3-22-2016

Bidirectional Data Collaborations in Distributed Research

Kevin Haynes
HealthCore, Inc., khaynes@healthcore.com

Nandini Selvam
HealthCore, Inc., nselvam@healthcore.com

Mark J. Cziraky
HealthCore, Inc., mcziraky@healthcore.com

Bidirectional Data Collaborations in Distributed Research

Abstract

Introduction: The need for collaborations with bidirectional data exchange within and across distributed research networks has increased.

Currently Existing Activities: This commentary will present currently publically available activities including the Sentinel Initiative, the Patient-Centered Outcomes Research Network (PCORnet), and the NIH Research Collaboratory.

Current Technical and Governance Challenges: Even with the advances made in this arena, several technical and governance challenges remain including the evolution of clinically rich data sources and modes of care, availability of longitudinal data resources through data linkage, and the processes to share data and link data resources while ensuring privacy and proprietary control of data.

Perspective: These activities will require enhanced levels of trust between entities involved in the delivery of healthcare (Trust 2.0) in addition to the trust health plans and health systems have with patients (Trust 1.0). Recent public funding announcements and public access to data resources will likely improve the landscape of bidirectional data collaborations in distributed research.

Disciplines

Clinical Epidemiology | Health Information Technology | Health Services Research

eGEMs
Generating Evidence & Methods
to improve patient outcomes

Bidirectional Data Collaborations in Distributed Research

Kevin Haynes, PharmD; Nandini Selvam, PhD, MPH; Mark J. Cziraky, PharmD[i]

ABSTRACT

Introduction: The need for collaborations with bidirectional data exchange within and across distributed research networks has increased.

Currently Existing Activities: This commentary will present currently publically available activities including the Sentinel Initiative, the Patient-Centered Outcomes Research Network (PCORnet), and the NIH Research Collaboratory.

Current Technical and Governance Challenges: Even with the advances made in this arena, several technical and governance challenges remain including the evolution of clinically rich data sources and modes of care, availability of longitudinal data resources through data linkage, and the processes to share data and link data resources while ensuring privacy and proprietary control of data.

Perspective: These activities will require enhanced levels of trust between entities involved in the delivery of healthcare (Trust 2.0) in addition to the trust health plans and health systems have with patients (Trust 1.0). Recent public funding announcements and public access to data resources will likely improve the landscape of bidirectional data collaborations in distributed research.

[i]HealthCore, Inc.

Introduction

Using health care data in the United States functions similarly to Heisenberg's uncertainty principle—the more precise our access is into deep clinical data (e.g., vital signs, inpatient medication administrations) from a hospital or clinical record, the less breadth of clinical data across health systems (e.g., pharmacy claims, knowledge of clinical encounters across different health systems) we have on an individual, and vice versa. The current state of health care delivery in the United States presents a fragmented system in which patients traverse numerous providers, provider organizations, administrative payers, and patient-reported outcome (PRO) portals over time. This poses problems for advancing our understanding of important health care related activities, but also presents an opportunity to create a new paradigm based upon bidirectional data collaborations to close both the gaps in knowledge and the gaps in care.

There is an increasing need for bidirectional data collaborations across data sources to support regulatory decision-making (safety surveillance), research (comparative effectiveness research: CER), quality improvement, and—ultimately—clinical decision support. A national research network of integrated data is critical to the future success of these broad, overlapping activities. There are precedents across the United States of integrated care delivery systems,[1] and networks designed to specifically address one particular activity (e.g., Sentinel Initiative's ability to respond to safety surveillance).[2] There is no national system established to support all three activities with integrated data provided from patients, clinical data from electronic medical records (EMRs) from individual providers or electronic health records (EHRs) from health systems, and administrative claims data from health plans. The need for integrated research and data networks will continue to grow as drug safety questions will increasingly come to rely on the collection of clinical data, CER will need to track patient outcomes longitudinally, and accountable care organizations will need to evaluate quality outcomes post discharge across multiple health systems.

Any activity involving the sharing or transferring of health data in a bidirectional manner requires increased levels of trust and transparency across organizations to maintain the high standards of privacy, security, and research integrity of personal health information that is expected by patients, providers, health plans, and regulators alike. Maintaining these high standards while simultaneously making innovations in the research environment by building new data sharing techniques will require a new level of trust among health plans and provider systems. Moreover, health plans and providers must remain mindful of the primary trust between themselves and their members or patients.

