Effective Health Care
Research Reports

Number 13

Design Specifications for Network Prototype and Cooperative To Conduct Population-Based Studies and Safety Surveillance

Jeffrey Brown, Ph.D.
John Holmes, Ph.D.
Judith Maro, B.A., B.S.
Beth Syat, M.P.H.
Kimberly Lane, M.P.H.
Ross Lazarus, M.B.B.S., M.P.H.
Richard Platt, M.D., M.S.

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

AHRQ
Agency for Healthcare Research and Quality
Advancing Excellence in Health Care • www.ahrq.gov

July 2009

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.

This report is based on research conducted by the DEcIDE (Developing Evidence to Inform Decisions about Effectiveness) Centers at the HMO Research Network Center for Education and Research on Therapeutics (CERT) and the University of Pennsylvania under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. HHSA29020050033I TO5). The AHRQ Task Order Officer for this project was Scott R. Smith, Ph.D.

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, Holmes J, Maro J, et al. Design specifications for network prototype and cooperative to conduct population-based studies and safety surveillance. Effective Health Care Research Report No. 13. (Prepared by the DEcIDE Centers at the HMO Research Network Center for Education and Research on Therapeutics and the University of Pennsylvania Under Contract No. HHSA29020050033I T05.) Rockville, MD: Agency for Healthcare Research and Quality. July 2009. Available at: effectivehealthcare.ahrq.gov/reports/final.cfm.

Contents

Author affiliations:

Jeffrey Brown, Ph.D.[a]
John Holmes, Ph.D.[b]
Judy Maro, B.A., B.S.[c]
Beth Syat, M.P.H.[a]
Kimberly Lane, M.P.H.[a]
Ross Lazarus, M.B.B.S., 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
[b]Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania
[c]Engineering Systems Division, Massachusetts Institute of Technology

Abstract

The use, cost, breadth, and depth of new medical technologies are growing rapidly, and health care stakeholders continue to seek emerging information about their relative risks and benefits. There is a need for a coordinated approach to generating valid scientific evidence about the harms and benefits of therapies to supplement what is known from clinical trials, and the related need for a sizable volume of longitudinal health care data to address many of the relevant questions. This report addresses the design specifications, such as technical design, key infrastructure components, and organizational structure, for a scalable, distributed health information network and research cooperative required for supporting large-scale, population-based studies.

1. Executive Summary

The use, cost, breadth, and depth of new medical technologies are growing rapidly, and healthcare stakeholders continue to seek emerging information about their relative risks and benefits. There is a need for a coordinated approach to generating valid scientific evidence about the harms and benefits of therapies to supplement what is known from clinical trials, and the related need for a sizeable volume of longitudinal healthcare data to address many of the relevant questions. The overall objective of this project is to design a scalable, distributed health information network architecture that will support secure data analyses for public domain research on the risks and benefits of therapeutics.

This report, the first of four for this project, includes a description of the technical design, key infrastructure components, and organizational structure of the network and research cooperative required for supporting large-scale, population-based studies on therapeutics.

The scalable network architecture will be designed to have the capability to securely submit both simple queries and complex analytic programs to run against data sources controlled and secured by individual data owners. The capability to distribute and execute arbitrarily complex queries to run against data that remain in control of the data owners, and to do so within a scalable and secure network, is novel.

Any distributed network design must address issues of scalability, transparency, autonomy, data heterogeneity, security, sustainability, and parsimony; inherent design trade-offs among these principles exist, as it is not always possible to accommodate all desired features when designing a complex system. Stakeholders, including current and potential future partners and information technology [IT] experts identified data autonomy, security, and the ability for fine-grained security and authorization and strong authentication as paramount concerns.

We, therefore, propose 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. This design addresses the issues and concerns noted above and supports the following capabilities: 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. IT experts confirmed that the proposed architecture is technically feasible and cautioned that governance and policy-making will likely pose the biggest barriers to implementation. A highly phased approach to implementation is recommended.

A strong oversight and management structure for such a network is essential. Concerns regarding risk mitigation, patient privacy and the Health Insurance Portability and Accountability Act (HIPAA), Institutional Review Board (IRB) and human subjects review, and legal and proprietary issues must be addressed as part of a viable network design and architecture.

2. Introduction and Scope

2.1. Background and Significance

The use, cost, breadth, and depth of new medical technologies are growing rapidly, and healthcare stakeholders continue to seek emerging information about their relative risks and benefits. There is a need to make better use of available, observational data to narrow the knowledge gap in post-marketing evidence. The Agency for Healthcare Research and Quality's

(AHRQ) Developing Evidence to Inform Decisions about Effectiveness (DEcIDE) program seeks to improve public knowledge about health outcomes more quickly than traditional approaches. AHRQ's goals include assessing the advantages and limitations of distributed database networks — defined as a system for remotely accessing clinical and other data resources secured and controlled by data owners — to support population-based research through the network itself or as part of a cycle (or continuum) that also includes studies conducted outside of the network. The potential uses of a distributed network to support population-based research include: describing therapies used across a range of practitioners, settings, and demographic groups and the associated therapeutic outcomes observed; identifying research questions about the intended and unintended effects of treatments and technologies used in clinical practice; generating and testing of hypotheses about the balance of harms and benefits of therapies; assessing long-term outcomes of treatments; and collecting new information about therapies at the point of care.

Recently, the Institute of Medicine's (IOM) Learning Healthcare System expressed dissatisfaction with the current approach to generating clinical evidence: "The prevailing approach to generating clinical evidence is inadequate…[It] takes too much time, is too expensive, and is fraught with questions of generalizability. How much can some of the problems…be obviated by…linked clinical information systems that might allow information about safety and effectiveness to emerge naturally in the course of care?"[1]

Recognizing that even very large individual databases and registries may not have a sufficient number of cases to detect treatment differences between various population subgroups, an aim of this project is to test the feasibility of using a distributed database network to support the generation and synthesis of new scientific information. This work is being conducted as part of AHRQ's Effective Health Care program, which was created after the enactment of the Medicare Prescription Drug, Improvement, and Modernization Act (MMA) of 2003.[2] Section 1013 of the MMA expanded AHRQ's research mission to include studies of clinical and comparative effectiveness. Such a network will contribute to the ability to close the knowledge gap regarding the safety, "real world" effectiveness, and comparative effectiveness of therapies.

There is a growing demand for using routinely collected healthcare information to rapidly develop scientific evidence and new analytic tools to assist healthcare providers, patients, and policy makers with making informed decisions about the clinical effectiveness, comparative effectiveness, appropriateness, safety, and outcomes of healthcare items and services. A distributed database network with an efficient, reusable infrastructure to assemble and analyze health and related information could assist with meeting these disparate needs. Such an infrastructure could provide a more rapid means to generate data about utilization and outcomes of care to support decision making by patients, providers, and policy-makers.

The Need for a Large, Distributed Research Network

As described above, AHRQ and others (e.g., the IOM, Food and Drug Administration [FDA]) have identified the need for a coordinated approach to generating valid scientific evidence about the harms and benefits of therapies to supplement what is known from clinical trials, and the related need for longitudinal healthcare data to address many of the relevant questions. The routine collection of such data, in the form of electronic medical records (EMRs), administrative data (demographic, eligibility, and claims), and registries makes it opportune to develop ways to use this information more effectively. However, information from more than one electronic data system may be needed in order to attain sample sizes that provide sufficient statistical power to ensure timely recognition of risks and benefits, the reproducibility and generalizability of findings, and to identify high-risk sub-groups. A distributed research network model would allow observation and study of a very large cohort of patients, thereby increasing the ability to quickly gain and disseminate best practice knowledge. A research network model also would improve study efficiency through re-use and standardization of study infrastructure (e.g., data models, study protocols, programming algorithms, IRB submission templates, contracting, and governance) and relationships.

In principle, generating valid scientific evidence about the harms and benefits of therapies can be accomplished by implementing either a distributed network or through creation of a large centralized data repository. The report's authors and others[3] believe the best way to satisfy the requirements of network users and participants, such as data owners, is to develop a distributed data network that allows data owners to maintain confidentiality and physical control over their data, while permitting authorized users to address questions important to public health. A distributed network is preferred over a centralized database because it avoids or reduces security, proprietary, legal, and privacy concerns, many related to the Health Insurance Portability and Accountability Act (HIPAA).[4] In addition, a network approach can give data owners complete control over both uses and users of their data. A distributed network model will therefore encourage participation by data owners who might refuse to provide data to a centralized system.

There are various and often conflicting definitions of distributed and federated networks. In this report we use "distributed network" to refer to a system in which data owners maintain possession of their proprietary data and allow controlled network access (e.g., querying) to the data for various purposes. A "federated query" can be disseminated across a distributed network to be executed locally against the available data resources, with results centrally aggregated and returned to the requestor. As envisioned, a distributed network will not require the routine transfer of large amounts of protected health information; rather, analyses will be executed locally within an environment secured and controlled by the data owner, and only the minimal necessary information will be transferred by the data owner, and only upon approval.

Limitations of Current Multi-Center Health Data Systems

Though several research networks exist (e.g., the Centers for Disease Control and Prevention's [CDC] Vaccine Safety Datalink comprised of eight Health Maintenance Organization Research Network [HMORN] member sites), none is large enough in size or broad enough in scope to address some of the most important public health needs facing providers, patients, and policymakers. Study-specific multi-center studies are not uncommon, but they are extremely time-consuming to create due to barriers regarding the use of data, including data quality and volume, technical limitations, and permissions. The need to share protected health information also limits current multi-center studies, as data owners are often unwilling to provide confidential and/or proprietary data. Multi-center studies also can be fraught with inefficiencies, including system heterogeneity, variation in human subject protection rules, trust building, contracting and coordination, and study governance policies. Additionally, technical issues, such as data being inaccessible and/or in many different formats, have impeded multi-center studies. Consequently, it is not possible to take full advantage of AHRQ networks, such as DEcIDE, to support these public health research needs.

