clinical decision support tool and data querying functionality to support local quality improvement efforts. The FInDiT tool catalogues type, quantity, and quality of the data that are available across the LC Data QUEST data sharing architecture. These tools facilitate the bi-directional process of translational research.

Introduction and Background

Data sharing across disparate ambulatory care based electronic medical records is necessary to facilitate comparative effectiveness research (CER) in primary care. Several national on-going efforts to develop federated data sharing architectures across electronic medical record systems have targeted open-sourced solutions, including i2b2, caBIG, HMORN, and DARTNet.[1-4] DARTNet in particular has a successful track record for facilitating numerous CER projects across primary care settings.[5]

Adoption of electronic medical records in primary care, from single to large group practices, has reached critical mass in the last decade. Practices, in response to national Health Information Technology incentives that promote efforts such as meaningful use and patient-centered medical homes, are engaging architectural solutions to conduct data sharing. This engagement is building the data sharing capacity critical to promote research. The Clinical and Translational Science Award (CTSA) institutions are well positioned to promote these architectural solutions and are engaged in developing data sharing capacity across large population bases. Our CTSA efforts at the University of Washington's Institute of Translational Health Sciences (ITHS) include the Locally Controlled Data QUery, Extraction, Standardization and Translation (LC Data QUEST) pilot project aimed at creating data sharing capacity within the Washington, Wyoming, Alaska, Montana, Idaho region across primary care based practices. LC Data QUEST is a collaboration between the Biomedical Informatics and Community Outreach and Research Translation Cores of the ITHS. The LC Data QUEST architecture was designed to facilitate translational research by increasing the accessibility to clinical health data captured in electronic medical record systems in primary care clinics serving rural populations, in order to accelerate the integration of new findings into care practices.[6]

We explored the perceptions, priorities, and concerns of our partner practices and tribal communities regarding data sharing and research.[7] The design of the technical data sharing architecture and tools that grew from these explorations within our CTSA and with our partner practices and tribal communities are discussed here.

Major technical components of a successful data sharing architecture include: 1) the extraction, transformation, load (ETL) process that transfers and aligns health data from the local electronic medical records (EMR) to a separate repository, 2) a set of end-user applications that can deliver data appropriately to users,[8] 3) data quality management, and 4) metadata management. Data quality management plays a critical role as the data may be of poor initial quality and if not controlled, can rapidly degrade over time due to interfacing with external data sources (i.e., laboratory or pharmacy services) that are often unknown *a priori*.[9,10] Furthermore, metadata management, including data cleansing specifications and mapping rules, are necessary to align and document data provenance for

effective data sharing.[8] Although these technical components are fundamental to data sharing systems, the methods for addressing these components vary from fully manual to fully automated. We selected the best technical solutions to carry out these activities that met our system requirements while staying within the scope of our financial and human resource constraints.

Methods

We began by seeking out practices and tribal communities willing to partner with us during the formative stages of this project. Partners who understood that our project focused on analyzing and designing a scalable technical architecture that would support a data sharing network and partners that were willing to endure this pilot process were sought. We brought together a multidisciplinary CTSA team that represented clinical, community engagement, ethics, and informatics interests to clarify and articulate a common vision to communicate to potential community partners. Ten practices and tribal communities were visited and evaluated. We developed a feasibility assessment using a semi-structured interview designed to determine partner interest and technical readiness. Results were collated and evaluated by our multidisciplinary team, and six partner sites were selected, all of which agreed to participate.

From these engagements, a set of systems requirements were developed that emphasized the need for local control of the data repository and the need to vet and authorize each query. We concluded that a federated model, where individual data repositories reside at each site, was the appropriate architecture for the governance and security concerns of our partners. Given that the nature of scientific research often prioritizes immediate benefit to the researcher and not the community partner, a locally owned solution that allowed vetting of all queries was deemed necessary. Our partners were wary of data fishing of sensitive disease based issues that could result in exploitation or might stigmatize communities and reveal practice quality issues. In addition, our tribal partners as sovereign nations had the authority to regulate and review any research conducted on their land and, as a condition of their partnership, required that all control over the data reside with each tribal partner. To address these sensitivities, our system requirements included the ability to evaluate and approve every query before execution, including aggregated anonymized queries.

