Frey LJ et al. J Am Med Inform Assoc 2015;0:1–7. doi:10.1093/jamia/ocv009, Case Report

Virtualization of Open-Source Secure Web Services to Support Data Exchange in a Pediatric Critical Care Research Network

RECEIVED 4 November 2014
REVISED 12 January 2015
ACCEPTED 21 January 2015

Lewis J Frey[1], Katherine A Sward[2], Christopher JL Newth[3], Robinder G Khemani[3], Martin E Cryer[4], Julie L Thelen[5], Rene Enriquez[5], Su Shaoyu[5], Murray M Pollack[6], Rick E Harrison[7], Kathleen L Meert[8], Robert A Berg[9], David L Wessel[10], Thomas P Shanley[11], Heidi Dalton[12], Joseph Carcillo[13], Tammara L Jenkins[14], J Michael Dean[15]

ABSTRACT

Objectives To examine the feasibility of deploying a virtual web service for sharing data within a research network, and to evaluate the impact on data consistency and quality.

Material and Methods Virtual machines (VMs) encapsulated an open-source, semantically and syntactically interoperable secure web service infrastructure along with a shadow database. The VMs were deployed to 8 Collaborative Pediatric Critical Care Research Network Clinical Centers.

Results Virtual web services could be deployed in hours. The interoperability of the web services reduced format misalignment from 56% to 1% and demonstrated that 99% of the data consistently transferred using the data dictionary and 1% needed human curation.

Conclusions Use of virtualized open-source secure web service technology could enable direct electronic abstraction of data from hospital databases for research purposes.

Key words: electronic health record; secure web services; grid; virtualization; pediatric critical care; data governance; pediatric network; virtual machines; learning health care system

CASE REPORT

Enhancing the Learning Health care System (LHS) is a national goal,[1] where data obtained during care and operations contribute to the development of knowledge and, in turn, translates into evidence-based practice and health system improvements. Efforts to advance the LHS have accelerated through the establishment of PCORnet[2] (www.pcornet.org) via the Patient Centered Outcomes Research Institute with the promise to deliver a national network for clinical outcomes research. The objective is to improve outcomes by leveraging electronic health records (EHRs) as a national research resource. However, to do so, networks need to resolve issues of data governance, central for the LHS and a crucial aspect of interoperability,[3,4] so that large scale collaborative initiatives are not overwhelmed by the diversity and multitude of clinical data along with competing stakeholder interests.[5] Success of these and other LHS initiatives will be measured by the speed with which new networks establish scalable multisite collaborations for observational studies[6] while not being overcome by data governance and interoperability issues. We present a case report on the design, deployment, and initial evaluation of a federated infrastructure for a national pediatric network based on a virtual machine (VM) framework that can be used in a LHS to addresses data governance and interoperability.

Background

Data sharing via federated networks for the purpose of conducting clinical studies include costly, complex, time consuming, and potentially error-prone activities. Even after creating written protocols, clinicians must dedicate substantial time and effort collecting data and communicating to achieve data consistency.[7] Data governance is a disciplined method through which resources are formally managed with a focus on data consistency and quality.[8] An approach that can improve data governance has multiple implications for clinical research, including an emphasis on the interoperability of clinical data and the need for rigorous approaches to ensure the utility and validity of clinical data used for research purposes.[9] Interoperability, composed of both syntactic and semantic components, has been identified as a mechanism to support national health systems initiatives.[10] Interoperability supports collaboration between organizations,[11] having demonstrated beneficial impact on clinical trials, EHR information

Correspondence to Dr Lewis J Frey, Associate Professor, Biomedical Informatics Center, Department Public Health Sciences, Medical University of South Carolina, 135 Cannon Street, Suite 405K, MUSC 200, Charleston, SC 29425, USA; frey@musc.edu

For numbered affiliations see end of article.

systems,[12,13] and maintenance of patient data across health care organizations.[14] There are multiple approaches to data governance and interoperability, ranging from technologies that are data model agnostic to locally focused models to globally constrained models.