Characteristics of a Distributed Network

Based on the experience of the authors, a distributed network capable of efficiently conducting clinical effectiveness, comparative effectiveness, appropriateness, safety, and outcomes studies of healthcare technologies and services should include these features:

- Distributed analytic capabilities: Allow secure, automated distribution and execution of computer programs (e.g., SAS programs) and aggregation of results sets.
- Minimal data transfer: Data is only transferred as needed to address study questions, and only the minimal necessary amount of data is transferred.
- Extensibility/ scalability: The network should be designed to allow for expansion and growth, including more analytic and system capabilities, incorporation of additional data types and sources, and inclusion of new data and software standards.
- Fine-grained authorization and permissions: Data owners maintain complete control over both uses and users of their data.
- Strong security and authentication: Incorporate strong standards for security and authentication (defense in depth strategies).
- Automated, extensive auditing: All network use should be monitored and be auditable.
- Multiple topics: Network should be broad enough to address multiple topics of interest to avoid creation of multiple, single-purpose networks.
- Standardized, reusable data model: A standardized data model will improve efficiency as projects will be able to learn from each other; would permit reuse of validated programs and algorithms, project templates, and standard operating procedures for quality assurance of the data.
- Menu-driven query: Incorporate an easy-to-use interface to conduct menu-driven federated queries for feasibility assessments or public health surveillance and monitoring.
- Project dataset created by distributed program: Network should have the capability to create, via a distributed program, project-specific datasets based on complex inclusion and exclusion criteria for use in approved studies.
- Pre-approved Institutional Review Board (IRB) and Data Use Agreements (DUAs) for standard queries, single IRB or delegation, DUA templates: A network should have agreements in place to facilitate access to information, with varying levels of authorization and pre-approvals dependent on the type of information requested.

Potential Uses of a Distributed Network

A distributed network should support a wide range of observational studies, and it should also include the potential to prospectively collect new information at the point of care. Observational studies include: (a) monitoring the dissemination of new medical technologies; (b) assessing the clinical effectiveness, comparative effectiveness, and safety of new and competing technologies; and (c) public health (including drug, device, and vaccine safety) surveillance. These include “quasi-experimental” studies that evaluate the consequences of system-wide changes in practice, such as substitution of one drug for another in a restricted formulary. Such data can also support evaluation of cluster randomized trials that intervene at the level of clinician, facility, or region. For such interventions, routinely collected healthcare data can provide both baseline information on health status and follow-up information on outcomes that result in healthcare utilization. In some cases, this information may directly inform policy and decision-making. In other cases, the information may assist in the design of other studies to generate additional evidence about a hypothesized relationship. Additionally, a research network will reduce the overall effort needed to conduct a study, especially the effort required for study start-up, database development, data quality checking, and analysis. This saving of valuable time and effort could result in completion of more queries and studies within a shorter time window and at a lower total cost than under the current ad hoc research model.

Examples of the kinds of capabilities required for a network include: (a) identifying individuals' first occurrence of a condition within a specified period (e.g., first diagnosis of hepatitis B within two years); (b) identifying initial treatment for a condition (e.g., first anti-hypertensive drug dispensing within six months); (c) obtaining complete follow-up over periods ranging from days to years; (d) calculating simple attack rates (e.g., seizure within 14 days of immunization); and (e) calculating comparative incidence rates (e.g., difference in the readmission rates for heart failure among patients treated with two different beta-adrenergic blockers). It should also permit adjustment for baseline differences between groups and for individuals' different lengths of follow-up. Lastly, the system should allow calculation of population-based estimates of rates of treatments, medical conditions, and health outcomes.

The network should allow non-technical users to ask simple questions without assistance (e.g., count people between the ages of 65 and 74 with a PET scan in 2008), but be flexible enough to allow sophisticated users to perform complex analyses (e.g., compare risk adjusted survival curves for breast cancer patients treated with tamoxifen as adjuvant chemotherapy, compared to those who did not).

2.2. Objectives and Goals of the Distributed Research Network Project

The current partnership with AHRQ provides the opportunity to develop a large distributed network prototype that can ultimately encompass most types of electronic health data from all willing and eligible holders of such data. Our primary goal is to focus on information exchange that is relevant to AHRQ's Effective Healthcare Program.

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. The two key network architecture design products are:

- Design specifications for network prototype and research cooperative. This product will include the technical design, key infrastructure components, and organizational structure of the network prototype required to support large-scale, population-based studies on the risks and benefits of therapeutics.
- Network prototype evaluation and testing. This product will include testing the implementation of a network prototype of the architecture designed in the aforementioned step.

The system architecture will comply with all privacy, security, and legal requirements, including current state and federal laws.

2.3. Outline of Report

This report, the first of four for this project, includes a description of the technical design, key infrastructure components, and organizational structure of the network prototype and research cooperative required for supporting large-scale, population-based studies on therapeutics. Section 3 of this report describes existing biomedical research models. Section 4 outlines the approach employed in our stakeholder discussions regarding a distributed research model for public health research, and Section 5 provides a report of findings from the interchange with these stakeholders. Section 6 delineates governing design principles for a network and discusses the inherent design tradeoffs associated with implementing these principles. Section 7 discusses options for high-level organizational management of such a network, with particular emphasis on the most pressing governance concerns of the stakeholders. Section 8 describes perceived or actual biases and limitations of this report based on the authors' own ideas or experiences with the HMORN. Section 9 previews the summary specifications for a proof-of-principle demonstration of the network design architecture. Finally, Section 10 outlines the prototype evaluation plan.

2.4. Outline of Future Reports

As part of the current project, three additional reports will be prepared and are described below.

Report 2: Network prototype evaluation and testing. This report will evaluate the proof-of-principle demonstration of the network architecture and characterize the needs, challenges, and barriers to creation of a distributed research model. During this phase, a network prototype will be built to demonstrate some functions of the network architecture described in this report.

Report 3: Research report—pilot project. In conjunction with AHRQ, and in parallel with the prototype project, a study protocol on the comparative effectiveness and safety of second-line anti-hypertensive agents will be developed, finalized, and implemented to further develop and evaluate the utility of a network with the characteristics being designed. A report will be provided in manuscript format based on the findings from implementing the study protocol. Specific research questions to be assessed during the pilot project are to:

1. Assess the comparative effectiveness of different 2nd line anti-hypertensive medications in achieving blood pressure control in patients whose blood pressure remains uncontrolled after initiation of 1st line therapy with a thiazide diuretic.
2. Assess the comparative effectiveness of different 2nd line anti-hypertensive medications in preventing long-term outcomes of myocardial infarction, stroke, heart failure, and chronic kidney disease in patients whose blood pressure remains uncontrolled after initiation of 1st line therapy with a thiazide diuretic.

Report 4: Blueprint/prospectus. AHRQ will be provided with a blueprint to guide future development of a larger network based on the experience of designing and testing the prototype. The report will include lessons learned from administrative, governance, technical, research, and political components of the project. It will particularly focus on the scalability of the system.

3. Overview of Existing and Planned Systems

3.1. Selection and Review Criteria

To inform our design process the available literature on existing distributed biomedical information networks was reviewed. To identify distributed biomedical networks the following search terms were used in PubMed:[5] federated data network, distributed data network, health data network, and federated query data network. When the existence of a distributed biomedical network was uncovered, such efforts were further researched using each network's publications and website. Review of non-medical querying systems or networks (e.g., travel search engines) is beyond the scope of the study. Health information exchange networks also were not reviewed because those networks are designed to provide clinical information (e.g., current medication lists) regarding specific individuals to specific providers (e.g., emergency room clinicians) to improve the quality of care being provided to the individual at the time of care; these networks are specifically designed to not allow the types of activities required for a distributed research network as described in this report.

Centralized—as opposed to distributed—biomedical networks and distributed networks in other scientific disciplines (e.g., physics, computer science, synthetic biology) were not reviewed because they were considered beyond the scope and of less direct relevance to the current project. Although the design considerations of centralized and non-biomedical networks or systems may be significantly different from the focus of this project, assessment of those approaches and lessons learned may be appropriate.

The review of existing networks targeted the purpose, current status, and lessons learned of the networks reviewed.

3.2. Existing Networks

Five networks met our inclusion criteria:

1. Electronic Primary Care Research Network[6,7,8]

Purpose. The ePCRN was initially created to facilitate patient recruitment for randomized controlled trial (RCT) studies among community medical practices. The architecture now supports clinical trial management system functionality and clinical trial analysis. The Federation of Practice-based Research Networks (FPBRN) has provided additional expertise and support for development. Specific aims were to perform real-time identification of potential subjects for RCTs, to link primary care clinics with potential investigators, and to speed the translation and dissemination of research findings into clinical practice. The ePCRN was funded primarily through the National Institutes of Health (NIH).

Technical description. ePCRN is a suite of applications that support a wide variety of functionalities, including clinical trial management and analysis. The eligibility software uses a client-server based Web Services architecture enabled by Globus® middleware and Internet2/Abilene networking technology. Specifically, "ePCRN Gateways" are Globus® servers that provide authentication and authorization services, receive query requests from "client" computers at community-based practices, enable communication protocols for further patient recruitment, and make available some advanced clinical decision support software and applications. These servers contain secure, de-identified patient registries in standardized XMLbased Continuity of Care Record/Documents (CCR/CCDs). The CCR/CCDs can be exported from a number of proprietary electronic medical record (EMR) formats.

The ePCRN utilizes the Primary Care Research Object Model, a standards-based model that is consistent with HL7, the Cancer Biomedical Informatics Grid[9] (caBIG™), and the Clinical Data Interchange and Standards Consortium (CDISC). By making use of metadata and controlled vocabularies, the ePCRN has reduced complications associated with cross-institutional variation in definitions and representations for clinical, demographic, or other concepts. The ePCRN uses Open Grid Service Architecture-Data Access Integration (OGSA-DAI) software to provide secure communication with the member practices. The wide variation in data structures is addressed through two mechanisms: (1) generalization of queries using CaBIG tools, and (2) standardization of data upon entry. Although neither mechanism is satisfactory alone, together these methods contribute to substantially enhancing accommodation to cross-institution data variability.