This commentary discusses the need for further integration across several existing research and data networks to improve both the depth and the breadth of the integrated data across health systems. The commentary illustrates these needs by highlighting the Food and Drug Administration's (FDA) Sentinel Initiative, the Patient-Centered Outcomes Research Institute (PCORI) National Patient-Centered Outcomes Research network (PCORnet), and the National Institutes of Health (NIH) Health Care systems Research Collaboratory. Other examples of distributed data networks are emerging on a regular basis—i.e., the Academy of Managed Care Pharmacy (AMCP) Biologics and Biosimilars Collective Intelligence Consortium (BBCIC) and Reagan-Udall Foundation for the FDA's Innovation in Medical Evidence Development and Surveillance (RUF-IMEDS)—but are not expanded on in this commentary.

Currently Existing Activities

The FDA Sentinel Initiative, and the recently completed pilot program, Mini-Sentinel, were created in response to a congressional mandate in the FDA Amendments Act (FDAAA) of 2007. The Sentinel Initiative is designed to monitor the safety of regulated medical products by utilizing existing electronic health care data from multiple sources, including large repositories of administrative claims submitted by health care providers to insurance companies.[3] This highly collaborative model between academic and private organizations has developed the capacity to rapidly respond to the FDA by performing active surveillance of marketed medical products, including drugs, biologics, and medical devices.[4] The system serves as a model leveraging large sources of administrative data to address important drug safety questions, while preserving privacy by minimizing the transfer of protected health information and proprietary data. Data partners serve as full collaborators in implementing the safety surveillance and retain full autonomy in the control of their data, allowing the partners to determine their participation level in specific surveillance activities. The initiative is also fully transparent, with public access to the specifics of the common data model (CDM) creation and active safety-surveillance system tools.[3] While the Initiative includes approximately 190 million patients in defined enrollment segments, the clinical depth of the data is limited. The majority of the data partners contribute administrative claims data electronically with access to medical records for validation studies. Among the Health Care Systems Research Network (HCSRN) data partners who have a long history of providing integrated data for research, there is diversity and differential access to deep clinical encounter level data (e.g., sites that own their hospitals).[1] The ability to integrate administrative claims data from large health plans with clinical data from EHR systems will advance the ability of Sentinel to respond to medical product safety surveillance.

PCORI, authorized by the Patient Protection and Affordable Care Act (PPACA) of 2010, is tasked with providing evidence to assist patients, clinicians, purchasers, and policymakers to make better-informed health decisions through the conduct of both observational and interventional CER. PCORI launched a funding initiative to support the development and sustainability of PCORnet, a national infrastructure of clinical data research networks (CDRNs) and patient-powered research networks (PPRNs).[5] PCORnet CDRNs are built on a foundation of EHR data from health systems, providing electronic health information recorded during routine patient care.[5] PCORnet PPRNs are registries of patients for the study of specific disease states of interest with the ability to collect PROs for large numbers of patients.[6] The primary goal of phase 1 was to establish the necessary data infrastructure, largely building off of the lessons learned from the CDM developed for Sentinel through specific enhancements for providing deeper clinical data from health systems. PCORnet is entering phase 2, which will expand the network to broader participation. With the infrastructure aspect mostly complete, PCORI announced funding to identify the extent of overlap between patients in CDRNs and members in large health plans—largely to support demonstration projects and enhance network sustainability. PCORnet demonstration projects such as the Aspirin Dosing: A Patient-Centric Trial Assessing Benefits and Long-Term Effectiveness (ADAPTABLE) will require data integration with health insurance plans to capture events that occur at health systems outside of existing CDRN sites. Likewise the Obesity Observational Research Initiative will require access to longitudinal claims to properly characterize exposure and to capture health outcomes. PCORnet

has begun to integrate clinical data, from EMRs and EHRs, with Medicare data. The addition of administrative claims data from large health plans will advance the conduct of research, and ultimately will provide the best evidence to help patients and health care providers make informed decisions. A recent funding announcement demonstrates the recognition of the importance of the involvement of health plans.

The NIH Health Care Systems Research Collaboratory is supported by the Common Fund at the NIH to improve clinical trial conduct by creating a new infrastructure for collaborative health care system research, with the goal of ensuring that health care providers and patients can make decisions based on available clinical evidence.[7] The Collaboratory has been focused on advancing the use of pragmatic clinical trials (PCTs), which seek to compare clinically relevant interventions in diverse patient populations.[8] Currently the sites for the NIH Distributed Research Network (DRN) are different from the sites for pragmatic clinical trials within the Collaboratory. However, administrative claims data from large health plans can facilitate the conduct of large PCTs through the rapid characterization of baseline characteristics of individual patients or provider networks for cluster randomized trials,[9] as well as providing the necessary longitudinal follow-up using consistent data capture within defined enrollment periods. Mini-Sentinel and the Clinical Trials Transformation Initiative (CTTI) have proposed a framework for conducting clinical trials within a DRN.[10]