The Mini-Sentinel network[15,16] deploys a distributed query architecture based on submit-run-return procedures using the data model agnostic PopMedNet[17] (www.popmednet.org) technology. The Mini-Sentinel approach does not employ a centralized database; instead, members of the network are responsible for their own data that are linked by PopMedNet. Queries are broadcast to PopMedNet client sites and are then reviewed and run by site-level data stewards. The stewards review the results and securely return them to the requesting investigator. Multiple networks within PCORnet use the PopMedNet query architecture.[18,19]

The Scalable Collaborative Infrastructure for a Learning Healthcare System (SCILHS) was designed to avoid limitations of global top-down solutions by using locally focused solutions with a vibrant user and developer base.[20] Components of SCILHS's PCORnet instantiation include the widely adopted Shared Health Research Information Network (SHRINE)[21] and Informatics for Integrating Biology and the Bedside (i2b2).[22] The i2b2 system stores data in a locally informed star schema relational model to simplify query strategies.[22] The SHRINE is composed of an aggregator/adaptor model that broadcasts queries to adaptors at multiple sites.

The Translational Research Informatics and Data Management Grid (TRIAD)[23] uses a grid approach to create a homogenous view of multisite data sources through a repository of object models, thus supporting a global view instead of a strictly local one. TRIAD integrated the caGrid middleware,[24,25] a global model technology developed for the Cancer Biomedical Informatics Grid (caBIG).[26] A grid architecture approach is essentially a federated collection of heterogeneous and geographically dispersed information systems. The dangers of integrating the caGrid architecture include limited adoption, lack of end-user facing applications, external dependencies on National Cancer Institute systems, and concerns of scalability of the technology.[23] The demonstrated ability to support data governance in federated applications[12] and securely manage data transmission are strengths of the grid architecture. A method for easily deploying the caGrid infrastructure was elusive, and the complexity of the grid architecture increased the difficulty of deployment. Hence, the traditional resource intensive process was cost prohibitive for all but a handful of well-funded projects.

We believed that the costs and complexity of grid deployment could be overcome through virtualization technology. In this paper, we describe the development and assessment of open-source secure web services, built using the caGrid middleware used by TRIAD,[23] deployed at 8 children's hospitals in the Collaborative Pediatric Critical Care Research Network (CPCCRN).[27] A root problem in the traditional caGrid deployment was the tight coupling of the local data sources, such as administrative databases or EHRs, with the complex web service. Each site needed to link their data elements to interoperable components and then build a grid service. This was a complex, multistep process that required collaboration between domain experts, informaticists, and software engineers. To overcome this, we developed a new approach using VMs running the full caGrid technology stack, along with a database (the shadow database) that served as a limited version of an institution's data.

METHODS

Using caGrid tools, experienced informaticists and domain experts built a platform-independent virtualized secure web service called the Pediatric Intensive Care Unit Grid (picuGrid). The system was designed to support an ongoing CPCCRN observational study called the Core Clinical Data Project (CCDP), conducted with Institutional Review Board approval and data sharing agreements. The data consist of descriptive elements such as patient demographics, length of hospital and Pediatric Intensive Care Unit (PICU) stays, procedure codes, and diagnosis codes. The data are typically aggregated on an annual basis and describe the characteristics of PICU stays at the CPCCRN Clinical Centers. Annual CCDP data supports hypothesis generation, preliminary power analyses, and patient recruitment projections for CPCCRN studies.[27]

The virtualized system was developed, tested, and deployed at the CPCCRN Data Coordinating Center (DCC) at the University of Utah. The system, depicted in figure 3, was then deployed at the following 8 CPCCRN Clinical Centers: Children's Hospital Los Angeles, Children's Hospital of Michigan, Children's Hospital of Philadelphia, Children's National Medical Center, Mattel Children's Hospital at UCLA, Phoenix Children's Hospital, CS Mott Children's Hospital at the University of Michigan, and University of Pittsburgh Medical Center. We evaluated consistency and data quality of the data set to assess potential effects of the picuGrid virtualized environment.

