Effective Health Care Program
Research Reports

Blueprint for a Distributed Research Network To Conduct Population Studies and Safety Surveillance

Jeffrey Brown, Ph.D.
Beth Syat, M.P.H.
Kimberly Lane, M.P.H.
Richard Platt, M.D., M.S.

Research from the Developing Evidence to Inform Decisions about Effectiveness (DEcIDE) Network

The DEcIDE (Developing Evidence to Inform Decisions about Effectiveness) network is part of AHRQ's Effective Health Care program. It is a collaborative network of research centers that support the rapid development of new scientific information and analytic tools. The DEcIDE network assists health care providers, patients, and policymakers seeking unbiased information about the outcomes, clinical effectiveness, safety, and appropriateness of health care items and services, particularly prescription medications and medical devices.

The findings and conclusions in this document are those of the authors, who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, no statement in this report should be construed as an official position of AHRQ or the U.S. Department of Health and Human Services.

None of the authors has a financial interest in any of the products discussed in this report.

Suggested citation:

Brown J, Syat B, Lane K, et al. Blueprint for a distributed research network to conduct population studies and safety surveillance. Effective Health Care Research Report No. 27. (Prepared by the DEcIDE Centers at the HMO Research Network and the University of Pennsylvania Under Contract No. HHSA29020050033I T05.) Rockville, MD: Agency for Healthcare Research and Quality. June 2010. Available at: http://effectivehealthcare.ahrq.gov/reports/final.cfm.

Contents

Author affiliations:

Jeffrey Brown, Ph.D.[a]
Beth Syat, M.P.H.[a]
Kimberly Lane, M.P.H.[a]
Richard Platt, M.D., M.S.[a]

[a]Department of Ambulatory Care and Prevention, Harvard Medical School and Harvard Pilgrim Health Care

Abstract

A distributed research network with an efficient, reusable infrastructure to assemble and analyze routinely collected healthcare data and related information could assist AHRQ by providing the means to rapidly generate reliable information about utilization and outcomes of care needed to support decision making by patients, providers, and policy-makers. A distributed network is preferred over a centralized system because it allows data holders to maintain physical and logical control over their data. At the highest level, a functioning distributed research network should be able to securely perform the following tasks: (1) distribute queries through network software; (2) execute the queries against the local data; and (3) return aggregated results to the end-user. A network should also support a variety of study types, including observational studies, quasi-experimental studies, clinical trials, and registries. Finally, a network should support both simple, menu-driven querying and complex queries using customized analysis code.

A large-scale distributed research network will require a substantial investment in administrative and governance infrastructure along with the investment in information technology. Issues such as security, proprietary, legal, privacy, and cost will present substantial challenges to implementation and maintenance of a network.

The current report serves as a blueprint to guide future development of a distributed research network based on the experience of designing and testing a network prototype. The report includes lessons learned from administrative, governance, technical, and research components of the project, and it emphasizes the scalability of the system.

A phased, systematic approach to implementation is recommended for creation of a viable distributed research network. In general, the first phase should focus on the most commonly-used and best understood data types, rely on simple technical requirements, and include targeted functionality. Additional phases would expand the network by adding new data sources, accommodating new data types, and expanding network functionality. This phased approach will enable research to be conducted in parallel with the development of the network, which will in turn help to inform continued development and improvements. Only a coordinated, well-supported, and step-wise approach is likely to garner the support necessary to build a viable and sustainable distributed research network.

1. Introduction

1.1. Objectives and Goals of the Distributed Research Network Project

The overall objective of this project is to design a scalable, distributed health information network that will support secure data analyses on the risks and benefits of therapeutics. Two key network architecture reports have been completed as part of the overall project:

- Specifications of network architecture and research network cooperative (Report #1).[1,2] That report recommended a technical design, key infrastructure components, and organizational structure required for a network to support large-scale, population-based studies on the risks and benefits of therapeutics.
- Proof-of-principle demonstration and evaluation (Report #2). That product has been completed and included a proof-of-principle implementation of a network prototype demonstrating some of the design features described in Report #1.

1.2. Purpose of this Report

The current report (Report #4) serves as a blueprint to guide future development of a distributed research network. The report includes lessons learned from administrative, governance, technical, and research components of the project. It emphasizes the scalability of the system.