In summary, the goals—of Sentinel, to provide an active safety surveillance system for the FDA; of PCORnet, to provide an infrastructure to conduct patient-centered CER for PCORI stakeholders; and of the NIH Collaboratory, to provide a rich infrastructure to conduct clinical research—overlap in the need for longitudinal, integrated clinical data. Additionally,

AMCP's BBCIC[11] and RUF-IMEDS[12] are further examples of the growth in distributed database research. Legislative initiatives like the 21st Century Cures Act continue to promote the innovations necessary to improve research collaborations.[13]

Current Technical and Governance Challenges

The integration of large administrative health care data with the clinical data repositories of Sentinel, PCORnet CDRNs and PPRNs, and the NIH Collaboratory represents a first step in providing integrated data in a national infrastructure. Even with the continued development of PCORnet, a large number of health care delivery systems will remain outside of existing PCORnet CDRNs. Conversely, many smaller regionally based health plans remain outside of the Sentinel Initiative. Current challenges include the integration of clinically rich data sources with available longitudinal administrative data resources through data linkage, and the technical and governance processes to share data and link data resources while ensuring privacy and proprietary control of data. Table 1 highlights some of the technical and governance challenges.

Technical Challenges

The first technical challenge will be to keep pace with the evolution of health care delivery and the availability of data. For example, individuals obtain care through telemedicine programs, urgent care facilities, and clinical care within pharmacies.[14] These clinical encounters may or may not be processed as administrative claims, or transmitted on health information exchanges, which would not allow these encounters to be observed in current distributed research environments. New sources of data will also be introduced that did not previously exist or were not able to be integrated into retrievable information systems (e.g., patient-recorded biometric data). Distributed research networks (DRNs) will need to

Table 1. Governance and Technical Challenges

	SENTINEL	PCORNET	NIH COLLABORATORY
Longitudinal capture of clinical encounters over defined periods	+++	+*	++
Depth of clinical data	++*	+++	++*
Patient-reported outcomes (PROs)	+*	+++	++*
Public health authority	+++	–	–
Data standardization	+++	+++	+++

*Exception: Participating integrated delivery systems, such as Kaiser Permanente, differ from large administrative health plans in their ability to provide longitudinal capture of clinical encounters over defined periods and to obtain depth of clinical data as well as patient-reported outcomes (PROs).

find ways to obtain the depth of clinical encounters outside of traditional observational research repositories as data sources and health care delivery evolve. From a technical perspective all organizations that collect health information maintain patient-level identifiers that can be used directly or through anonymous linkage[15] techniques to link data across data partners over defined periods.

A second technical challenge in developing a DRN is how to increase the person-time follow-up as individuals move through health insurance companies. Individuals may change health insurance companies for a variety of reasons: open-enrollment decisions, change of employers, change to a spouse's coverage, loss of employment resulting in subsequent enrollment in gap coverage insurance or Medicaid, and retirement resulting in coverage change or eligibility for Medicare. Additionally, legislative changes can alter the access to insurance coverage, providing another source of variability in the ability to follow individuals longitudinally. The existing fragmented payment system prevents obtaining long follow-up time in secondary data vital to many safety surveillance, CER, and quality improvement activities. Of particular note is the disruptive nature in longitudinal follow-up of payer administrative data when individuals become eligible for Medicare at age 65 or for chronic disease eligibility. Building the capacity to routinely and consistently link commercial claims data with Medicare data presents an opportunity to increase both the number of eligible members for an analysis and person-time available for follow-up. From a technical perspective, the ability to link data longitudinally across data partners is possible either directly or through anonymous linkage[15] to link data longitudinally as patients move from one administrative health plan to another, or from one health system to another.

A third technical challenge will be to include health care providers, health plans, and other related entities that historically have not utilized existing data or infrastructure in a collaborative manner for secondary safety surveillance, CER, and quality improvement activities. Developing the necessary education and the outreach efforts to these groups to demonstrate the important benefits realized by the use of a CDM in advancing evidence generation and improving data integration will be a challenging step in expanding a nationally representative research network. These efforts will require current leaders in bidirectional research networks

to identify the unique business needs of each of these organizations, to understand the factors that have limited or prevented their participation in collaborative research networks, and to develop an arrangement that is mutually beneficial to all parties.