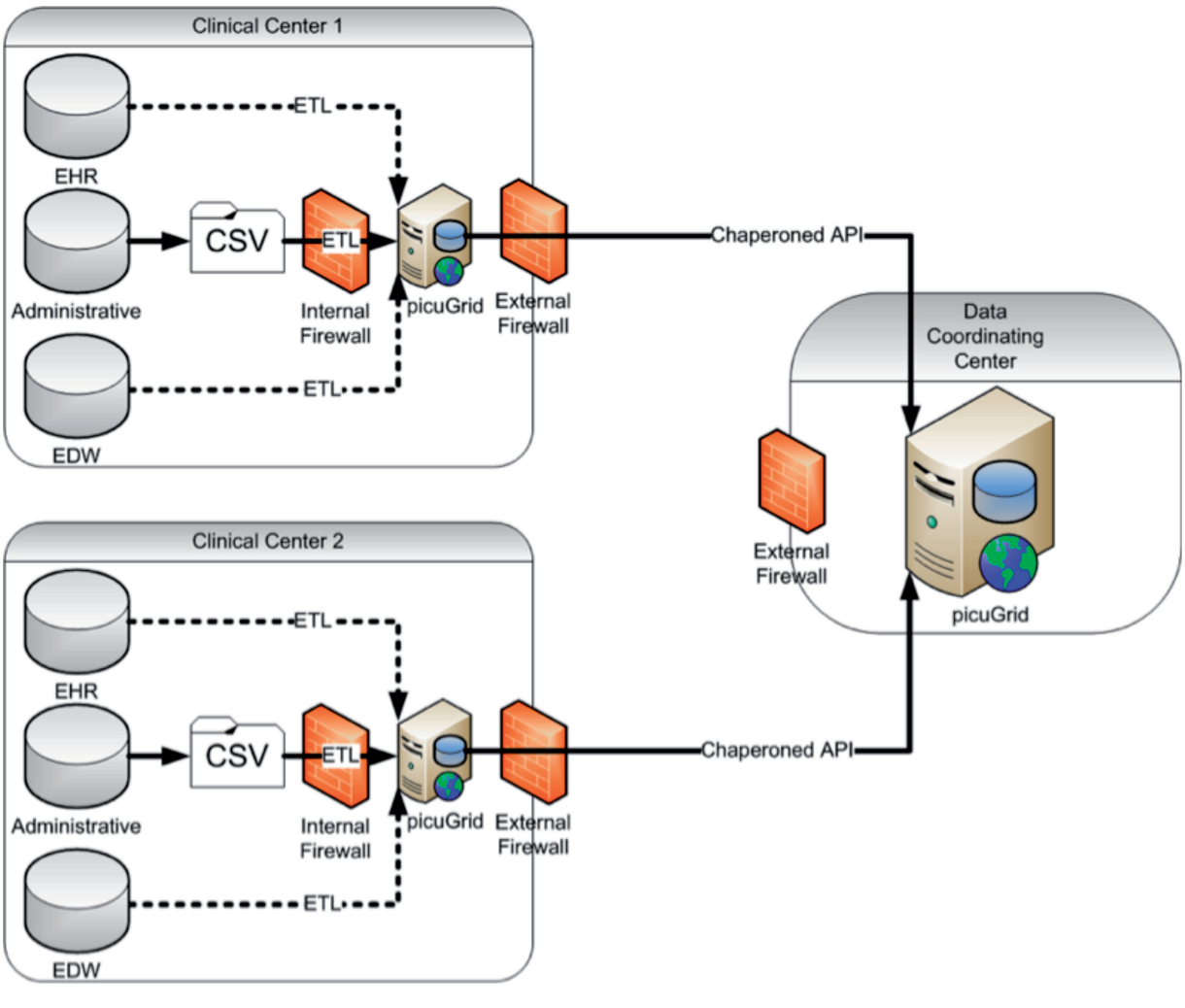
Evaluation

Since the extraction, transform, and load (ETL) process that populates the shadow database used the 2011 CCDP data set, we assessed the process using the 2012 data obtained via picuGrid in parallel with the traditional CCDP file submission process. The total evaluation set across the 8 sites consisted of 18 551 rows. A row was flagged by the ETL process as having an error if the row did not load properly because of at least 1 of the 54 fields having a format inconsistency (ie, format misalignment) and/or a nonvalid dictionary value.