The rationale, goals, potential uses, and challenges associated with a distributed research network have been described by the authors in Reports #1 and #2 and are briefly reiterated in the following sections.

1.3. Rationale for a Distributed Research Network for Public Health Activities

There is growing demand for using routinely collected healthcare information to rapidly develop scientific evidence and for new analytic tools to assist healthcare providers, patients, and policy makers to make informed decisions about the clinical effectiveness, comparative effectiveness, appropriateness, safety, population health, and outcomes of healthcare items and services. A distributed research network with an efficient, reusable infrastructure to assemble and analyze routinely collected healthcare data and related information could assist with meeting these disparate needs and address knowledge gaps in the areas noted above.

As compared to the development of multiple independent and single-purpose networks, a *single multi-purpose network* would reduce the total burden for data holder participation and for network infrastructure. The latter structure improves efficiency, especially through the reuse of common network components and savings on resources devoted to building data interfaces. It is unlikely that data holders, especially large data holders, will find it feasible or desirable to participate in multiple health data networks, raising the likelihood that independent, single-purpose networks will be difficult to sustain.

1.4. Overview of a Distributed Research Network

1.4.1. Distributed Approach

In principle, either a distributed network or a large centralized database could meet the specified system requirements. In the authors' opinions, the best way to create a viable and sustainable system is through a distributed network. A distributed network is preferred because it allows data holders to maintain logical and physical control over their data; without this control, in our experience, they are unlikely to voluntarily participate. A distributed system can mitigate security, proprietary, legal, and privacy concerns, many of which are regulated by the Privacy and Security Rules of the Health Insurance Portability and Accountability Act (HIPAA).[3] A distributed approach can eliminate the need to create, secure, maintain, and manage access to a complex central data warehouse; further, it can minimize the need to disclose protected health information (PHI) to users other than the participating covered entities. In addition, a distributed network allows data holders to accurately assess, track, and authorize query requests, or categories of requests, on a case-by-case basis. Ensuring that data holders can retain local control over all uses and users of their data is a key issue in securing their participation. Finally, a distributed network also avoids the need to repeatedly transfer and pool data in order to maintain a current database, which is a costly undertaking each time updating is necessary.

1.4.2. Attributes

At the highest level, a functioning distributed research network should be able to perform the following tasks:
- Distribute queries through network software,
- Execute queries against the local data, and
- Return aggregated results to the end-user.

A viable network should have access to health status, medical care, and outcome information from large populations, be able to incorporate new kinds of data as they become available, and allow a study protocol to be implemented identically and efficiently across the network. A network also should be designed to minimize data exchange, maximize local control of data and uses, and support standardized, reusable components to improve efficiency and learning.

A network also should support a variety of study types, including **observational** studies (e.g., hypothesis testing, adoption and diffusion of new medical technologies, and effectiveness studies), **quasi-experimental** studies, **clinical trials** (e.g., collect long-term outcomes for participants in randomized controlled trials and serve as sole data source for design and evaluation of cluster randomized trials of effectiveness), and **registries** (e.g., adding baseline and follow-up data to prospectively collected registry information).

Finally, a network should support both simple, menu-driven querying and complex queries using customized analysis code. Menu-driven queries should facilitate extraction of simple counts, such as the number of individuals receiving a treatment or surgical procedures performed by age, sex, region, and year. Menu-driven queries could be executed on an *ad hoc* or routine basis. The local execution of complex analyses, typically one-time programs in support of a formal research protocol, also should be possible through the network.

1.4.3. Potential Uses

A distributed research network would address the comparative-effectiveness research investments and activities identified by the Federal Coordinating Council for Comparative Effectiveness Research, including research (e.g., comparing medicines for a specific condition) and data infrastructure (e.g., developing a distributed practice-based data network, linked longitudinal administrative and claims or electronic health record databases, or patient registries).