LC Data QUEST was funded as a five-year pilot project to develop a proof-of-concept infrastructure. We targeted a low-cost, self-sustaining architecture to facilitate its continued sustainability and expansion. We evaluated existing solutions for both the ETL and end-user applications based upon their ability to meet our system and feasibility requirements of local control, self-sustainability, and low-cost. We examined existing solutions and processes used by other research networks by researching and installing open source, academic, and vendor tools.[1,3,11-13] To thoroughly evaluate these solutions, an IRB approved set of clinical data was constructed to use as a real-world test bed. The cost and benefits of developing ETL and end-user application solutions within the ITHS Biomedical Informatics Core were compared directly against existing open source technologies and outsourcing efforts to vendor(s).

During our evaluations we continued to engage our prospective partners by openly discussing potential architecture solutions and barriers and facilitators to data sharing.[14] These iterative engagements directly informed our design processes and were critical to defining and establishing data sharing governance and building a foundation of community collaboration.

Results

We evaluated six end-user applications and research network processes for data sharing. The applications and processes did not meet the security requirements for local control and vetting of individual queries. Several processes required unavailable local expertise at the clinics such as programming and database staff to manually execute queries. Given the infancy state of end-user applications that supported our data sharing security and governance needs and the enormity of developing the requisite software ourselves, we limited the project scope to implementing the ETL process without an end-user application layer. Thus, for the pilot project, data sharing across sites would be supported manually through ITHS and vendor collaborations, rather than through an end-user application. As we developed the ETL process, we also developed data quality and metadata management approaches to establish a foundation for a data sharing architecture that could apply an end-user application layer in the future.

Our sites use diverse EMR products without agreed upon data standards and data practices. Among the 10 evaluation sites, four different EMR products were in use with variations across ownership and physical location of EMR data for our six selected partner sites. This complicated cost and governance issues to gaining access to perform extractions. Three sites had physical access and ownership over their EMR data, providing the best ease of access. Two sites had only ownership or physical access, complicating access.

Our last site had neither physical access nor ownership over their EMR data, postponing our ability to include them in this new architecture until they migrated to a new EMR system.

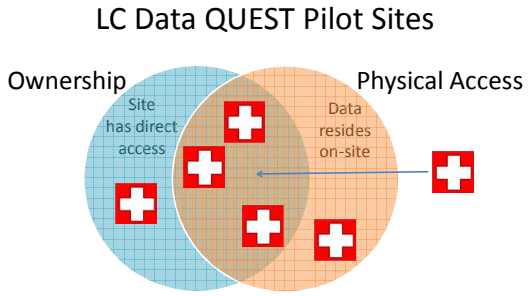

Figure 1. LC Data QUEST pilot sites diversity in EMR data access.

We recognized that engaging a vendor who specialized in ETL across diverse practice settings would be the most cost-effective method for performing ETL at multiple sites with multiple vendor supplied EMRs. Therefore, vendors who had previous experience with ETL processes at primary care clinics, experience with multiple EMR products, and data quality and semantic alignment strategies were sought. Discussions with national CTSA colleagues building similar architectures and Practice Based Research Networks (PBRNs) led us to focus on point-of-care based clinical decision support (CDS) tools for our data quality approach. Employing a clinical decision support tool at sites offered two key features: 1) a natural data quality feedback loop to sustain the usability of the repositories by actively using and iterating the extracted clinical data in practice; and 2) immediate benefit to our partners. Therefore, we pinpointed vendors based on their experience with delivering ETL services in medical settings, point-of-care CDS tools, and solutions to semantic alignment. These requirements, in addition to our original requirements of including a federated architecture and remote management, comprised the core set of system requirements that we used to evaluate vendors.

Specifically, we identified and evaluated four vendors (W, X, Y, Z) with experience in providing data services to medical settings. Table 1 summarizes our system requirements and vendor evaluations. All vendors had the ability to remotely manage their systems. Vendors W and X's primary business was primary care CDS tools using an extracted data repository. Vendor W extracted EMR data into repositories located at practice sites while Vendor W extracted the data into a repository located remotely at their own facilities. Vendor Y was a clinical data warehousing consulting firm with extensive experiences performing custom ETL projects at large hospitals, but not small primary care based clinic settings. Vendor Z specialized in health data exchange services. Both Vendors Y and Z lacked the necessary ETL experience and a CDS tool. Vendor W met all of our criteria, met our budgetary constraints, and brought additional expertise in national health guidelines and delivering data extractions to support comparative effectiveness research.