In terms of security, ePCRN practices a "defense-in-depth" strategy that closely correlates with the caBIG™ security suite. Thus, the program includes public key infrastructure (PKI) certificates, role-based access control, Transport Level Security (TLS)-based secure communication, and message encryption to ensure membership of individual Globus® servers. Citrix and "RSA key" hardware authentication are used at the research portal to ensure security for generation and distribution of research queries.

Distributed analytic capabilities, as defined above, are not incorporated into the current design.

Current status. Initial funding was provided to link 11 practice-based research networks (PBRNs), which consist of 50+ practices and 260 authorized primary care physicians. Each PBRN has a Research Director who trains researchers and physicians in the use of the ePCRN. Each Research Director is charged with managing the authorization policies within their PBRN. Over one million patients are covered within the ePCRN.

Lessons learned. The ePCRN has successfully overcome many of the trust issues that will be important for a distributed network. The ePCRN focused on a local data model with local control of the server (i.e., an autonomous system). It uses role-based restrictions to define their authorization policies as well as table- and column-based data restrictions for a finer degree of control. Much of the responsibility for maintaining authorization policies lies with the Research Director. Local institutional review boards and human subjects committees approve all research protocols and any data sharing arrangements, and typically, protected health information (PHI) on human subjects is available only locally. Additionally, significant effort has been put into developing a sustainable economic model that reimburses local practices when their locally held data is used. The University of Minnesota's Office of Business Development licenses both the "ePCRN Gateway" software and "ePCRN Research Portal" software in part to advance sustainability, promote standardization, and ensure security.

2. Harvard Catalyst (NIH Clinical and Translational Science Award [CTSA] at Harvard)[10]

Purpose. The Harvard Catalyst (Harvard NIH Clinical and Translational Science Center [CTSC]) is developing a distributed network among participating entities in order to facilitate patient cohort identification and to identify the location of appropriate biological materials for research studies. It is funded through a Clinical and Translational Science Award (CTSA)[11] granted by the NIH. This network will create the Shared Health Research Information Network

(SHRINE)—a universal querying system—that will be built using the foundations of i2B2 and the Shared Pathology Information Network (SPIN).[12] i2b2 (Informatics for Integrating Biology and the Bedside)[13] is a biomedical informatics infrastructure that facilitates clinical and translational research by providing integrated access to observational clinical care data in conjunction with newly available genomic data developed through genome wide association studies and other projects. The i2b2 infrastructure uses a common data model that will be used to facilitate querying across the individual clinical data warehouses at four of Harvard's teaching hospitals in addition to some related medical practices and health plans.

Technical description. Within the Harvard CTSC, an Internet portal (CONNECTS) will be used to connect researchers and data. Institutions that transformed their data into the i2b2 common data model will be included in the network; a SHRINE adaptor will allow the data to be retained and managed locally, but queried remotely by authorized CONNECTS users. Initially, the SHRINE query language will be restricted to a small subset of possible data types.

i2b2 was developed using a Web-based service-oriented architecture. It is platform independent and relies on web-based languages and protocols for describing metadata, sending communications, and registering services (i.e., software functions). The backbone of i2b2 is provided by the "Hive" which is a bundled set of software services that are loosely coupled to manage data and applications. There are five core "cells" that make up the Hive which provide the bare minimum of functionality for end-users. These are: (1) Project Management, (2) Ontology Management, (3) Identity Management, (4) File Repository, and (5) Data Repository (Clinical Research Chart). This CRC is a particular focus of the CSTC because it creates a common data model that is optimized for clinical genomic research. The data model is a "limited dataset" (i.e., a dataset that contains identifiable patient information but that has "facial" identifiers such as name and address removed) per HIPAA. That is, the information remains coded and linkable. Aside from the five core cells, there are a number of peripheral cells that can extend usability or meet particular research or development needs. Finally, the i2b2 infrastructure is accessed using a workbench. The workbench – enabled by Eclipse, an open-source, Java-based software development platform – communicates with the "Hive" using web services protocols.

Distributed analytic capabilities, as defined above, are not incorporated into the current design.

Current status. Selected clinical and encounter data from Harvard-affiliated clinical facilities are expected to populate the Harvard CTSC network. Data from medical practices and local health plans will be added to the network over time. The number of authorized users is likely to be high, and each independent entity (i.e. individual hospital) will be tasked with authorization of their employee-users. Thus, overall, the CTSC will adopt a federated security model in which collaborating sites agree to trust each other globally instead of establishing a trust relationship with each individual user.

Additionally, the i2b2 CRC infrastructure has been adopted outside of Harvard Medical School and its affiliated teaching hospitals. Early external adopters include the University of Massachusetts Medical School, Cincinnati Children's Hospital, the Morehouse School of Medicine, and potentially the University of Utah.

Lessons learned. The first demonstration of the Harvard CTSC network occurred during summer 2008. Thus far, reports are that governance concerns remain the sticking points in initial adoption. Each teaching hospital has had comparable concerns regarding patient privacy and security. Harvard's CTSC management currently plans to limit SHRINE queries to pre-approved options. There will be no ad hoc querying capability and no distributed analytic capabilities.

3. Cancer Biomedical Informatics Grid (caBIG™)[9]

Purpose. The Cancer Biomedical Informatics Grid (caBIG™) was funded to integrate information from clinical trials, genetic studies, protein studies, and other biological models related to cancer using a common networked interface. Using a common interoperable infrastructure, researchers, physicians and patients are encouraged to share information that may improve cancer detection, diagnosis, treatment or prevention. It is funded primarily by the NIH. caBIG™ recently initiated the BIG Health Consortium™[14], a multi-stakeholder collaboration designed to foster an integrated and interactive ecosystem (or "mega-community") of previously-unlinked sectors within life sciences and health care with a particular focus on personalized medicine. This new initiative will leverage the technical design of caBIG™.

Technical description. The caBIG™ tools employ a Web-services architecture that makes use of carefully controlled ontology and metadata to overcome barriers of data heterogeneity. caBIG™ is particularly noted for the broad range of distinctive data types that it seeks to add to the network, including computational models, images, gene array expressions, standard case report forms, and clinical records. Its ontology incorporates concepts that have been defined by medical standards groups, such as the CDISC as well as established data standards such as HL7, LOINC, ICD-9, SNOMED, and MedDRA. caBIG™ employs Globus®-based middleware for its communication protocols, security infrastructure including authorization policy management, and data integration. Bundled packages of middleware are released in both beta and production mode. caBIG™ actively encourages the development of new applications and the registration of new data. Data that is registered with caBIG™ generally must be cleaned and labeled with metadata at its origin.

Distributed analytic capabilities, as defined above, are not incorporated into the current design.

Current status. caBIG™ is used in over 50 cancer centers and institutions. There are no fees to join the network. caBIG™ is organized into workspaces that are charged with network management functions in addition to domain-specific development. Recently, in an effort to expand the network, a new program - the Enterprise Support Network (ESN) – has been designed to support and extend caBIG™ tools and infrastructure. The ESN includes support for six domain-specific Knowledge Centers consisting of experts in a particular facet of caBIG™ that will aid new organizations during caBIG™ adoption, and certification of caBIG™ Support Service Providers that provide services across four caBIG™ domains.

Lessons learned. caBIG™ released a summary report[15] that included some important lessons learned. First, the caBIG™ team emphasized the importance of managing expectations. Initial phases of caBIG™ were highly technical and focused on interoperability and software development. The network encountered a wide range of expectations on what could or could not be reasonably achievable given particular timeframes and budgets. Additionally, the caBIG™ team encountered a cultural clash when software developed in an academic setting was converted to a production model for broad distribution. In short, academic developers have skill sets that make them adept at finding innovative solutions to software problems. However, they do not have experience with the testing and rollout of commercial grade software that is high quality and cost-effective. There was general consensus that private entities needed to be engaged at earlier stages to ensure smooth transitions during software engineering.

4. Vaccine Safety Datalink (VSD)[16]

Purpose. The Vaccine Safety Datalink (VSD) was created in 1990 to conduct epidemiologic studies to monitor immunization safety using data from eight managed care organizations. All eight organizations are part of the HMO Research Network. Management of the VSD is centralized at the CDC. Participating managed care organizations assemble standardized datasets using a common data model. Many VSD activities are conducted using a distributed approach in which analytic programs are sent to participating sites for execution and results are returned for aggregation and analysis.

Technical description. The VSD does not employ sophisticated network software to handle tasks, such as federated querying, authentication, authorization, permissions, and auditing. Rather, these tasks are managed through password-protected access to a central portal (hub) for a very limited group of personnel. Each organization transforms their local data to a common data model and stores the data as SAS datasets; all analyses are conducted via distributed SAS programs. The programs are distributed through a central portal and can only be submitted by a few trusted partners. Some sites allow automated execution of distributed SAS programs and others require manual execution. VSD's central hub allows for secure, temporary storage of site-specific program output. These outputs are combined for analysis. Cross-institutional variability is handled through adoption of common software and hardware platforms and a common data model.

Although distributed analytic capabilities, as defined above, are not incorporated into the current design, the VSD system shares many of the characteristics of such as system.

Current status. The VSD has a defined population of approximately 8 million patients. All VSD health plan members can access patient medical charts as needed to validate exposure and outcome data.

Lessons learned.[17] The VSD model is greatly facilitated by the trusting relationships that have been built between the funder (CDC) and the eight participating health plans. It should be noted that several years ago a disputed vaccine study resulted in public demands for access to the data used by the VSD and resulted in the development of a VSD data-sharing program. The heart of the data-sharing conflict reflects a competing need to conduct transparent studies in order to assure reproducibility and validity of results coupled with a need to maintain the privacy of personal medical records.