Potential uses of a distributed research network include:

- Evaluation of medical product utilization patterns, including the adoption, diffusion, and ongoing use of new medical products
- Drug effectiveness studies
- Comparative effectiveness studies
- Assessment of trends and patterns of off-label or non-approved medical product use
- Assessment of disease burden and changes in clinical practice or population health
- Active medical product adverse event signal detection and signal strengthening
- Safety surveillance data mining (hypothesis generation)
- Confirmatory safety studies (hypothesis evaluation)
- Augmentation of registry information (e.g., medical devices)
- Calculation of background incidence rates for outcomes of interest
- Improving evidence regarding the predictive value of diagnosis codes of interest in automated health care data
- Potential to identify immediate adverse effects such as transfusion-related acute lung injury (TRALI) and transfusion-associated circulatory overload (TACO)
- RiskMAP effectiveness
- Evaluation of biomarkers for adverse event risk

1.4.4. Use Case Examples

The following selected use cases provide examples of the possible uses of a distributed network.

- *Disease surveillance*: identify the first diagnosis of hypertension within two years of an index date.
- *Treatment assessment*: describe the first anti-hypertensive drug dispensed within two years of an index date.
- *Outcomes*: identify a new diagnosis of angioedema after start of antihypertensive treatment or a hospitalization for a myocardial infarction or stroke after first treatment for hypertension.
- *Simple rates*: identify the rate at which hypertensive patients are dispensed antihypertensive therapy or the rate at which newly treated hypertensive patients discontinued therapy over a specified period such as a single year.
- *Case-mix adjusted comparisons*: calculate incidence rates of myocardial infarction and stroke among patients receiving a beta-blocker versus an angiotensin-converting enzyme (ACE) inhibitor as second line antihypertensive therapy, omitting those with an indication or contraindication for either agent, and adjusting for baseline health status.

1.4.5. Challenges

A large-scale distributed research network will require a substantial investment in administrative and governance infrastructure along with the investment in information technology. The administrative and governance infrastructure must enable a complex oversight structure of advisory and supervisory boards and be able to address issues such as network maintenance and usage, study oversight, monitoring, access, standardization of proposals, protocols, and multi-site agreements, including data use agreements. In addition, data standardization across organizations will be a substantial challenge requiring dedicated resources to achieve an acceptable level of consistency across data sources. Data standardization will be an ongoing resource demand as data systems change, coding standards evolve, and health information technology and exchange mechanisms mature and need to be accommodated.

Issues such as security, proprietary, legal, privacy, and cost will present substantial challenges to implementation of a network. In addition, concerns regarding risk mitigation, patient privacy and HIPAA, Institutional Review Board (IRB) and human subjects review must be addressed as part of a viable network design and architecture. It also will be necessary to develop a persuasive business case in order to convince data holders that the benefits of participation outweigh the real and potential costs of participation.

2. Blueprint for Implementation

2.1. Architecture

The high-level architecture of a proposed distributed network is illustrated in Figure 1. The infrastructure components identified as being key features of a viable and scalable research network include a central portal (hub) that manages:

- Network security
- Authentication
- Authorization
- Workflow
- Auditing
- Query monitoring, distribution, and aggregation
- Administration of governance policies.

These features of a re-usable and scalable technical infrastructure are required components applicable to any distributed research network regardless of database model, querying modality (menu-driven or distributed analytic programs), governance approach, research objective, or topic. As a rule, network design decisions should minimize the burden on data holders as a way to encourage broad participation.

2.1.1. Network Datamart

The simplest implementation of a distributed research network involves each site, as the data holder, creating and controlling a uniformly structured network "data mart database" that is physically separate from the data holder's primary data repositories. This network datamart resides in an isolated area inside the data holder's institutional Internet firewall. The network datamart adheres to a common data model that ensures identical file structures and data element definitions across all data holders.

Most implementations of common data models require each contributing partner (in this case, each data holder) to transform its data into a data model, either virtually or physically. Physical transformations are referred to as an extract, transform, and load (ETL) process. Implementation of a data model using an ETL procedure would greatly facilitate initiation of the network, but it is not a requirement of a network.