ETL Requirement	Vendor
Data exported to separate repositories (federated solutions vs. centralized data sharing solutions)	W, X, Y
Repository stored locally at site	W, Y
Point-of-care clinical decision support tool available	W, X
Remote management (no onsite support staff needed)	W, X, Y, Z
Previous ETL experience with primary care clinic based EMRs	W, X

Table 1. Vendor evaluation matrix. Vendors A and B specialized in clinical decision support products. Vender C was a clinical data warehousing consulting firm. Vendor D specialized in health data exchange. The system requirements are listed in the left column and the vendors who met the requirement are listed in the right column.

Figure 2 summarizes the resulting ETL and data quality components of our technical architecture. Our vendor extracts a set of common data elements into individual LC Data QUEST repositories located at each practice site. Once the data is loaded into the LC Data QUEST repository, it can be shared with researchers to support various research related activities, including cohort discovery, randomized control trials, and comparative effectiveness research. Individual practices can also analyze their own repository data to target quality improvement initiatives or to support any individual practice based activity, using a registry tool licensed by the vendor. Data are owned by the individual partner sites and no data are shared to outside collaborators unless explicitly approved by the site.

At the bottom of the local data loop, a program generates a point-of-care report that includes CDS for recommended national guidelines of care. Patients and practitioners review the point-of-care report during visits and can correct data errors.

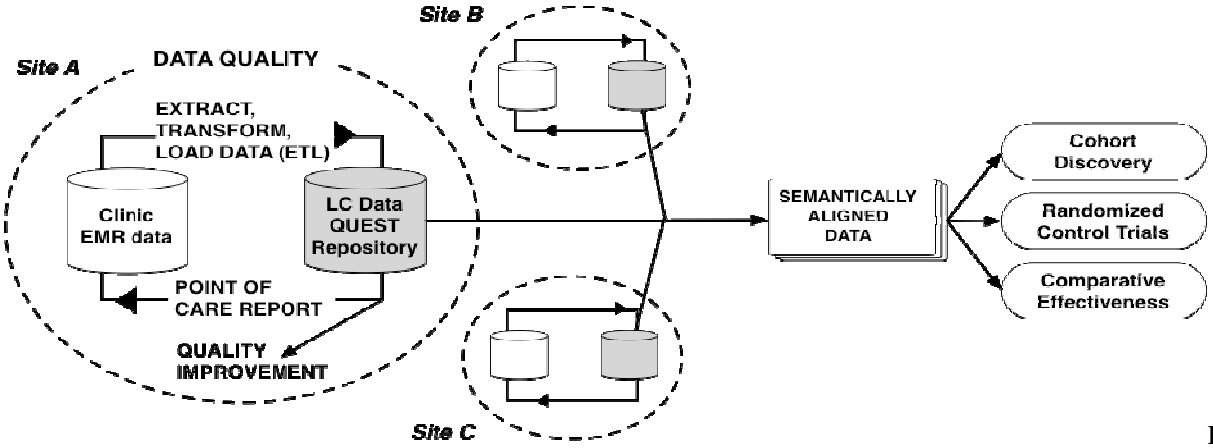

Figure 2. LC Data QUEST technical system architecture illustrating ETL and data quality management activities. At each practice, a standard set of EMR data elements are batched daily into a local repository. The repository supports generation of point-of-care decision support reports and quality improvement queries. Practitioners can identify errors in the EMR via the point-of-care reports and develop workflow processes to correct the data. This architecture repeats at each practice site. Site data can be semantically aligned and combined for health research.

Our initial federated data sets included variables that support management of common guideline supported diseases, although the design is scalable for expansion to other domains that sites find desirable or are needed by data sharing projects in the future. However, as an initial proof-of-concept, providing decision support using national clinical guidelines was of immediate benefit to practices, while the extracted data allowed us to test our data quality and metadata management strategies. Two of the six pilot sites have implemented ETL, with two in the installation process, and two with some delay due to administrative and technical issues (i.e., governance requirements and EMR migrations). LC Data QUEST supports three funded research projects outside of the initial pilot funding, with several projects in development and growing collaborations.

