Pains and Palliation in Distributed Research Networks: Lessons from the Field

Michael G. Kahn MD, PhD1 (moderator), Jeffrey Brown PhD2, Lisa Dahm PhD3, Daniella Meeker PhD4, Lisa Schilling MD, MSPH1

1University of Colorado Denver CO, 2Harvard Pilgrim Health Care Institute Boston MA, 3University of California Irvine CA, 4Rand Corporation Los Angeles CA

Abstract

Large-scale comparative effectiveness research studies require detailed clinical data collected across disparate clinical practice settings and institutions. Distributed research networks (DRNs) have been promoted as one approach to wide-scale data sharing that enables data sharing organizations to retain local data ownership and access control. Despite significant investments in distributed data sharing technologies, clinical research networks using distributed methods remain difficult to implement due to a broad range of organizational and technical barriers. The panelists represent four different research networks are in different stages of implementation maturity and are leveraging different informatics technologies. Challenges common to all DRNs include governance, semantic interoperability, and identity management. This panel will describe some of the critical challenges and experimental solutions to implementing, expanding, and sustaining DRNs. Each panelist will focus on a specific challenge that requires new informatics tools to reduce barriers to participation and data sharing.

Problem Description

With the growing deployment of electronic health records, the volume of detailed clinical observations collected during routine clinical care in standard clinical practices is increasing rapidly. Comparative effectiveness research (CER) focuses on evaluating the relative effectiveness of clinical interventions as observed in actual clinical practice rather than under artificial ideal conditions created in controlled clinical trials. Thus, access to detailed clinical data across a wide range of clinical practices is necessary for CER to compare clinical outcomes across differing clinical settings, workflows, constraints, and populations. A substantial investment has been made in creating large practice-based research networks. Many require technical staff to perform study-specific data extractions which are sent to a central coordinating center for analysis. DRNs have been promoted as an alternative approach than enables data contributing sites to retain control over data access and release. In a distributed network, data are extracted to a local data store whose content and access is controlled by local data administrators. Data requests are managed via a centralized data portal or coordinating center, are sent to all data sites who have agreed to participate in the data sharing activity. Local sites may deny access to their data at any time. Distributed data sharing models have held the promise of enabling a wider range of clinical sites to participate in national CER efforts. Yet, despite successes in non-medical fields, distributed data sharing methods have struggled to achieve wide spread adoption and sustainability. Significant technical and non-technical challenges remain. Existing informatics tools are too cumbersome to enable broad deployment in many clinical practices, even those with electronic health records. New tools and methods are required that substantially reduce barriers to data sharing for the vision of CER to be realized.

This panel brings together leaders of multiple DRN projects. These leaders bring real-world experiences with deploying and sustaining early-stage and more-mature networks. The diversity of technologies and approaches provides an opportunity to learn from these differing settings and experiences and provide an opportunity to challenge the Clinical Research Informatics community to develop the next-generation data sharing platforms and tools that enable smaller clinical practices to contribute to a national CER network.

Moderator: Michael Kahn, MD, PhD

Panelist: Jeffery Brown, PhD, PopMedNet

PopMedNet is an open-source software tool used as the secure distributed querying platform for several distributed networks, including the Scalable Partnering Network for CER: Across Lifespan, Conditions, and Settings (SPAN), FDA Mini-Sentinel, the HMO Research Network, and MDPHnet. These networks vary across a range of important characteristics, including purpose (comparative effectiveness research, public health surveillance), governance, implementation architecture, state and federal regulations, data models, and query interfaces. The
guiding principles for each network are to enable secure distributed querying of health data that allows data partners
to retain physical control of their data and its uses, while permitting approved requesters to send queries for local
execution, with information returned to the requester. Key issues faced across the networks include security,
governance, trust, and query mechanisms.

Regarding security, some data partners in one network wanted to encrypt all results that would reside on the
network portal for any amount of time. The PopMedNet network portal serves as the conduit for sharing of results
between data partners and the requester. The network portal is hosted in a secure Tier III data center and operations
are FISMA compliant. Nevertheless, these partners were unwilling to allow unencrypted files to reside on the
network portal for fear of unauthorized access by the IT vendor or secure hosting facility staff. The project
developed governance and software solutions to incorporate automated encryption for all results transmitted to the
portal, but these solutions create several other workflow problems.

A second, more common, issue that has arisen across several networks is the need for more fine-grained access
control. In response to requests from users, we developed several new user roles, and developed a more flexible
approach for creating new roles and modifying the rights given to users. New user roles include "observer", "query
administrator", "group administrator", and "analyst". To further address improved fine-grained access control we are
implementing a new architecture that allows data partners to set querying permissions and response settings based
on any combination of user, query type, and network project.

Panelist: Daniella Meeker, PhD, SCANNER

The SCAlable National Network for Effectiveness Research (SCANNER) project was one of a series of AHRQ
projects intended to enhance DRN infrastructure for multisite comparative effectiveness research. The multi-
institutional partnership led by UCSD developed requirements around demonstration studies that include both
observational and intervention designs. Founding data partners are the Tennessee VA, Brigham & Women’s
Hospital, and The UCSD Medical Center. SCANNER is designed to enable data management and analysis for
multiple studies within the same network, where each study might focus on different clinical domains, have varying
data needs and data models, and might have different data sharing policies, and with different combinations of sites
and investigators participating in each study. Each node in the SCANNER network is designed to execute services
for data analysis and data transformation. In order to facilitate standardized practices for developing new services
and studies, SCANNER has adopted OMOPV4 as a common data model. Other requirements include role-based
access control, flexible models for hosting and processing data, and support for heterogeneous and extensible
models for governance policies.