5. Biomedical Informatics Research Network (BIRN)[18]

Purpose. The Biomedical Informatics Research Network (BIRN) was conceived as a shared biomedical infrastructure to support the computational demands, large data requirements, and shared laboratory resources for neuroscience and neuroimaging research. BIRN consists of a coordinating center, a brain morphometry "test bed" (i.e. research group), a clinical functional imaging "test bed," and a neurodegenerative disease mouse model "test bed." BIRN is funded by the NIH.

Technical description. BIRN uses an Abilene or Internet2 backbone coupled with a standardized hardware rack that is purchased and installed through the coordinating center. In addition to standard hardware configurations, the coordinating center supplies integrated software bundles that include both the middleware and applications required for BIRN. Software configurations are controlled via semi-annual releases and all BIRN members utilize the same version of the standard software. BIRN's middleware builds on the Globus® Toolkit particularly with regard to security solutions, authorizations, and credentials managements. BIRN uses a "mediator" architecture in which an end-user makes use of a "smart agent" to perform concept-based queries that retrieve information from databases that are held primarily in Oracle™. All data sources are annotated using a controlled BIRN terminology that incorporates medical standards such as SNOMED and LOINC. This terminology is referred to as BIRNLex. The mediator or smart agent interrogates the metadata that is used to "wrap" the databases to determine whether the information contained within them matches the criteria for the query. BIRN is distinguished from the other networks under study as the most thoroughly managed and standardized.

Distributed analytic capabilities, as defined above, are not incorporated into the current design.

Current status. 31 university hospitals and 39 research groups participate in BIRN.

Lessons learned. BIRN's high degree of standardization requires strong management capabilities and dedicated staff for sustainability. Data variability concerns are managed through adoption of common hardware and software configurations. BIRN has computational needs that far exceed what is expected for a distributed research network as envisioned in this report. Additionally, because the primary data for much of BIRN is imaging, BIRN has a greater need to establish the capacity to handle very large data files.

3.3. Summary of Existing Networks

Much can be learned from the activities of the networks described above, especially in the areas of federated querying, authentication, authorization, and security. It seems clear that developing any type of functioning network, whether to identify patients for a clinical trial or study large cohorts of patients for vaccine safety, is a large, expensive, and time-consuming endeavor. The ePCRN, developed over several years with several million dollars of funding, focuses on identification of cohorts for enrollment in clinical research studies. Its security and privacy policies and approaches can be useful in developing a distributed research network. ePCRN does not have the capability to extract and store analytic datasets and remotely execute complex queries against them: that is, distributed analyses. The Harvard Catalyst distributed network project (SHRINE) will be based on the i2B2 and the SPIN platforms (also developed over several years with millions of dollars of funding), and like ePCRN, focuses on cohort or case identification and also will not be capable of distributed analytics. Both ePCRN and SHRINE have (or plan to have) simple menu-driven querying capabilities. The initial phase of the caBIG project (2003 – 2007; $60 million in funding for 2004-2006) focused on development of a broad array of tools to enable distributed research within cancer surveillance and research. These tools could be used as part of the infrastructure needed to create a distributed network like the one envisioned here, but no such network using caBIG tools currently exists.[15] The current VSD data model, built over almost 20 years of collaboration and consistent funding, is capable of conducting epidemiologic and surveillance activities for vaccine safety, but it lacks some of the ability for menu-driven querying, fine-grained authorization, and scalability. The BIRN, also a long-term project, has many of the attributes of a distributed research network (e.g., security, common data model), but is limited in scope. There is much to learn from these biomedical networking activities about implementation of a distributed network.

3.4. Planned Networks

Several biomedical distributed research networks are currently in various stages of development or discussion. Three such networks with sufficient publicly-available information are:

1. **Sentinel Network.**[19] In May 2008, the FDA published its Sentinel Initiative strategy highlighting its plans to develop a distributed, queriable, and secure network based on existing electronic clinical databases.[20]
2. **AHIP Quality Measures.** The Robert Wood Johnson Foundation is working with America's Health Insurance Plans (AHIP) to access data across multiple institutions and providers to assess quality of care.[21] The planned network would allow a broad range of users to evaluate selected quality measures at various provider levels.
3. **MediGRID.**[22] As part of the German D-Grid Initiative, the MediGRID is a planned biomedical informatics grid that leverages the cyberinfrastructure foundation laid by other D-Grid Projects. The MediGRID research project is divided into four methodological modules (middleware, ontology, resource fusion and eScience) that will focus on incrementally developing a Grid infrastructure while taking into account the need of the biomedical users. Three end-user communities are represented: biomedical informatics, image processing and clinical research. The MediGRID project is funded for three years and is in the initial year of funding currently.

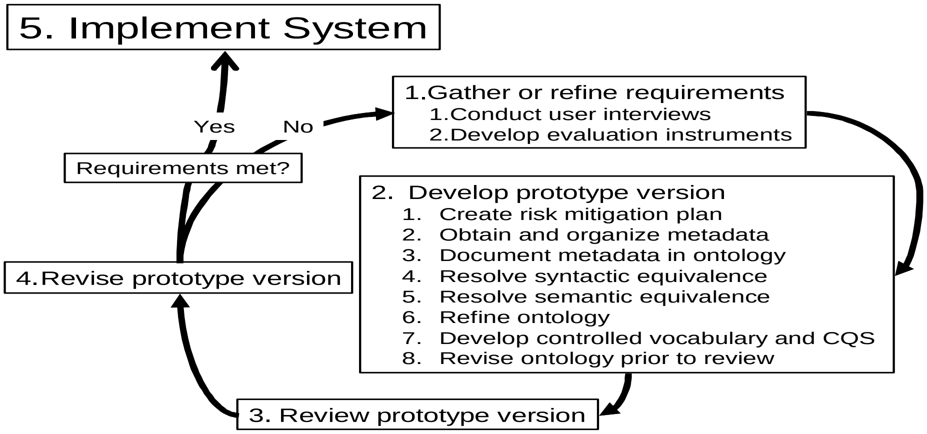
Other initiatives that have similar goals include the Data Extraction and Longitudinal Time Analysis System (DELTA)[23] that is proposing to use standardized registry data for surveillance of medical device safety, the Observational Medical Outcomes Pilot (OMOP),[24] and the eHealth initiative (eHI).[25]

4. Stakeholder Analysis: Methods

4.1. Gather System Design Parameters and Requirements

To ensure the information system architecture is feasible and appropriate, considerable effort was dedicated to developing a robust information systems model that fully reflects the disparate needs of stakeholders. A prototyping systems development life-cycle approach was employed to develop the system architecture and design and implementation specifications (Figure 1).[26,27] Based on this approach, a simplified system specification (i.e., a use case) was developed for discussion with stakeholders and iteratively refined as additional stakeholder input was collected. Thus, the specifications were developed through a series of stakeholder discussions and webinars anchored by evolving use-cases (i.e., use-cases were continually revised based on information obtained during previous stakeholder discussions). Each stakeholder discussion and webinar was updated based on information gathered from prior stakeholder interactions.

Figure 1. Prototyping system development life-cycle


5. Implement System
Yes
No
1.Gather or refine requirements
1.Conduct user interviews
2.Develop evaluation instruments
Requirements met?
2. Develop prototype version
1. Create risk mitigation plan
2. Obtain and organize metadata
3. Document metadata in ontology
4. Resolve syntactic equivalence
5. Resolve semantic equivalence
6. Refine ontology
7. Develop controlled vocabulary and CQS
8. Revise ontology prior to review
4.Revise prototype version
3. Review prototype version

Key stakeholders include current partners (i.e., AHRQ and HMORN sites), information technology (IT) experts ("vendors"), and potential future partners (e.g., individuals affiliated with FDA, CMS). Partners and future partners encompass data owners and/or the research community (i.e., those who would generate network queries). The initial design specifications were based on our experience within the HMORN, a review of the literature and existing networks, and preliminary discussions with stakeholders.

4.1.1. Development and Iteration of Use Cases

To prepare for discussions with the stakeholders, the Informatics team created a Network Overview and Use Case Summaries document and a slide deck describing: (1) the general architecture and requirement areas of the network; and (2) network guiding principles. The Network Overview and Use Case Summaries document contained background material and examples (i.e., use cases) of the potential network capabilities.

Use cases were developed and continually revised based on stakeholders' advice to understand the needs and capabilities of an ideal system. The initial use case presented to stakeholders traced the activities of a comparative effectiveness study from inception to analysis. The example illustrated a study comparing the rate of hospitalization (a measure of effectiveness) for two drugs within the same drug class. To help illustrate network capabilities, the steps described a range of activities from development of initial preparatory to research summary counts to complex analyses (e.g., comparing the effectiveness of different therapies). The first iteration of the use case is presented in Table 1.

Table 1. First iteration of use case for stakeholder discussions

For this example, assume that all participating organizations have agreed to allow authorized queries against their data and return approved results sets. Results range from de-identified summary counts to, upon IRB approval with required agreements, patient-level datasets. Some organizations will allow automated transmission of summary data to approved users; others will require approval before the summary data are transmitted. All will require approval before transmission of patient-level data.

Step 1: Preparatory To Research Request: Summary Counts Based On Patient and Encounter Level Source Files

An authorized user creates a query using the standard interface. The query includes a series of Boolean operations to define a cohort based on clinical and demographic parameters. For this example, the query identifies all users of the two study drugs under comparison. The query runs at each participating site, and result sets are aggregated and returned to the requestor. The results set could give de-identified summary counts of all individuals exposed to the study drugs, stratified by drug, age, sex, and year of exposure.

Step 2: Subsequent Querying of Saved Results

An authorized user submits a query to run against the drug users identified in Step 1. This would allow additional sub-setting of query results; for example results could show comorbidities across the two study drugs.

[Note: the data generated in Steps 1 and 2 are typically used to prepare a research proposal for funding. Once funding or other approvals are obtained, a researcher could undertake the next steps for the planned analyses.]