A ,method for managing inventory and the complex set of shared clinical data has led to the development of the Federated Information Dictionary Tool (FInDiT), specified to catalogue type, quantity, and quality of the data that are available across the LC Data QUEST data sharing architecture. This design allows for easy addition of future sites by: 1) defining a set extract format for aggregated data content and metadata needed from any additional federated repository wishing to be added; 2) allowing for simple upload of this extract into a SQL database; and 3) dynamic access to the data via the web-based front-end graphical user interface.

Discussion

Data sharing architectures built in the context of primary care are often driven by project specific needs (i.e., specific randomized control trials, comparative effectiveness research, etc.).[1] Our CTSA built this pilot project as a proof of concept infrastructure to facilitate bridge building and engage communities in translational research. This infrastructure was intentionally proposed without a defined research project to support our core philosophy of engaging communities in effective collaboration as a strategy for facilitating translational research. Critical to this process was delivering immediate benefit to our partners. Therefore we created an architecture and set of tools customized to the needs of our community partners. Finding and engaging partners was our primary focus during the initial phases of this project and occurred simultaneously with the technical evaluations.

In building a data sharing network *de novo*, we first needed to establish a foundation of trust, which involved time intensive interactions with partner sites and between our internal partners within the ITHS. These interactions were crucial to iterating our technical requirements and developing the toolsets needed. Evaluation of individual clinical work flows, data standards, and data semantics was necessary to build effective ETL processes. Engaging a vendor to partner with our informatics team who brought experience in understanding diverse primary care environments and EMR technologies was not only essential to developing our ETL process, but facilitated stronger engagements with our partners.

LC Data QUEST aimed to promote comparative effectiveness research capacity within community settings through increasing data sharing capacity. Use of a point-of-care CDS tool supported the bi-directional process of translational research and mission of the CTSA to deploy proven biomedical

applications and knowledge into clinical practice.[15] The CDS tool not only provided natural data quality upkeep, but could also be used as a vehicle to facilitate practice change that can directly improve patient care and outcomes.

Vendors who develop CDS tools are incented to keep current with national guidelines, extract and align data sets meaningful to practices for their own efficiency, and ensure their tools and services are financially sustainable. Partnership with a vendor permitted us to take advantage of their expertise efficiently and cost-effectively. However, empirical evaluation is needed to assess return on investment for practices, given fee structures long term would need to be maintained by the individual partner sites, rather than via grant funding. Evaluation efforts are underway to quantify return on investment with our pilot sites.

Metadata management tools are needed to facilitate linkages between research and practice communities. The CTSA program is in the unique position to build bridges across practices and communities by creating research networks and developing a metadata management strategy across sites. Maintaining high quality metadata is the basis for collaborations with local quality improvement officers, researchers, and other research networks and CTSAs who share common interests and priorities with LC Data QUEST practices and communities. Collaborations with other primary care based data sharing networks is crucial for aligning efforts nationally and promoting utility of these networks for research and dissemination of practice standards. FInDiT will be instrumental in communicating and facilitating collaborations between research and community based practice partners by offering detailed, meaningful information needed to understand the data sharing capacity across our sites and to attract researchers to use our architecture.

Key lessons learned from our experience with LC Data QUEST include the importance of sustainability and growth when building data sharing networks in practices and tribal communities. To achieve sustainability, solutions must be cost effective to be financially viable. It was also essential that we bring immediate benefit to our partners to entice engagement, rather than propose a passive model of data sharing. Practice sites have large barriers to engaging in data sharing efforts such as resource contention or psychologically based reluctance due to harms from research practices in the past. Bringing immediate benefit and ensuring local control over the data was needed to overcome these barriers. Assessing and accounting for governance/ownership issues and physical location of data as potential barriers to access are important considerations, given they can complicate timelines, costs, and effort.

Next Steps

With our established architecture, we are expanding our data sharing network beyond the original six sites through current funded research projects and future grant proposals. We are partnering with other primary care based research networks and CTSAs to explore end-user tools and metadata management, as well as alignment strategies across networks to include sites working outside our vendor model.