One area where SCANNER has focused efforts is addressing the problems with the practice of sharing limited
data sets across sites by sending extracts to a central analysis site for pooling. Limiting shared data to only aggregate
statistics often eliminates the need for data use agreements for HIPAA compliance and is more palatable for IRBs
and institutional risk managers. However, sharing only univariate or bivariate aggregates like counts and means does
not support multivariate regression methods frequently employed in CER. While conventional meta-analysis
methods most commonly applied in systematic reviews can be employed in distributed networks to yield
informative results, these approaches do not optimally utilize the network data resources, and result in loss of
information that would not occur in a pooled approach. By applying principles of parallel distributed processing,
SCANNER researchers have demonstrated a system for implementing logistic regression that iteratively converges
to the same parameter estimates as would be observed in an estimation using pooled data. This privacy preserving
method allows sites to share only intermediate, aggregate statistics by locally computing estimations iteratively with
a master-worker architecture. The same approach is valid for all generalized linear models (GLMs), enabling a
broad range of multivariate analyses to be conducted across sites. Further work includes implementation of
SCANNER distributed analytic services for GLMs that include site fixed and propensity matching methods.

Panelist: Lisa Schilling, MD, MSPH, SAFTINet

SAFTINet (Scalable Architecture for Federated Translational Inquiries Network) is a multi-state partnership of
academic, clinical and Medicaid organizations working to create a distributed data network of clinical and claims
data to support comparative effectiveness research (CER) and quality improvement for safety-net stakeholders and
investigators. Partners involved in data sharing and prototype CER include Denver Health and Hospital Authority,
The Colorado Community Managed Care Network, and Cherokee Health Systems. The project is also focusing on
the acquisition of patient reported outcomes and their use for both clinical care and research. SAFTINet is
collaborating with Ohio State University to deploy TRIAD-based grid services and a single central query portal to
allow authorized data access. We have worked with the Observational Medical Outcomes Partnership (OMOP) to modify their prior data model to support a broader range of CER questions and the multi-level analytic methods required to assess the impact of provider, practice and organizational effects on outcomes. Recombinant Data Corporation is our partner in the creation of a system, ROSITA, which facilitates data transformation and harmonization from disparate EHRs, and data formatting to the common data model. In addition, ROSITA supports essential data quality metrics and will eventually make these available via a parallel data quality grid service.

An instructive SAFTINet challenge has been the ability to facilitate the sharing of Medicaid claims data in a reusable and sustainable way. The decision to share data requires substantial commitment to explore and tackle the issues that are potential barriers. These issues include the following:

1) Giving up control of large amounts of data to support a CER infrastructure, in contrast to more typical requests of a specific data set for a single study covered under a single DUA.

2) Needing to provide identified patient-level claims data to clinical partners to support record linkage for use that includes research in addition to QI - even when the data eventually made available for research conforms to a HIPAA-defined limited data set.

3) Potentially providing proprietary cost or reimbursement information about one entity to another entity.

Many of these challenges centered on the fact that we were requesting the use of personal clear text identifiers to perform record linkage and that we were requesting a data set with breadth enough to build a CER infrastructure and not merely elements required for a specific study. These factors presented greater risk to the Medicaid entities due to the potential loss of control of the information on behalf of their beneficiaries and concerns regarding HIPAA non-compliance. We therefore shifted to a data sharing mechanism that would allow Medicaid to share data without exposing clear text identifiers and instead using privacy protected record linkage.

Panelist: Lisa Dahm, PhD, UCRex

The University of California set out to develop a system to encourage collaboration among its 5 Medical Center campuses. A key component of this system is a tool that identifies cohorts of research participants across campuses. Each campus has established an instance of i2b2 (Informatics for Integrating Biology and Bedside) onsite. SHRINE (Shared Health Information Network) allows the propagation of federated queries across instances of i2b2. The UC Research Exchange (UCReX) is (presently) a SHRINE network. By leveraging i2b2 and SHRINE technologies across participating academic medical centers, the University of California system has stood up a resource that allows access to the de-identified records of over 12 million patients across the state. This setup provides a governance structure, infrastructure, and front end that enables research teams to validate the existence and size of potential research cohorts through a user friendly UI while eliminating the possible exposure of Protected Health Information (PHI). Through numerous deployments across multiple academic and clinical organizations it has been determined that an architecture that connects an i2b2 system to SHRINE within the organization’s clinical firewall environment makes the most sense from a security perspective. These pieces are exposed to the other SHRINE nodes through either a peer-to-peer network approach or through a hub and spoke setup. UCReX implemented a hub and spoke model because it is more straightforward to setup and manage. This deployment routes all messages from each node to the SHRINE Management Server housed in the University of California Davis Data Center. Recombinant Data Corporation wrote proprietary software that allows the SHRINE Management Server to create audit reports of all queries that have been run, maintain a heartbeat signal for each organization, alert a support team upon system down time, and facilitate the review or editing of configuration files at each UCReX site.

A significant issue related to the development of the UCReX SHRINE network that requires additional informatics work is how to most effectively expand functionality beyond the provision of counts from simple cohort discovery. How can we expand functionality to create a collaborative environment that supports cross-institutional IRB submissions and access to PHI data? Another issue common to any shared network project is semantic interoperability and development of a common ontology. When sharing information between multiple health care organizations the terminology used to describe a medical concept must be aligned between the sites. This can be handled in more than one way. In the UCReX project, the organization underwent a data harmonization effort to identify areas of knowledge that are aligned/misaligned with respect to the terminology standard used for the organization.

All panelists have agreed to take part in the panel presentation.