Step 3: Creation and Storage of Detailed Intermediate Datasets

An authorized user submits a dataset creation query to run against the list of drug users identified in Step 1. The query initiates the creation of intermediate, patient-level datasets that are saved behind each organization's firewall. For this study, the intermediate datasets may be limited to inpatient care for users of the drugs of interest for certain diagnoses codes over a specified period. Dataset creation is only permitted under an approved research protocol; the system must verify that the requestor has obtained necessary IRB approval(s).

Step 4: Processing Analytic Code against Detailed Intermediate Datasets, Summary or Patient-Level Data Returned

Once the data are created as described in Step 3, an authorized user can submit complex analytic code against the detailed intermediate datasets and create final tables for return to the user (either summary tables or patient-level data). The summary tables may include the number of incident and prevalent users, and an accounting of the number of hospitalizations of interest. This activity requires IRB and research approvals. The summary data are typically used to further define a detailed study protocol and assess final sample size and power. The patient-level data only would be transferred if pooled analyses were required.

Based on information obtained during the initial stakeholder discussions, the proposed network architecture, flow of network operations, and use cases were updated and presented during future stakeholder discussions (Figures 2-5). Figures 2 through 5 are a depiction of the current use cases for a distributed network. The term "datamart" is used to distinguish the information available to the network from the local data warehouses maintained by the individual data owners; datamarts are typically sub-sets of institutional data warehouses.

Figure 2. Creation of a network datamart

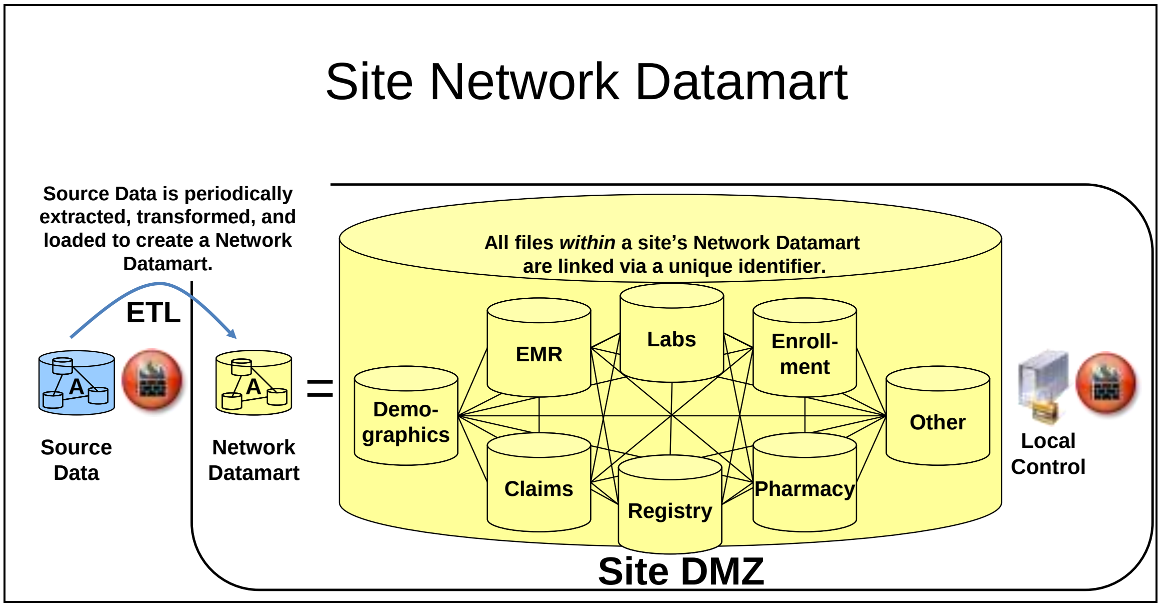

Site Network Datamart
Source Data is periodically extracted, transformed, and loaded to create a Network Datamart.
ETL
A
A
Source Data
Network Datamart
=
All files *within* a site's Network Datamart are linked via a unique identifier.
EMR
Labs
Enroll-ment
Demo-graphics
Other
Claims
Registry
Pharmacy
Local Control
Site DMZ

Figure 2 depicts the most straightforward implementation of a distributed health data network, as each data owner (site) creates and controls a network-usable version of its primary data through an extract-transform-and load (ETL) process using a common data model. Site “A” creates a network datamart (via an ETL process based on the network common data model) from its source data (local data warehouse) and makes the datamart available to the network through the site’s demilitarized zone (DMZ) – a secure area under control of the data owner with restricted access available through the network. The datamart can be physically located in the DMZ, or the DMZ can be used to communicate with the datamart stored elsewhere within the site. The datamart is accessible only to authenticated and authorized end-users through access to the DMZ. Sites are not required to have all types of data shown. “Other” data is a placeholder for new types, such as imaging, or genomic data.

Figure 3. Basic flow of network operations

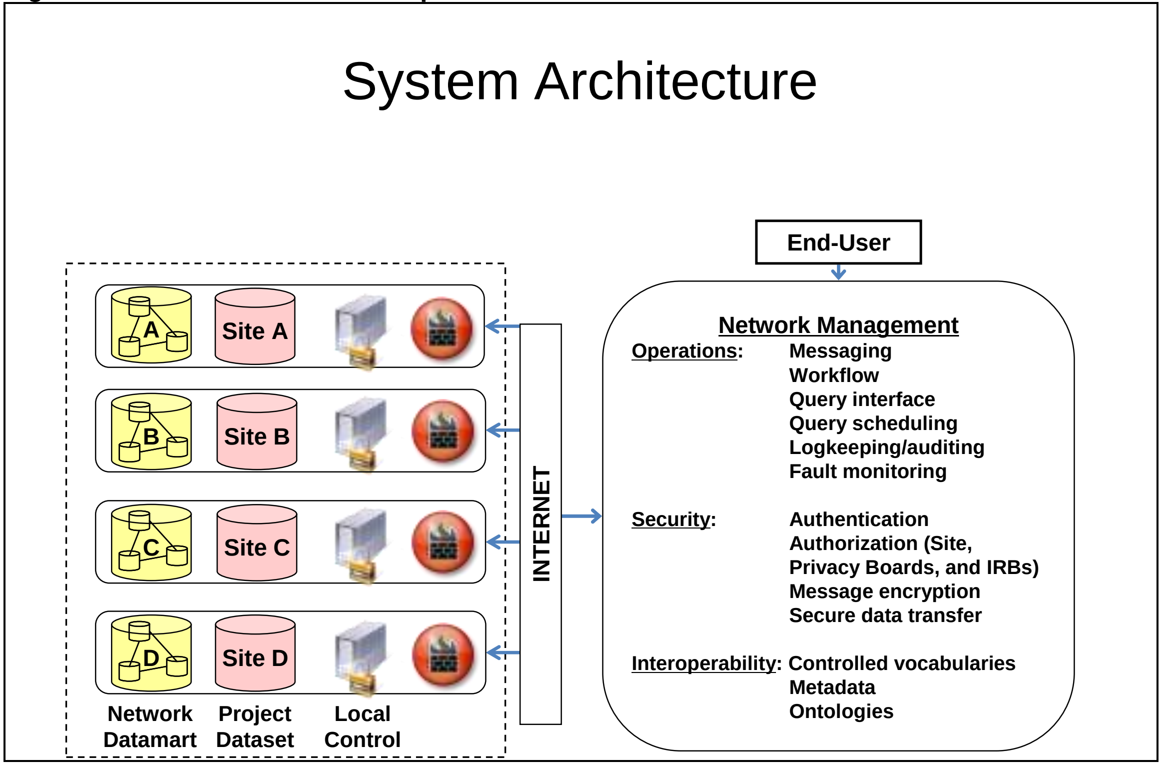
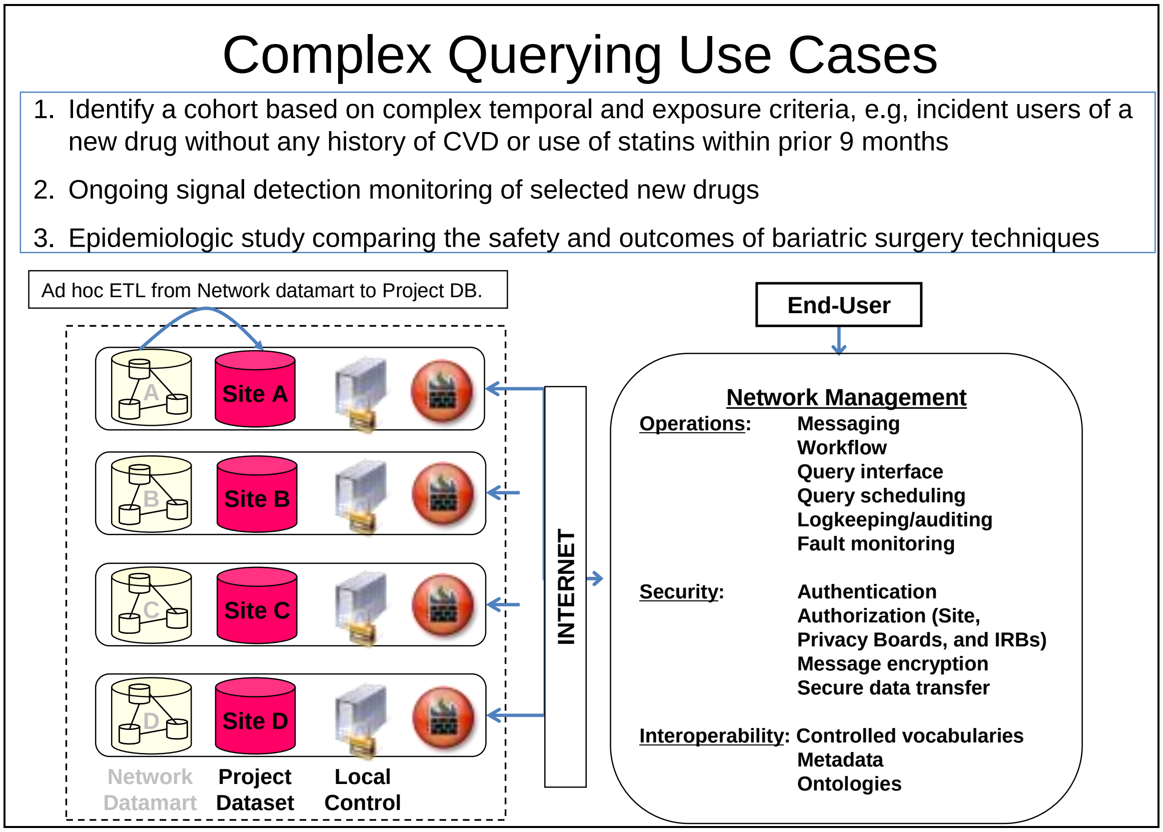


Figure 3 depicts a high-level view of the system architecture, including many of the network management functions identified during stakeholder interactions. The solid boxes show each site's DMZ under local control. The dashed box indicates the resources potentially available to the network (i.e., data from four sites) via management software.