Governance Challenges

The technical challenges of data linkage are relatively easy when compared to the governance challenges around data utilization. The first governance challenge for the research community will be to develop the methods to preserve patient privacy while creating a multipurpose DRN with the capability of utilizing data across data partners. Linkage strategies, such as anonymous linkage[15,16] and distributed regression,[17-20] will likely need to be employed to protect patient privacy. These privacy preserving methods will be vital regardless of whether the activity is a public health surveillance, CER, or quality improvement activity.

The second governance challenge will be to provide data governance that protects the proprietary business interests of participating organizations that initially collect health care data. These organizations already exchange data in the forms of treatment, benefit, and operations; governance will need to protect the business case of each stakeholder who is engaged—to ensure that the data linkage activities will not adversely affect the business interests of the organizations involved. The research, provider, payer, and patient communities need to build the collaborations and engender the trust necessary to address the data needs for the betterment of public health, CER, and quality improvement.

Perspective

The recent PCORI board decision to fund up to two health insurance companies to demonstrate the value of payer data in providing longitudinal data is a step toward creating a national research infrastructure. Full engagement of the health plan stakeholders in patient-centered research is a major advancement toward an infrastructure that supports multiple safety surveillance, CER, and quality improvement initiatives. As we move toward a learning health care system that places increased demands on clinicians to electronically document the care provided, as patients begin to capture health data through wearable technology, and as payment models introduce new complexities, the demands on both the data governance and technical capacity to link data sources will become increasingly challenging. Collaboration with large regional or national health plans will be essential for broad access to longitudinal data.

The ability to link patients across health plans and ultimately into Medicare will be pivotal in the transformation of true longitudinal patient follow-up and the ability to study long-term outcomes. If an investigator wants to study the effects of pediatric exposures on the onset of adult outcomes, the data networks available today are insufficient to address the discovery of these potential associations. Longitudinal data transformation will require additional governance to protect the proprietary business interests of the entities that initially collected the information. The decision by the Centers for Medicare & Medicaid Services (CMS) will allow innovators and entrepreneurs access to Medicare claims, which opens the door to potential further integration.[21] Researchers will be granted access to the CMS Virtual Research Data Center (VRDC), which contains granular privacy-protected CMS data files. A major innovation will be to utilize the VRDC in extending the follow-up of commercial health plans to generate a longitudinal patient record ideally linked with clinical data from either PCORnet or other clinical data repositories.

The preservation of patient privacy will be paramount in building trust among patients for

these activities. These collaborations will need to involve researchers and governance resources across data resources—and not simply be an extraction of an entire data set, but establish clear data use agreements for bidirectional data exchange of the minimum data necessary to respond to an inquiry. Activities will need to clearly demonstrate benefit to public health through either active surveillance or the generation of comparative effectiveness of therapeutic modalities. Ethical and regulatory oversight will continue to evolve, as data queries will require access to multiple resources. Within health systems, observational clinical research is often performed within a single "covered entity" and can thus seek a waiver of informed consent according to the Health Insurance Portability and Accountability Act (HIPAA).[22] Such protocols stress, according to the Institutional Review Board (IRB) process, that the research poses no more than minimal risk, and that the only risk is loss of confidentiality. In situations requiring data linkage with external resources to the specific covered entity, provisions such as anonymous linkage will be needed to maintain patient privacy.[15,16] There are other examples of linking EMR data with claims data from health insurers while maintaining patient privacy through a trusted third party.[23] Additionally, proposed changes to the Federal Policy for the Protection of Human Subjects (known as the "Common Rule") will likely have an impact on the conduct of public health surveillance and observational clinical research in secondary data.[24] Regardless of the strategies employed, major governance and technical innovations are essential to alleviate the burden and costs associated with conducting linkage for public health surveillance and clinical research.

The era of "big data" will shift from the discussion of how many patients or members are in the database, to how linked is the data to capture the full picture of the patient's interactions with the health care systems, and over what period is the patient followed. Undoubtedly, health care will continue to need big data, but arguably in the form of "deeper data" to address clinical research questions and patient care. The question is how broad and deep are the data on an individual, both in terms of longitudinal follow-up and depth of detail on clinical encounters, with health systems. These are exciting times for both public health surveillance and patient-oriented research as both funding opportunities and data access have increased. Now we need to build not only the infrastructure but the trust to conduct these large-scale data endeavors.