RESULTS

The results focus on (1) an overview of the picuGrid implementations, (2) comparing field format misalignments using the traditional approach with field format misalignments using the picuGrid approach, (3) examining levels of curation needed to load the 2012 data into the shadow database (reflecting the extent to which ETL scripts need to be modified), and (4) an analysis of scalability of the VM client and server grid architecture.

Figure 1: The picuGrid architecture was designed using a chaperoned Application Program Interface (API); firewall settings were controlled by the centers with picuGrid being instantiated between the external and internal firewall of the site, and local IT departments could set additional security restrictions to limit connections to the VM. Secure data transmission between the sites and the DCC was enforced through caGrid credentials within each VM that were validated by a third party credentialing service. Unlike traditional grid architecture, we limited the system so that only the DCC could access data and clinical sites and the other Clinical Centers could not view or access other sites. All data up to and including the shadow database were under the direct control of the local site personnel. The shadow database had a dictionary table for updating value sets for each site. The DCC could pull data using the chaperoned API but could not access the shadow database directly. The solid arrow shows data pulled from the administration database to a comma separated values (CSV) file and then pushed past the internal firewall and into the picuGrid shadow database. Many clinical research studies use data from the active EHR or the enterprise data warehouse (EDW). Pulling data such as laboratory test results or vital signs would be beneficial to most of the network clinical studies. The dotted lines from the EHR and EDW represent those desired future data sources. Since each site has a MySQL database, the training needed to access the data is the standard querying of databases (ie, Structured Query Language, SQL). Each site received a user guide to facilitate installation and support of the system.



CASE REPORT

picuGrid deployment overview

The picuGrid VMs were successfully deployed to all CPCCRN site hospitals; deployment took roughly 3 hours for each site. Establishing the ETL process to load 2011 data into the shadow database ranged from 1 to 4 hours per site. The sites reused the 2011 ETL processes and did not require a new ETL script for the 2012 CCDP data set. Successful secure data transfer was demonstrated for all 8 CPCCRN sites.

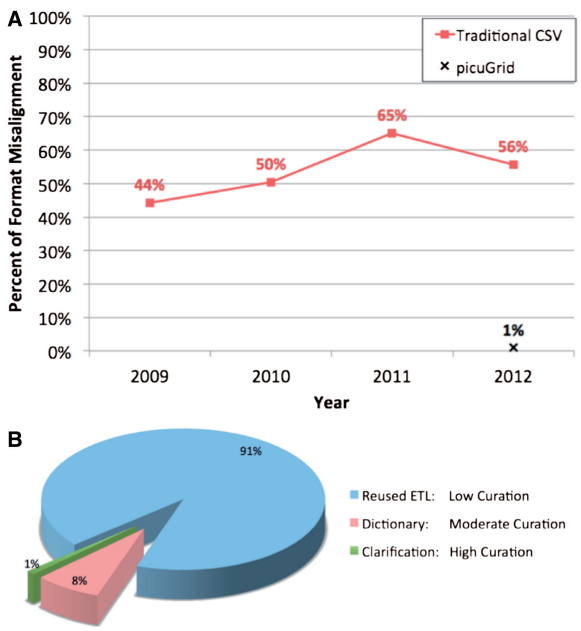


Figure 2: (A) A field was defined as being misaligned if at least 1 row had incorrectly formatted values. Format misalignments were measured as the percentage of incorrectly formatted fields out of a total of 54 fields specified by the research protocol. (B) We assisted sites to load their 2012 data into the picuGrid system. If the row of 2012 data loaded with no ETL process content errors, then the record was counted as "Reused ETL: Low Curation." If the dictionary table in the picuGrid shadow database needed to be updated to account for a new value, then the record was counted as "Dictionary: Moderate Curation." If a human needed to clarify and potentially change the data in the data file, then the record was labeled as "Clarification: High Curation." We assisted 1 site in reconfiguring their ETL process. This change was necessary due to a field being conjoined from 2 fields in the 2012 data set. The fields were separated and loaded through a simple change to the ETL process.