Figure 4. Use cases for stakeholder discussions: complex querying



Complex Querying Use Cases

1. Identify a cohort based on complex temporal and exposure criteria, e.g, incident users of a new drug without any history of CVD or use of statins within prior 9 months
2. Ongoing signal detection monitoring of selected new drugs
3. Epidemiologic study comparing the safety and outcomes of bariatric surgery techniques

Ad hoc ETL from Network datamart to Project DB.

End-User

Network Management

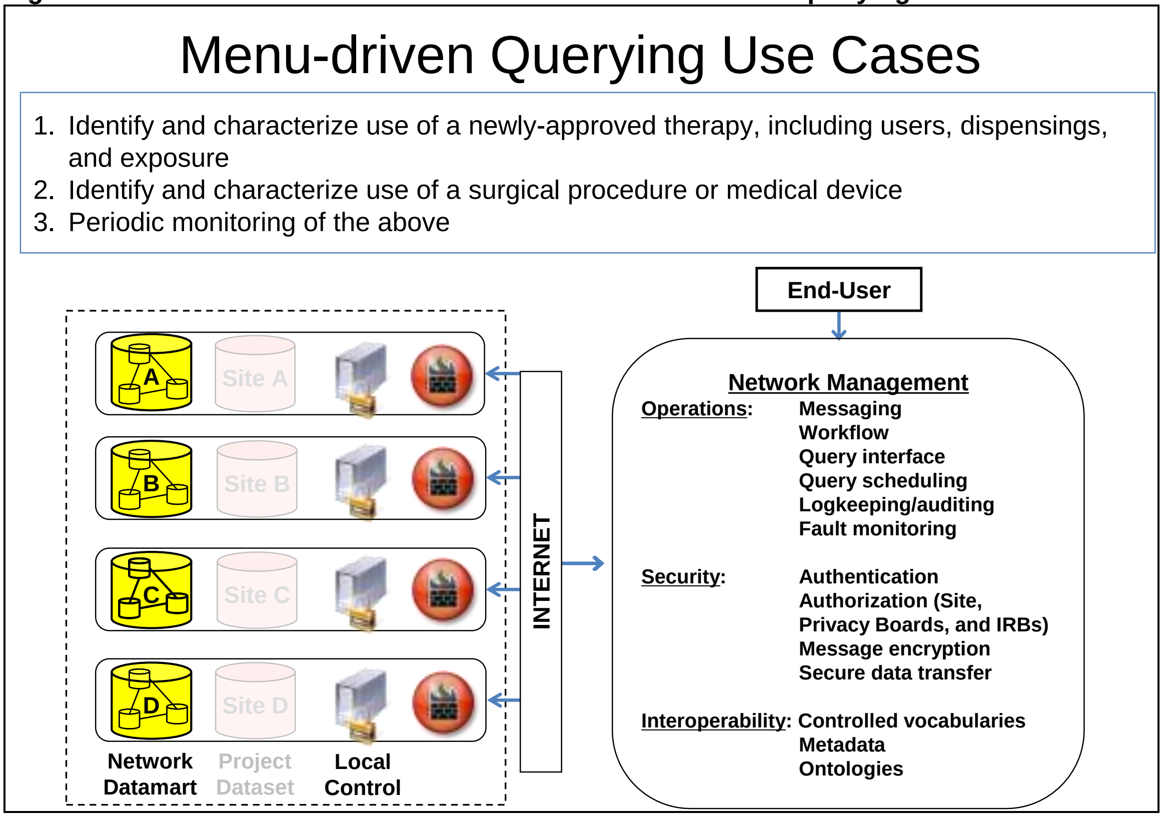
Operations: Messaging
Workflow
Query interface
Query scheduling
Logkeeping/auditing
Fault monitoring

Security: Authentication
Authorization (Site, Privacy Boards, and IRBs)
Message encryption
Secure data transfer

Interoperability: Controlled vocabularies
Metadata
Ontologies

Menu-driven systems cannot easily incorporate complex exposure and outcome definitions or temporal relationships. Figure 4 is used to depict use cases that require complex cohort identification or full-scale epidemiologic studies, and that can run against either the network datamarts or project-specific datasets.

Figure 5. Use cases for stakeholder discussions: menu-driven querying



Menu-driven Querying Use Cases

1. Identify and characterize use of a newly-approved therapy, including users, dispensings, and exposure
2. Identify and characterize use of a surgical procedure or medical device
3. Periodic monitoring of the above

End-User

Network Management

Operations: Messaging
Workflow
Query interface
Query scheduling
Logkeeping/auditing
Fault monitoring

Security: Authentication
Authorization (Site, Privacy Boards, and IRBs)
Message encryption
Secure data transfer

Interoperability: Controlled vocabularies
Metadata
Ontologies

Figure 5 is used to depict use cases that can be addressed using a menu-driven query. In this scenario, a credentialed end-user accesses the system via a web portal to which a user must authenticate (i.e., prove their identity) and be authorized (i.e., have permission to access the portal). The end-user is then able to query available data resources (i.e., network datamarts) based on a set of permissions using a menu-driven query interface.

4.1.2. Development and Distribution of an Environmental Scan

In conjunction with the Overview and Use Case Summaries document, an Environmental Scan ("Scan") was developed and provided to stakeholders to gain an understanding of their organization's research, information technology, and governance needs and expectations with respect to a network. The Scan served to facilitate a discussion about: (1) data owners' specific policies and procedures governing access to and use of data for research purposes; (2) information technology and security issues (e.g., security requirements, architecture, software platforms, analytic capability) related to data access; and (3) analytic capabilities desired by end-users (e.g., research investigators, research staff).

Section 1 of the Scan included questions completed by decision makers at the stakeholders' organizations (i.e., those responsible at strategic levels for establishing and/or enforcing policy to make data available for research). Questions centered on each stakeholder's current participation in distributed data networks, secondary uses of their clinical or administrative data, their organization's participation in multi-center data activities, history of data sharing and associated policies, and their IRB- and HIPAA-related policies.

Section 2 of the Scan was completed by IT personnel (i.e., those responsible at tactical and logistical levels for operationally making data available for research). Section 2 questions provided a structure for describing the capabilities and relevant policies of each organization; questions aimed to assess each organization's data policies and procedures and potential network architectures. To supplement their answers, stakeholders were asked to provide copies of any relevant documentation, such as their data security policies or system user agreement and policies. Finally, advice and views were solicited about any additional technical or governance issues related to practical aspects of implementing the network.

Section 3 of the Scan was completed by research investigators / research staff (i.e., end-users). This section addressed questions related to (1) expected network users within their organization (e.g., for internal queries or multi-center research); and (2) desired analytic capabilities.

To gather information using the Scan, the Informatics team contacted representatives at each targeted stakeholder institution and provided them with the Detailed Network Overview and Use Case Summaries and Scan documents. The representatives were told that the data collected would be used to inform and refine our approach to developing the network architecture. Stakeholder representatives identified colleagues to review the Network Overview and Use Case Summaries document and complete the three sections of the Scan (i.e., Section 1 to be completed by decision makers; Section 2 to be completed by IT personnel; Section 3 to be completed by research investigators / research staff). Once completed, the stakeholder contacts compiled their organization's opinions on the Scan and coordinated their availabilities for participation in a webinar or face-to-face meeting with the Informatics team.

4.1.3. Description of Key Stakeholders

4.1.3.1. Current Partners

AHRQ

As a project collaborator, stakeholder, and end-user, AHRQ provided ongoing input regarding their areas of interest and research needs as they relate to a distributed research model. AHRQ input helped to inform the use cases and the system architecture iterations. The Informatics team travelled to AHRQ (July 2008) to present the system architecture for discussion. AHRQ continues to provide substantial feedback on the network design.

University of Pennsylvania DEcIDE Center

A team at the University of Pennsylvania's (Penn) DEcIDE center reviewed the Overview and Use Case Summaries document, completed appropriate sections of the Scan, and participated in a one-hour webinar with the Informatics team. The Penn team works with the various databases that comprise the Penn Health System's information systems. The Penn database resources include: (1) Medicaid data from five large programs linked to Medicare and the Social Security Administration Death Master file; (2) the Pediatric Health Information System (PHIS); (3) University of Pennsylvania Health System (UPHS) databases, including Pennsylvania Integrated Clinical Research Database (PICARD) and EpicCARE; and (4) Veterans' Affairs Database. The Penn data resources are substantially different (e.g., availability of detailed inpatient data) than the resources found at many of the other health plan partners, thereby providing the Informatics team with a unique perspective of the challenges of designing an extensible and scalable distributed research model.

HMORN Sites

The HMO Research Network (HMORN)[28] is a consortium of geographically distributed health plans with a combined population of 11+ million members, record linkage systems, electronic medical records, and access to both clinicians and members. Six HMORN health plans are participating in this task order and served as the initial test sites for the project. The six health plans are Geisinger Center for Health Research (Danville, PA), Group Health Cooperative (Seattle, WA), Harvard Pilgrim Health Care (Boston, MA), HealthPartners Research Foundation (Minneapolis, MN), Kaiser Permanente Colorado (Denver, CO), and Kaiser Permanente Northern California (Oakland, CA). Starting our discussions with the participating HMORN sites allowed us to focus on essential communications and query capabilities by taking advantage of the HMORN's existing data infrastructure and collaboration agreements.