2.1.2. Network Operations

The basic flow of network operations begins when an end-user authenticates to the network portal by supplying credentials to establish his or her identity. Role-based access control (see: http://csrc.nist.gov/groups/SNS/rbac/) would be used to allow access to individual applications only if the authenticated user has appropriate permissions (i.e., authorizations).[4] An authorized user can then query available data resources based on the specific privileges associated with his or her identity. Data holders can set authorization policies for each user and for each query type; these include mandatory approvals from appropriate HIPAA privacy boards and IRBs. The user will submit a query, which may be a simple menu-driven operation or an executable analysis program, which then will be stored in a queue for retrieval by the appropriate data holders. Local policies will determine whether the query is automatically executed or manually reviewed for approval. Query results can be automatically encrypted and returned to the central website, or they can be queued for manual data holder approval before being returned. Application software on the portal will consolidate results and make them available to the user in aggregate across all data sources. Details of each step are recorded for monitoring and auditing.

2.1.3. Key Features of the Network

The key features of the network are described below.

Distributed Architecture

The network will employ a distributed architecture in which data holders maintain local control over their data and its uses. This approach avoids the need to centralize confidential or proprietary data.

Scalability

Scalability is a crucial aspect of the architecture because it is likely that a network would be built in distinct phases over several years and, therefore, requires the ability to incorporate new data holders, data types, and functionality.

Hub-and-Spoke Design

A client-server architecture with a central portal (also known as a hub-and-spoke design) is preferred. In a client-server network, all nodes are connected to a central server and do not necessarily know of the existence of any other nodes and do not need to communicate with, interact with, or verify the authenticity the other nodes. The client-server architecture with a central portal minimizes data holder IT responsibilities, provides for a more straightforward security implementation, and focuses network management tasks at the central portal.

Central Portal

A distributed system that features a central portal that performs network functions, such as operations (e.g., workflow, policy rules, auditing, query formation and distribution) and

security (e.g., authentication, authorization) and distributed data marts that remain under the control of the data holders is recommended. This design supports important capabilities, such as secure communications and data protection, auditable processes, a simple query interface that enables menu-driven and complex queries, and fine-grained, locally-managed security, authentication, authorizations, and permissions.

Query Distribution

A "pull" mechanism (also described as publish-and-subscribe or polling) in which data holders are notified of waiting queries (or routinely poll for queued queries) and retrieve them from the central portal for execution is preferred over a "push" mechanism in which queries are sent directly to data holders for automated execution. The "pull" approach for query distribution would obviate many of the security and access concerns of data holders.

Data Holder Autonomy

Each data holder will maintain control over all uses of its data, and will be responsible for establishing and enforcing its own policies and procedures with respect to data access and user and use audits.

Protection of Proprietary Information

Data holders maintain the responsibility of protecting their data and ensuring appropriate use within federal, state, and institutional patient protection and privacy guidelines. These restrictions and requirements are not unique to a distributed research network and would be undertaken for any secondary use of data. Potential additional issues include anti-trust policies and intellectual property concerns in the event that the network develops new intellectual property. Data holders also must agree to the adequacy of network safeguards against competitors accessing their proprietary data.

Protection of Patient Information

Keeping data under the control of data holders and avoiding a central data warehouse mitigates the potential for a large-scale release of personal health information (PHI), either by accident or through unlawful activity.

Strong, Central Coordination Role

The authors recommend a centralized approach to analytics in which queries, in the form of executable computer programs, are distributed to the data holders, who run them on locally held data that are stored in a common format. Requiring a common data model ensures a level of standardization of definitions, analytic approaches, and data quality that will otherwise be extremely difficult to achieve and verify. This approach also ensures that complex analytic approaches are implemented identically and the findings are comparable across institutions as long as the source data are comparably defined.

2.2. Network Coordination

In order to implement a distributed research network, the following key pieces must be in place.

2.2.1. Governance Structure

An operational governance structure that includes the roles and responsibilities of a coordinating center, data holders, and stakeholders must be organized. Network governance must include development of policies and procedures to address issues such as data holder protections, conflict of interest, external communications, priority setting, by-laws and governance rules, data security, accounting, network strategy, stakeholder issues, and HIPAA and human subjects protection.

2.2.2. Coordinating Center

A coordinating center should be established with responsibility for supporting and facilitating use of the network. This could include maintenance of network infrastructure, documentation, coordination, monitoring of data resources and contacts, documentation of lessons learned, data validity activities, and study implementation.