Syntactic and semantic interoperability: field format misalignments and data curation

A field format misalignment occurred when data were submitted in a format that differed from that specified by the research protocol. Figure 2A displays 4 years of format misalignment data for the 8 Clinical Centers, for their traditional submission process (solid line with square symbols). Over the 4 years, the average format misalignments drift upward from 44% to 65%.

The picuGrid's web service resulted in substantially reduced format misalignments (1% for 2012 across all Centers, "X" symbol on figure 2A). That reduction primarily resulted from the picuGrid ETL process correcting formatting before the data file was loaded into the database and submitted to the DCC.

We categorized rows in the dataset as needing low, moderate, or high levels of curation depending upon the maximum level of curation associated with any field in the row (figure 2B). Using the 2011 picuGrid ETL process, 91% of the rows in the 2012 data sets loaded correctly into the shadow database. An additional 8% of the rows of data needed moderate levels of curation that were addressed by updating the dictionary. Specifically, fields for race, ethnicity, admission type, discharge disposition, and payer had new values that were not in the 2011 dictionary. Adding the value for a new payer at one site accounted for 2% of the total data being correctly loaded, and updating an ethnicity value at another site accounted for 4% of the total data. The final 1% required clarification related to diagnosis codes and zip codes; values had to be examined by a human curator and corrected.

Virtual system scalability

A concern with the grid architecture is its ability to scale.[23] Given an open-source code base that can be replicated as needed without licensing costs, an advantage is the ability to cost-effectively create multiple clients and servers to scale the amount of data that can be queried. Figure 3 demonstrates the ability to improve response time through parallelization by increasing the number of virtualized servers and clients[28] for the picuGrid architecture.