Most HMORN sites have created a standard research datamart – the Virtual Data Warehouse (VDW). The VDW is a distributed standardized data resource designed to meet the needs of multi-site collaborative research projects. The VDW comprises: (1) computerized datasets stored behind separate security firewalls at each HMORN site, including variables with identical names, formats, and specifications; (2) informatics tools—hardware and software—that facilitate storage, retrieval, processing, and managing the datasets; (3) access policies and procedures governing use of VDW resources; and (4) documentation elements of the VDW. The VDW currently contains standardized data elements for enrollment, demographic characteristics, inpatient and outpatient utilization, diagnoses and procedures, and outpatient pharmacy dispensing.

Staff from each of the six participating health plans reviewed the Overview and Use Case Summaries document, completed the Scan, and participated in a one-hour webinar with the Informatics team.

4.1.3.2. IT Experts (Vendors)

In parallel with the HMORN site discussions, the Informatics team consulted with seven information technology (IT) vendors in order to vet the technical feasibility of the proposed network. The Informatics team held at least two discussions, including one-hour webinars with technical experts from each vendor. Prior to the one-hour webinars, an Informatics team member held an initial call with representatives from each vendor to describe the project, elicit initial reactions, and if possible, setup a follow-up call with the entire Informatics team.

The Informatics team presented the vendors with use cases (Figures 3, 4 and 5) to demonstrate the proposed uses and capabilities of an ideal system. The vendors were asked to discuss the architecture and tools that would be needed to build such a network. Some specific issues discussed included the following:

1. How important is the number of expected users?

2. What are some approaches to crossing firewalls?

3. What are some approaches to permissions (centrally and/or locally [data owner] controlled)?

4. What are some approaches to authorizations/authentication/security/audit trails?

5. What is an approach to aggregation of results sets?

6. What are some approaches to dealing with delays / lags in query results across sites (some will get results in 2 minutes, others may take 2 days, others will have their system down)?

7. What are some approaches to local control of timing of queries? (i.e., Demand queries wait until midnight to run? Delay queries if system use is high?)

8. What is the importance of how data are stored (DB2, Oracle, SQLserver, SAS, etc.)? Do they need to be standardized?

9. What are some pros and cons of ETL to a common data dictionary versus publication of local data schemas for mapping of queries back to the local schema?

10. What should be centrally controlled versus locally controlled? How should those controls be implemented (i.e., how much local effort required versus central effort)?

4.1.3.3. Potential Future Partners

AHRQ facilitated a meeting held at the FDA (September 2008) with several potential future partners to discuss the task order and present project accomplishments and future activities. Proposed attributes, capabilities, and uses of the network were presented, and examples were provided to illustrate its potential use with respect to studying diseases, treatments, and outcomes, and performing simple rates and case mixed adjusted comparisons. Presentation attendees included representatives from the FDA (including the Center for Biologics Evaluation and Research [CBER], the CDC, the Center for Drug Evaluation and Research [CDER], and the Critical Path Initiative), Centers for Medicare and Medicaid Services (CMS), the Department of Defense (DoD), the Office of the National Coordinator for Health Information technology (ONCHIT), the National Cancer Institute (NCI), and the United States Department of Veterans Affairs (VA). The goal of the meeting was to engage with potential future partners and elicit initial reactions regarding the proposed technical design, their desired analytic capabilities, and potential implementation challenges (technical- and governance-related).

5. Stakeholder Analysis: Findings

This section describes a summary of discussions with the stakeholders. The needs of the stakeholders and the level of importance for those needs will directly impact the software architecture and overall feasibility of creating a distributed research network.

5.1. Findings from Current Partners: AHRQ

Discussions with AHRQ representatives served to ensure that the network will be capable of supporting evidence development and new research studies for questions about comparative effectiveness within the DEcIDE and Center for Education and Research on Therapeutics (CERTs) programs. Further, discussions centered on the network's ability to support the research needs of AHRQ's Effective Health Care program and AHRQ's many partners. Examples of required research capabilities include:

a. Support systematic reviews by the Evidence-based Practice Centers (EPCs)

b. Fill evidence gaps identified by systematic reviews of the literature

c. Measure the impact of the Eisenberg Center initiatives on patients, providers, and policymakers

d. Address the needs of other AHRQ partners

Key discussions involved the ability to facilitate comparative effectiveness research, incorporate data from diverse data streams (e.g., registry data, EMRs, patient-reported outcomes, costs), and also the ability to develop novel data collection mechanisms (e.g., point of care systems) within a network. AHRQ also was interested in the potential to accommodate fully distributed analytics, such as conducting logistic regression analyses without requiring a centralized dataset. Issues related to data quality, data checking, and cross-site data consistency also were highlighted.

5.2. Finding from Current Partners: HMORN Sites

After reviewing the use cases and Scan, the HMORN sites identified and discussed their organization's research, information technology, and governance needs and expectations with respect to a distributed research model. Discussions centered on uses and capabilities of an ideal system and facilitators and barriers to implementation, use, and maintenance of a system. These partners were especially concerned with data autonomy and security, and expressed great interest in having the ability for fine-grained (e.g., person-level as opposed to institutional-level permissions) security and authorization and strong authentication. The HMORN partners want the ability to review all network requests as the requests arrive and before the results leave the organization; a system that includes detailed workflows and approval mechanisms was suggested by the Informatics team, and this approach resonated with most plans. These plans also felt that detailed auditing and active monitoring of network use would be valuable.

Questions also were raised about cross-institutional trust for network authentication (i.e., trusting that network institutions will maintain the security of their networks and tightly manage identities). Because most network designs would require some level of institutional trust to verify identities (e.g., site A would have to trust that a person with a site B login certificate is really an authorized site B user), it is very important that participating institutions maintain the security of their networks. The example of how well an institution restricts local network access to a newly fired employee was raised by multiple plans. Many plans also raised questions about implications for IRB approval, control of data access and use, and the level of local management and oversight needed to maintain involvement in a research network. It became clear that strong internal advocates would be needed to promote participation in a network.

Although most of the HMORN sites' reactions were consistent, there was some divergence of opinion, procedures, and policies. Some partners preferred a publish-and-subscribe architecture in which sites would periodically "look" in a pre-specified place within a central portal for appropriate queries published by a central authority and pull those queries back through their firewall. This approach obviates the need to open a port into their local systems to accept distributed queries. Other sites felt comfortable opening and securing a port to receive external messages (e.g., queries). The health plans differed regarding their internal policies and procedures for evaluating, approving, and implementing a system such as the one described, and there were some differences in the availability of the necessary IT expertise to implement and manage the required system architecture. A summary of the overall comments from the HMORN sites is presented in Table 2.

Table 2. Overall findings from HMORN sites

Major Areas to Consider	Comments from Participating HMORN Sites
Security	• Permission: important to know who can access the data; recommend local control of permissions that are fine-grained • Software: concerned about compatibility at various sites • Allowing a hole in the firewall would require a major effort, as sites' IT departments are likely to deny requests to do so • Concerned about unauthorized users/hackers • Web services function capabilities: some sites would prefer to "look" for queries published by a central authority and pull those queries back to be run locally (publish and subscribe) • Issues may arise if sites are required to execute a code that is not trusted (e.g., sites might not allow running another site's SAS code) • Auditing and monitoring: need for local workflows to allow for approval of incoming messages and outgoing information
Data Issues	• Expect/require intellectual input • Research process cannot impede clinical care • Voiced the need for local knowledge in order to ensure proper use of data; easy to misunderstand and misuse the data • Some data elements are sensitive; recommend controlling access to proprietary data (e.g., enrollment patterns) and private data (e.g., mental health, HIV, alcohol dependence) • Must adhere to state regulations regarding data access • No requirement to store the site datamart in the DMZ
IRB Process	• Need to educate each local IRB about infrastructure • Can envision global IRB approval for certain queries, but likely difficult for other queries • Anticipate that certain levels of queries would require specific IRB

Major Areas to Consider	Comments from Participating HMORN Sites
	approvals • Intent for research is very important: preparatory to research vs. research; research would require full IRB submission • Likely no need for preparatory to research form if no personal health information is involved
Governance	• Discussed issues, such as who will be able to access the system, the need for standard operating procedures, how data requests will be administered, and prioritization of multiple network requests • Highlighted that site decision-makers will need to know *why* queries are being made • Remarked that governance issues will likely prove more challenging than technical issues
External Support	• Expressed the need for a stable, long-term funding mechanism for the network • Sites are more likely to invest in a longer-term system
Internal Support	• Voiced concerns about the need to identify resources in order to participate (e.g., IT, infrastructure, system management, and oversight)

During discussions with the HMORN stakeholders, the Informatics team specifically assessed four key areas: (1) desired network uses; (2) preferred capabilities; (3) facilitators to implementation, use, and maintenance of system; and (4) barriers to implementation, use, and maintenance of system. A summary of the findings from those discussions is presented in Table 3.