Developing methods and user application layers for sharing data securely, responsibly, and respectfully will be a key consideration as we continue to evolve this architecture to include front end data sharing tools. Governance issues across a federated system are complex and must be carefully considered within system requirements to maintain trust and participation by partners. CTSAs are well-suited for developing end-user tools for data sharing as they have incentive and skills to facilitate collaborations with communities and engage academic researchers. Using an iterative process with our community and practice based partners is crucial to developing end-user system requirements for front-end tools that support streamlined access to data across sites.

Conclusion

We have presented a technical architecture for community-based practice data sharing across disparate primary care settings. Our system architecture design was a result of partnerships between multiple stakeholders including our CTSA, community practices and tribal partners, and national research communities. Developing the LC Data QUEST data sharing architecture involved significant time and effort in creating and sustaining relationships among all partners involved and required an iterative process to allow stakeholders to give valuable input into system requirements. Our LC Data QUEST data sharing architecture met three of the four primary technical components of data sharing: 1) ETL; 2) data quality; and 3) metadata management. The field of biomedical informatics remains challenged with developing solutions for the fourth primary technical component, end-user tools that can support secure data sharing across the dense, rich datasets available in ambulatory care based EMR systems. Developing data sharing architectures involve a complex socio-technical confluence requiring tangible immediate benefits for each stakeholder to ensure participation, sustainability, and scalability.

Acknowledgements

The research was supported by Grant Number 1 UL 1 RR 025014-01 from the National Center for Research Resources, NIH and Training Grant T15LM07442 from the National Library of Medicine, NIH.

References

1. Pace WD, Cifuentes M, Valuck RJ, Staton EW, Brandt EC, West DR. An electronic practice-based network for observational comparative effectiveness research. Ann Intern Med. 2009 Sep 1;151(5):338-40.
2. Murphy SN, Weber G, Mendis M, al. e. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010;17(2):124-30.
3. Vogt TM, Elston-Lafata J, Tolsma D, Greene SM. The role of research in integrated healthcare systems: the HMO Research Network. Am J Manag Care. 2004;10(9):643-8.
4. Lin K, Daemer G. caBIG Security Technology Evaluation White Paper: National CAncier Institute. 2006;January 23, 2006.
5. Pace WD, West DR, Valuck RJ, Cifuentes M, Staton EW. Distributed Ambulatory Research in Therapeutics Network (DARTNet): summary report. Rockville, MD: Agency for Healthcare Research and Quality, 2009:1-36.
6. Zerhouni EA. US biomedical research: basic, translational, and clinical sciences. JAMA. 2005 Sep 21;294(11):1352-8.
7. Lin CP, Black RA, LaPlante J, et al. Facilitating Health Data Sharing Across Diverse Practices and Communities. 2010 AMIA Summit on Clinical Research Informatics. San Francisco, CA: AMIA, 2010.
8. Sen A, Sinha AP. A comparison of data warehousing methodologies. Commun ACM. 2005;48(3):79-84.
9. Carlo B, Cinzia C, Chiara F, Andrea M. Methodologies for data quality assessment and improvement. ACM Comput Surv. 2009;41(3):1-52.
10. Roth CP LY, Pevnick JM, Asch SM, McGlynn EA. The challenge of measuring quality of care from the electronic health record. Am J Med Qual. 2009 Sep-Oct;24(5):385-94.
11. Vogt TM, Elston-Lafata J, Tolsma D, Greene SM. The role of research in integrated healthcare systems: the HMO Research Network. Am J Manag Care. 2004 Sep;10(9):643-8.
12. Peterson KA, Fontaine P, Speedie S. The Electronic Primary Care Research Network (ePCRN): A New Era in Practice-based Research. J Am Board Fam Med. 2006 January 1, 2006;19(1):93-7.
13. Anderson N, Chilana, P, Tarczy-Hornoch, P. Challenges of Implementing Anonymized Cross-Institutional Federated Querying for Clinical Translational Research. Computer Human Interaction Conference (CHI). Boston, 2009.
14. Stephens KA, Anderson N, Lin C. Developing best practices for evaluating federated data sharing: Approaches from academic hospital and primary care clinic networks. Annual Meeting of the American Medical Informatics Association. Washington, DC, 2010, November.
15. Reis SE, Berglund L, Bernard GR, et al. Reengineering the National Clinical and Translational Research Enterprise: The Strategic Plan of the National Clinical and Translational Science Awards Consortium. [Miscellaneous]. (1040-2446).