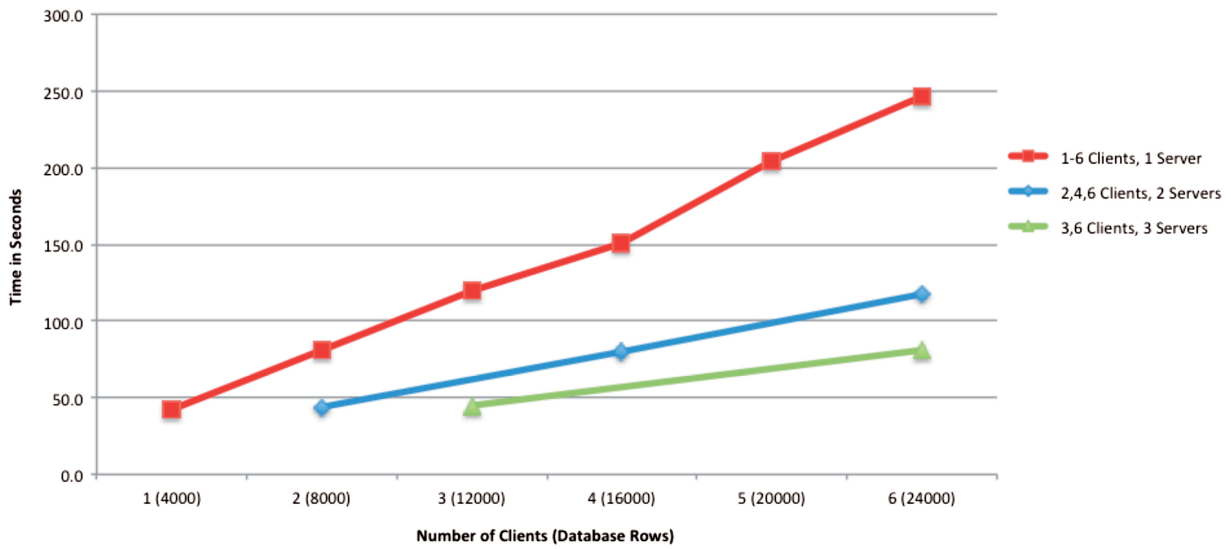
DISCUSSION

We successfully leveraged machine virtualization to ease the deployment of complex grid technologies. The virtualized picuGrid system reduced the deployment time from months to hours, thus allowing hospital deployment teams to have, within minutes, a fully operational secure grid web service. Traditional caGrid implementation has been hampered by the direct connection between the logical model and the hospital databases as well as the informatics expertise required to generate the Application Program Interface (API). Virtualization and using a shadow database enabled us to eliminate many of the informatics requirements at the individual hospitals. By decoupling the web service from direct interactions with the organization's data, we also decoupled the need for the local hospital information technology team to learn the complex traditional grid deployment processes. In picuGrid, the API was centrally developed and eliminated the need for hospital technology teams to learn how to navigate the complex ecosystem of caGrid applications.

Limitations

This was a single point in time feasibility evaluation. While significant benefits have been hypothesized, the challenges and benefits of directly transferring data to the DCC from the Clinical Center environments were not evaluated. Although many of the caBIG tools were integrated into the new National Cancer Informatics Program and remain available as open-source code, future sustainability could be problematic with the retirement of the caBIG program in 2012; an alternative for sustainability is the public private partnership of TRIAD.[29]

Figure 3: The red line is the time for whole table queries to return for 1 to 6 virtual clients requesting data from 1 virtual server with each client requesting 4000 rows of data. The time increases linearly with the number of clients. The blue line consists of 1, 2, or 3 pairs of virtual clients requesting 8000 rows of data from 2 virtual servers. The green line consists of 1 or 2 virtual client triplets requesting 12 000 rows of data from 3 virtual servers. There is an initial cost to establishing the caGrid connection of around 50 seconds, which is a time lag that is more acceptable for picuGrid's batch architecture instead of a real-time system.



CONCLUSIONS

Using virtualization and open-source software, we were able to quickly and easily deploy a complex technology solution. We demonstrated the feasibility of securely moving data within the CPCCRN research network. Using semantically and syntactically interoperable secure web services, we showed potential improvements in data quality and data governance implications for LHS and PCORnet implementations.

Multisite research networks typically implement complex study protocols that involve abstraction of extensive data, including laboratory values, vital signs, demographics, medication, and study-specific data, with reentry of those values into a research database. The abstraction and data reentry process requires personnel and time and contributes to the costs of clinical observational and interventional studies. Use of virtualized secure web service technology with strong data governance could enable direct electronic data abstraction from hospital databases and speed adoption of the Learning Healthcare System.

CONTRIBUTORS

LJF was responsible for the conception, design, implementation, and analysis of the work. KAS contributed to the acquisition and consistency of annotated data for the work. CJLN and RGK facilitated the initial deployment of the system. MEC analyzed federated query system performance for the work. JLT, RE, and SS assisted in implementation and deployment of the system. JMD was the network champion for the work. The remaining authors assisted in data collection and were site champions for the work. All authors contributed to the writing of the manuscript and approved the final version.

FUNDING

This work was supported by the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, Department of Health and Human Services, through cooperative agreements (U10HD050096, U10HD049981, U10HD063108, U10HD063106, U10HD063114, U10HD049983, U10HD050012, and U01HD049934).

COMPETING INTERESTS

None.