2.3. Implementation Overview: Recommendations for the Future

A phased approach to implementation is recommended for creation of a viable distributed research network. In general, the first phase should focus on the most commonly-used and best understood data types, rely on simple technical requirements, establish the security model (e.g., role-based access control), and include targeted functionality. Additional phases would expand the network by adding new data sources, accommodating new data types, and expanding network functionality. A summary of the phased approach is presented below.

2.3.1. Phase 1: Network Initiation

The first stage of network development should target distributed access to commonly available data sources (i.e., administrative and claims data), an easy-to-use menu-driven query interface, and the ability to manually execute distributed analysis programs. The data holders included at this stage should be required to have appropriate experience and access to original medical records to validate the coded electronic information. These capabilities will meet many primary user needs and also be acceptable to a range of data holders.

This stage of development should rely on use of a common data model using an ETL procedure, as this would facilitate development of network queries and minimize the effort needed to respond to queries. This approach does not exclude other options going forward, but it would greatly facilitate short-term initiation of the network.

The choice of data model is a function of the data available to the network, the willingness of data holders to provide detailed versus summary data for querying, and the purpose of the system. In general, the more granular the data, the more flexibility it provides. For example, an encounter-based patient-level model (a model consistent with claims and electronic medical record (EMR) systems) enables implementation of most types of observational studies, whereas summary-level data are most useful for monitoring medical product use.[5]

With appropriate dedicated resources, these short-term goals could be completed within one to two years.

Data Sources and Types (Phase 1)

The authors recommend initial development of the network using data arising from defined populations, i.e., those for whom there is sufficiently complete information regarding therapies and outcomes that occur no matter where care is delivered during specified periods. The best sources are ones that maintain administrative and claims data together with EMR data. Although the combination of administrative, claims, and EMR data for defined populations is preferred, relatively few systems have all the necessary data resources.

The next priority should be public and private insurers with large defined populations in administrative and claims databases. This recommendation is based on the large fraction of the population covered by these organizations (well over 100 million people), existing cross-institutional standardization, the comprehensive capture of ambulatory drug dispensing and most other exposures and outcomes of interest, and extensive experience using these data sources for post-marketing safety analysis.[3,6-8] Although these data sources will be useful for assessing many types of questions, they are not well suited to evaluating medical devices and inpatient care that cannot be uniquely identified with administrative, claims, and EMR data.

Functionality (Phase 1)

In phase 1, functionality should include the ability to securely distribute analytic code for manual execution, secure project communications, and implementation of a simple "pull" mechanism to distribute queries. Development of a simple menu-driven interface to gather feasibility and aggregate information such as patient counts also should be considered as part of the initial implementation phase.

2.3.2. Phase 2: Network Expansion

The next phase, which should follow as soon as practicable, should expand the network to additional data sources and types, and include more advanced functionality. These medium-term goals could be completed within two to four years with appropriate dedicated resources.

Data Sources and Types (Phase 2)

Continued development should incorporate data from additional data sources and incorporate new data types. New data sources could include additional health plans with administrative and claims data. New data types could include inpatient data (needed to evaluate short-term outcomes of therapies like blood products, contrast agents, general anesthetics, and other products used primarily in hospitalized patients), registry data, and, stand-alone EMR data.

Functionality (Phase 2)

In phase 2, the flexibility of the menu-driven interface should be expanded to accommodate new data sources and data types, add additional features to the query interface, and allow automated execution of selected types of queries, such as feasibility requests.

2.3.3. Phase 3: Network Maturation

The final phase of network development could be completed within four to six years with appropriate dedicated resources. This phase would include new data sources and more fully developed networking capabilities and automation.

Data Sources and Types (Phase 3)

In addition to continuing to add new data sources, work in this phase should focus on incorporating new types of data that will allow a complete system with full capabilities to conduct all types of necessary research. These new data types include linkage to national registry data, other registries, and genomic information. Work also should be done to link individuals longitudinally across data systems while protecting patient privacy and confidentiality.[9]

Functionality (Phase 3)

In phase 3, the flexibility of the menu-driven interface should continue to be expanded to accommodate new data sources and data types, add supplementary features to the query interface, and allow automated execution of additional query types. Further, enhanced functionality could include point-of-care systems for primary data collection[10] and automated distributed regression analysis.[11-13]