Acknowledgements

References

1. Selby JV. Linking automated databases for research in managed care settings. *Ann Intern Med.* 1997;127(8 Pt 2):719-724.
2. Curtis LH, Brown J, Platt R. Four health data networks illustrate the potential for a shared national multipurpose big-data network. *Health Aff (Millwood).* 2014;33(7):1178-1186.
3. Platt R, Carnahan RM, Brown JS, et al. The U.S. Food and Drug Administration's Mini-Sentinel program: status and direction. *Pharmacoepidemiol Drug Saf.* 2012;21 Suppl 1:1-8.
4. Curtis LH, Weiner MG, Boudreau DM, et al. Design considerations, architecture, and use of the Mini-Sentinel distributed data system. *Pharmacoepidemiol Drug Saf.* 2012;21 Suppl 1:23-31.
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):578-582.
6. Fleurence RL, Beal AC, Sheridan SE, Johnson LB, Selby JV. Patient-powered research networks aim to improve patient care and health research. *Health Aff (Millwood).* 2014;33(7):1212-1219.
7. NIH Health Care Systems Research Collaboratory. 2015; https://www.nihcollaboratory.org/about-us/Pages/default.aspx. Accessed June 29, 2015.
8. Richesson RL, Hammond WE, Nahm M, et al. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. *J Am Med Inform Assoc.* 2013;20(e2):e226-231.

9. Anderson ML, Califf RM, Sugarman J, participants in the NIHHCSRCCRTW. Ethical and regulatory issues of pragmatic cluster randomized trials in contemporary health systems. *Clin Trials.* 2015;12(3):276-286.

10. Platt R, Tenaerts, P., Archdeacon, P., Chrischilles, E., Evans, B., Hernandez, A., McGraw, D., Walraven, C., Raebel, M., Rosati, K. Developing approaches to conducting randomized trials using the mini-sentinel distributed database. 2014; http://www.mini-sentinel.org/work_products/Statistical_Methods/Mini-Sentinel_Methods_CTTI_Developing-Approaches-to-Conducting-Randomized-Trials-Using-MSDD.pdf. Accessed 3 December, 2015.

11. AMCP Biosimilars Collective Intelligence Consortium. 2015; http://www.amcp.org/AMCPBiosimilarsCIC/. Accessed June 29, 2015.

12. Advancing Regulatory Science for Public Health. *Innovation in Medical Evidence Development and Surveillance* 2015; http://imeds.reaganudall.org/. Accessed June 29, 2015.

13. Jaffe S. 21st Century Cures Act progresses through US Congress. *Lancet.* 2015;385(9983):2137-2138.

14. Ashwood JS, Reid RO, Setodji CM, Weber E, Gaynor M, Mehrotra A. Trends in retail clinic use among the commercially insured. *Am J Manag Care.* 2011;17(11):e443-448.

15. Kho AN, Cashy JP, Jackson KL, et al. Design and implementation of a privacy preserving electronic health record linkage tool in Chicago. *J Am Med Inform Assoc.* 2015;22(5):1072-1080.

16. Weber SC, Lowe H, Das A, Ferris T. A simple heuristic for blindfolded record linkage. *J Am Med Inform Assoc.* 2012;19(e1):e157-161.

17. Wu Y, Jiang X, Kim J, Ohno-Machado L. Grid Binary LOgistic REgression (GLORE): building shared models without sharing data. *J Am Med Inform Assoc.* 2012;19(5):758-764.

18. El Emam K, Samet S, Arbuckle L, Tamblyn R, Earle C, Kantarcioglu M. A secure distributed logistic regression protocol for the detection of rare adverse drug events. *J Am Med Inform Assoc.* 2013;20(3):453-461.

19. Karr AF, Feng J, Lin X, Sanil AP, Young SS, Reiter JP. Secure analysis of distributed chemical databases without data integration. *J Comput Aided Mol Des.* 2005;19(9-10):739-747.

20. Wolfson M, Wallace SE, Masca N, et al. DataSHIELD: resolving a conflict in contemporary bioscience--performing a pooled analysis of individual-level data without sharing the data. *Int J Epidemiol.* 2010;39(5):1372-1382.

21. CMS announces entrepreneurs and innovators to access Medicare data. 2015; http://www.cms.gov/Newsroom/MediaReleaseDatabase/Press-releases/2015-Press-releases-items/2015-06-02.html. Accessed June 29, 2015.

22. United States. Congress (104th 2nd session : 1996). *Health Insurance Portability and Accountability Act of 1996 : conference report (to accompany H.R. 3103).* Washington, D.C.?: U.S. G.P.O.; 1996.

23. West SL, Johnson W, Visscher W, Kluckman M, Qin Y, Larsen A. The challenges of linking health insurer claims with electronic medical records. *Health Informatics J.* 2014;20(1):22-34.

24. NPRM for Revisions to the Common Rule. 2015; http://www.hhs.gov/ohrp/humansubjects/regulations/nprmhome.html. Accessed October 16, 2015.