ACKNOWLEDGEMENTS

The authors thank Jamie Bell, Drew DeMarco, Jeri Burr, and Emily Stock for their tireless efforts to support the picuGrid project and the DCC's mission of improving care for critically ill and injured children. We thank Ron Price and Derek Huth from the University of Utah Center for High Performance Computing for answering questions about the National Cancer Institute caBIG architectural infrastructure. We would also like to acknowledge the following champions and technical teams within the network: Carol Nicholson, John Berger, Athena Zuppa, Alan Abraham, David Benigni, Alan Korey, James Zullo, Brad Ottoson, Anthony Geoffron, Paul Vee, Jeni Kwok, Nimesh Patel, James Maszatics, Padma Vinukonda, Lihshwu Ke,

James Law, Dawn Brown, Shirley Fan, Moni Weber, David Higginson, Jesse Perlmutter, Jean Reardon, Amjad Chaudhry, Bill Lawrence, Robert Dennis, Patrick Alger, Kaiding Zhu, Ann Pawluszka, Paula Tank, Jerry Lowetz, Gregory Davis, Mary Ann Diliberto. and Robert Grundmeier.

REFERENCES

1. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. *Sci Transl Med.* 2010;2(57):57cm29.
2. 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;21:578–582.
3. Garde S, Knaup P, Hovenga EJ, Heard S. Towards semantic interoperability for electronic health records–domain knowledge governance for open EHR archetypes. *Methods Inf Med.* 2007;46(3):332–343.
4. Wollersheim D, Sari A, Rahayu W. Archetype-based electronic health records: a literature review and evaluation of their applicability to health data interoperability and access. *Health Inf Manag J.* 2009;38(2):7–17.
5. Adams L. Stewardship and governance in the learning health system. In: Institute of Medicine. Grossmann C, Powers B, McGinnis JM, eds. *Digital Infrastructure for the Learning Health System: The Foundation for Continuous Improvement in Health and Health Care: Workshop Series Summary*. Washington, DC: National Academies Press; 2011.
6. Collins FS, Hudson KL, Briggs JP, Lauer MS. PCORnet: turning a dream into reality. *J Am Med Inform Assoc.* 2014;21:576–577.
7. Sachdeva S, Bhalla S. Semantic interoperability in standardized electronic health record databases. *J Data Inf Qual.* 2012;3(1):1.
8. Ladley, J. *Data Goverance: How to Design Deploy, and Sustain an Effective Data Governance Program*. Boston, MA: Morgan Kaufmann Publishers; 2012.
9. Institute of Medicine (US) Roundtable on Evidence-Based Medicine. Olsen LA, Aisner D, McGinnis JM, eds. *The Learning Healthcare System: Workshop Summary*. Washington, DC: National Academies Press; 2007.
10. Hovenga, EJS. Importance of achieving semantic interoperability for national health information systems. *Texto Contexto-Enferm.* 2008;17(1):158–167.
11. Janssen M, Scholl HJJ. Interoperability for electronic governance. In: *ICEGOV '07: Proceedings of the 1st International Conference on Theory and Practice of Electronic Governance, Macao, China, December 10-13*. New York, NY: ACM; 2007:45–48.
12. Davies J, Gibbons J, Harris S, Crichton C. The CancerGrid experience: metadata-based model-driven engineering for clinical trials. *Sci Comput Program.* 2014;89:126–143.
13. Martínez-Costa C, Menárguez-Tortosa M, Fernández-Breis JT, Maldonado JA. A model-driven approach for representing clinical archetypes for Semantic Web environments. *J Biomed Inform.* 2009;42(1):150–164.
14. Sachdeva S, Bhalla S. Semantic interoperability in healthcare information for EHR databases. In: *Databases in Networked Information Systems.* Berlin, Germany: Springer; 2010:157–173.
15. Platt R, Carnahan R. The US Food and Drug Administration's Mini-Sentinel Program. *Pharmacoepidemiol Drug Safety.* 2012;21(S1):1–303.
16. Psaty BM, Breckenridge AM. Mini-sentinel and regulatory science-big data rendered fit and functional. *N Engl J Med.* 2014;370(23):2165–2167.
17. Toh S, Platt R, Steiner J, Brown J. Comparative-effectiveness research in distributed health data networks. *Clin Pharmacol Ther.* 2011;90(6):883–887.
18. McGlynn EA, Lieu TA, Durham ML, *et al*. Developing a data infrastructure for a learning health system: the PORTAL network. *J Am Med Inform Assoc.* 2014;21:596–601.
19. Kho AN, Hynes DM, Goel S, *et al*. CAPriCORN: Chicago Area Patient-Centered Outcomes Research Network. *J Am Med Inform Assoc.* 2014;21:607–611.
20. Mandl KD, Kohane IS, McFadden D, *et al*. Scalable Collaborative Infrastructure for a Learning Healthcare System (SCILHS): architecture. *J Am Med Inform Assoc.* 2014;21:615–620.
21. Weber GM, Murphy SN, McMurry AJ, *et al*. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. *J Am Med Inform Assoc.* 2009;16(5):624–630.
22. 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):124–130.
23. Payne P, Ervin D, Dhaval R, Borlawsky T, Lai A. TRIAD: The Translational Research Informatics and Data Management Grid. *Appl Clin Inform.* 2011;2(3):331–344.
24. Oster S, Langella S, Hastings S, *et al*. caGrid 1.0: a grid enterprise architecture for cancer research. *AMIA Annu Symp Proc.* 2007:2007:573–577.
25. Saltz J, Hastings S, Langella S, *et al*. A roadmap for caGrid, an enterprise Grid architecture for biomedical research. *Stud Health Technol Inform.* 2008;138:224–237.
26. Buetow KH, Niederhuber J. Infrastructure for a learning health care system: CaBIG. *Health Aff.* 2009;28(3):923–924.
27. Willson DF, Dean JM, Newth C, *et al*. Collaborative Pediatric Critical Care Research Network (CPCCRN). *Pediatr Crit Care Med.* 2006;7(4):301–307.
28. Cryer ME. *Hybrid Agent-Based Modeling of Healthcare Surveillance Networks* [dissertation]. Salt Lake City: The University of Utah; 2013.
29. Payne PR. Sustainability through technology licensing and commercialization: lessons learned from the TRIAD Project. *eGEMs.* 2014;2(2):2.