3. Summary/Conclusions

The blueprint described above presents a systematic approach to building the infrastructure necessary to help AHRQ generate the evidence needed to assist healthcare providers, patients, and policy makers to make informed decisions about the clinical effectiveness, comparative effectiveness, appropriateness, safety, and outcomes of medical products and services. This phased approach will enable research to be conducted in parallel with the development of the network, which will in turn help to inform continued development and improvements. The blueprint is consistent with the ideas proposed by the Federal Coordinating Council for Comparative Effectiveness Research and also the IOM report titled "Initial National Priorities for Comparative Effectiveness Research[14] which states:

"A large public-private CER enterprise will require a supporting infrastructure to efficiently move the science forward. In addition to the capacity to support high-efficiency, pragmatic randomized trials, the program will require large-scale clinical and administrative data networks that enable observational studies of patient care while protecting patient privacy and data security. New methods for linking patient-level data from multiple health care organizations will promote inclusion of populations frequently omitted from clinical trials." [14]

Only a coordinated, well-supported, and step-wise approach is likely to garner the support necessary to build a viable and sustainable distributed research network.

4. References

1. Brown JS, Holmes J, Maro J, et al. Report 1: Design Specifications for Network Prototype and Research Cooperative. Effectiveness Health Care Research Report No. 13 prepared by DEcIDE centers at the HMO Research Network Center for Education and Research on Therapeutics (HMORN CERT) and the University of Pennsylvania Under Contract No. HHSA29020050033I. Rockville, MD: Agency for Healthcare Research and Quality; January 30 2009. Available at: http://www.effectivehealthcare.ahrq.gov/reports/final.cfm.

2. Maro JC, Platt R, Holmes JH, et al. Design of a National Distributed Health Data Network. Ann Intern Med. Jul 28 2009;[epub ahead of print].

3. Moore KM, Duddy A, Braun MM, et al. Potential population-based electronic data sources for rapid pandemic influenza vaccine adverse event detection: a survey of health plans. Pharmacoepidemiol Drug Saf Dec 2008;17(12):1137-1141.

4. The Computer Security Division (CSD). Role Based Access Control (RBAC) and Role Based Security. National Institute of Standards and Technology, Information Technology Laboratory. Available at: http://csrc.nist.gov/groups/SNS/rbac/.

5. Brown JS, Lane K, Moore K, et al. Defining and Evaluating Possible Database Models to Implement the FDA Sentinel Initiative: U.S. Food and Drug Administration; 2009. Available at: http://www.regulations.gov/search/Regs/home.html#documentDetail?R=090000648098c282.

6. Platt R, Davis R, Finkelstein J, et al. Multicenter epidemiologic and health services research on therapeutics in the HMO Research Network Center for Education and Research on Therapeutics. Pharmacoepidemiol Drug Saf Aug-Sep 2001;10(5):373-377.

7. Schneeweiss S. Understanding secondary databases: a commentary on "Sources of bias for health state characteristics in secondary databases." J Clin Epidemiol 2007 Jul;60(7):648-650.

8. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol 2005 Apr;58(4):323-337.

9. Swire P. Application of IBM Anonymous Resolution to the Health Care Sector. IBM Available at: http://www.peterswire.net/anon.resolution.whitepaper.pdf. Accessed August, 2009.

10. Pace WD, Cifuentes M, Valuck RJ, et al. An Electronic Practice-Based Network for Observational Comparative Effectiveness Research. Ann Intern Med 2009 Jul 28.

11. Karr AF, Lin X, Reiter JP, et al. Secure regression on distributed databases. Journal of Computational and Graphic Statistics 2005;14(2):1-18.

12. Karr AF, Feng J, Lin X, et al. Secure analysis of distributed chemical databases without data integration. J Comput Aided Mol Des 2005 Sep-Oct;19(9-10):739-747.

13. Fienberg SE, Fulp WJ, Slavkovic AB, et al. "Secure" Log-Linear and Logistic Regression Analysis of Distributed Databases. In: Domingo-Ferrer, Franconi, eds. Lecture Notes in Computer Science. Heidelberg: Springer Berlin; 2006:277-290.

14. Institute of Medicine. Initial National Priorities for Comparative Effectiveness Research. June 2009.

Figure

Figure 1. High-level architecture of a proposed distributed network