AUTHOR AFFILIATIONS

[1]Biomedical Informatics Center, Department Public Health Sciences, Medical University of South Carolina, Charleston, USA

[2]College of Nursing; Department of Biomedical Informatics, University of Utah, Salt Lake City, USA

[3]USC Keck School of Medicine; Department of Anesthesiology and Critical Care Medicine, Children's Hospital Los Angeles, Los Angeles, USA

[4]Department of Biomedical Informatics, University of Utah School of Medicine, Salt Lake City, USA

[5]Department of Pediatrics, University of Utah School of Medicine, Salt Lake City, USA

[6]Phoenix Children's Hospital, Department of Pediatrics, University of Arizona Phoenix, Phoenix, USA

[7]Department of Pediatrics, University of California at Los Angeles, Los Angeles, USA

[8]Department of Pediatrics, Children's Hospital of Michigan, Detroit, USA

[9]Department of Anesthesiology and Critical Care, The Children's Hospital of Philadelphia, University of Pennsylvania Perelman School of Medicine, Philadelphia, USA

[10]Department of Pediatrics, Children's National Medical Center, Washington, DC, USA

[11]Department of Pediatrics, University of Michigan, Ann Arbor, USA

[12]Department of Child Health, Phoenix Children's Hospital, University of Arizona College of Medicine-Phoenix, Phoenix, USA

[13]Department of Critical Care Medicine, Children's Hospital of Pittsburgh, Pittsburgh, USA

[14]Eunice Kennedy Shriver National Institutes of Child Health and Human Development (NICHD), National Institutes of Health, Bethesda, USA

[15]Department of Pediatrics, Division of Pediatric Critical Care Medicine, University of Utah School of Medicine; NICHD Collaborative Pediatric Critical Care Research Network, Salt Lake City, USA