Table 3. Comments from participating HMORN sites regarding potential network uses, capabilities, facilitators, and barriers

Four Key Areas assessed by Informatics Team	Participating HMORN Sites' Comments
Desired Uses	• Ability to share information • Solve governance issues currently dealt with on an ad hoc basis • Routine and complex preparatory querying: - Menu-driven queries for feasibility counts - Complex querying when menu-driven is insufficient • Complex analytics (e.g., comparative and clinical effectiveness) • Monitoring and surveillance: periodic monitoring of trends in utilization; adoption and diffusion evaluation

Four Key Areas assessed by Informatics Team	**Participating HMORN Sites' Comments**
Preferred Capabilities	• Source data never leave site • Each site is able to check any query before it is run and again before it goes out • Site control over permissions for access and use of their data • Different levels of querying: - Menu-driven, such as: 1) identification and description of cohorts; and 2) monitoring the uptake and use of a medical intervention - Complex, such as: 1) comparative effectiveness; 2) excess risk of a medication-related adverse event; 3) active surveillance for adverse drug events; 4) identification of statistically unusual clusters of illness • Sites able to define different data access policies (build capability into system) • Prefer automated returns, as programmer support to do preparatory-to-research is expensive; however, some sites want the option for human checks at any point in the process, especially as they learn to use the system
Facilitators	• Different levels of queries • Different types of partners (trusted vs. not trusted) • Control over data • Sites to participate intellectually and financially in uses of their data • Fine-grained workflow control (i.e., ability to have person-level as opposed to institutional-level control over which queries are run) • Ability to approve all users, at least initially • Querying summary data less sensitive • Web services function capabilities: - Publish and subscribe architecture (i.e., “look” for queries published by a central authority and pull those queries back to be run locally): some would prefer to never open the port in order to eliminate exposure of data to potential hackers - Others prefer to open a port and defend it • Menu-driven queries; ensure only authorized persons can drop a query • Obtain standard IRB approvals for certain types of queries
Barriers	General: • Hesitancy on the part of health plans to participate in an area outside of their business needs
	Security: • Unfamiliarity with technology needed to implement a network; need to establish acceptance of and comfort with technology and infrastructure • Fear of unauthorized users/hackers

Four Key Areas assessed by Informatics Team	**Participating HMORN Sites' Comments**
	• Discomfort in execution of someone else's SAS code; difficult to prospectively assess what the code will do
	Data issues: • Need to establish comfort with performance of querying mechanisms (e.g., menu-driven queries) • Big concern is making sure sites know why queries are being made, as most political issues are related to "why" • Reluctance to allow access to enrollment information • Size of queries
	Costs and resources: • Anticipate high use of resources • Who will host the web portal? • Anticipate high upfront cost of data extraction system testing and data checking
	IRB: • Need to establish practical and minimally onerous rules for data access • Need to gain acceptance from local IRBs
	Governance challenges: • Engendering the cooperation, collaboration, and trust of all stakeholders in a unified solution • Overcoming proprietary and legal liability fears associated with data-sharing and reporting • Ensuring compliance with disparate patient privacy laws and IRBs • Establishing rights to access and use data

5.3. Findings from IT Experts (Vendors)

The Informatics team consulted with vendors of information integration systems in order to vet the technical feasibility of the proposed network. The team sought the vendors' guidance regarding the architecture and tools needed to build a network and potential challenges.

The vendors confirmed that examples of major network functions exist in other industries (e.g., telecommunications, travel, insurance), that the proposed architecture is technically feasible, and that all major stakeholder questions and concerns (as noted above) can be addressed through software design and governance policies. Each of the vendor teams cautioned that governance and policy-making will be time-consuming and will likely pose the biggest barrier to implementation. Considering the size and complexity of the proposed network, they recommended a highly phased approach to implementation. Finally, discussions centered on technical challenges, including:

- Integration of dissimilar databases into a cohesive and secure network

- Management of change
- Maintenance of security levels and dynamic permissions
- Variation in data owner's technical, human subjects, and IT expertise and policies
- Disparate data quality and inconsistent data availability
- Variation in coding practice

5.4. Findings from Potential Future Partners

Meetings were held with potential future partners to discuss the proposed technical design, desired analytic capabilities, and potential implementation challenges (technical- and governance-related) of a distributed research model. Potential future partners talked about the need to be attentive to variation among data owners (i.e., technical capabilities, human subjects' protection procedures and interpretation of regulations, IT expertise and policies, state regulations), disparate data quality, inconsistent data availability, and variation in coding practice.

Potential future partners brought up several possible challenges. Data-related challenges presented by the groups included establishing rights to access and use data, overcoming the proprietary and legal liability fears associated with data-sharing and reporting, and dealing with sensitive data (e.g., HIV, mental health, alcohol dependence). These groups also talked about the need to ensure compliance with disparate patient privacy laws and IRBs. Additionally, discussions centered on the challenges related to engendering the cooperation, collaboration and trust of all stakeholders in a unified solution as well as funding and supporting a comprehensive governance structure.

6. Technical Design of Distributed Research Network

The design and architecture of the proposed Distributed Research Network (DRN) described in this section is informed by the analysis of existing systems, the stakeholder analysis (Section 5), and our review of distributed research model features (Appendix B). In addition, the technical design defined here addresses the key characteristics of a network. The scope of the proposed DRN design and architecture is defined by a high-level network architecture, system hardware and software requirements, and security policies.

6.1. Applications and Functions

The DRN design and architecture will support several functions. These include study feasibility activities, such as identification of individuals who satisfy specific clinical, demographic, or health care plan enrollment criteria. In addition, users will be able to apply complex inclusion and exclusion criteria, including specification of temporal relationships, for creating study datasets to address specific research questions. The DRN design will also incorporate a mechanism to save queries distributed through the network, thereby allowing those queries to be executed at a later date, modified, or used as a reference for other users studying similar questions or using similar methods. By providing for the creation and storage of study-specific datasets and distributed queries, the DRN design addresses the need to re-create or revise analyses (reproducibility) using either the identical dataset used in the initial analysis, or creating new datasets based on the exact criteria used in the prior analysis. Because the data available through the network are expected to change often, these capabilities to reproduce analyses are crucial. Finally, the methodology for implementing distributed multivariate analyses has been described elsewhere and could be implemented through network infrastructure.[29,30]

6.2. Security, Network Governance, and Policy Rules

The role of the network portal system is central to the security and enforcement of policy of the network as a whole. All users must use the network through the portal, and the portal system is responsible for authenticating users, enforcing governance and permissions policies, and tracking and auditing all uses of the network. The DRN design allows for data owners to apply additional policies and procedures above those set at the network level; for example, a data owner could choose to set a policy in which all requests from a specific authorized user or from a specific institution are pre-approved for execution while requests from other users require manual review.

6.3. System Architecture—High-Level View of the Network

The DRN design supports the following capabilities:

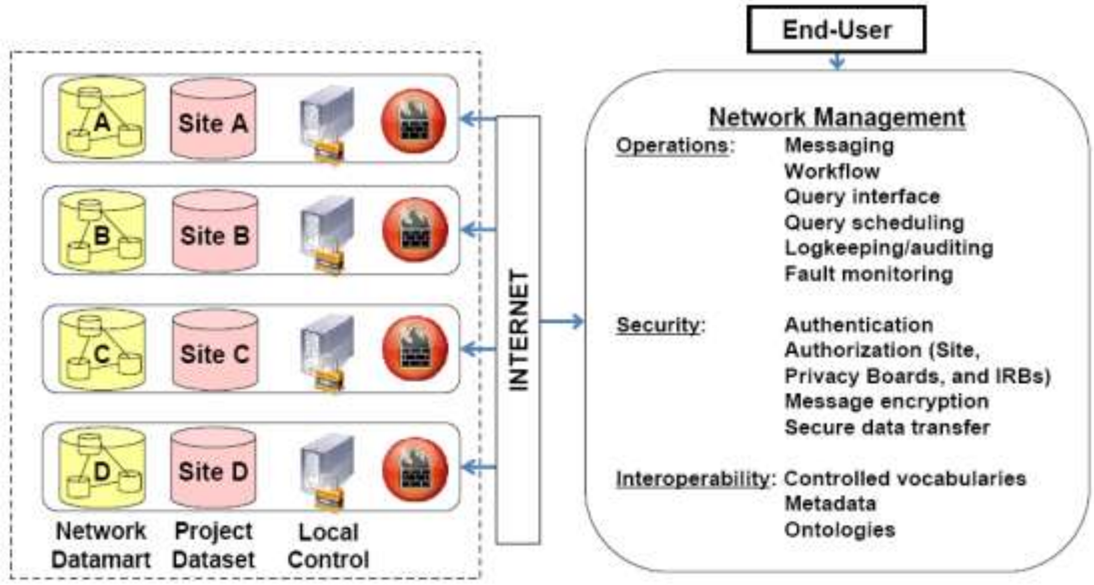
- Accommodate data heterogeneity
- Provide secure communications and data protection
- Support auditable processes
- Provide simple query interfaces that enable menu-driven and complex queries
- Enforce fine-grained, locally-managed security, authentication, authorizations, and permissions

Six basic principles derive from these capabilities and inform the overall design of the network. First, it will employ a distributed architecture, in which each data owner will maintain local control over their data and other resources, such as analysis software. This approach avoids the need to centralize confidential or proprietary data. Second, each data owner will maintain control over all uses of their data, and will be responsible for establishing and enforcing its own policies and procedures with respect to data access, and user and use audits. Third, the architecture will allow incorporation of multiple types of data from a range of organizations. Fourth, queries and analyses will be standardized to ensure identical implementations of research protocols across the network. Fifth, the network architecture will support any kind of statistical analysis through the distribution and execution of SAS or other programs. Finally, a network portal will be employed to distribute queries, aggregate results, and maintain auditable logs of network use. The portal will support a menu-driven user interface for creating simple or recurring queries.

A client-server architecture with a central portal (also known as a hub-and-spoke design), as opposed to a peer-to-peer design, was selected for several reasons. 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. Conversely, in a peer-to-peer network, all network nodes are connected to each other, and each functions as both a client and a server. From a security perspective, the client-server single hub model is more familiar to data owners and permits a simpler and more stable set of firewall rules and configurations (e.g., all communication will come from a single location that is known). The client-server architecture with a central portal minimizes data owner IT responsibilities, provides for a more straight-forward security implementation, and focuses network management tasks at the central portal.

A high-level graphical schematic of the proposed system (Figure 6) provides a less technical view of the network architecture. The overall technical design of the DRN illustrated as a set of graphical models, using the Unified Modeling Language (UML) convention, is provided in Appendix C.

Figure 6. Proposed architecture for the Distributed Research Network



6.4. Hardware Requirements

The hardware requirements of the DRN design are straightforward. A network server and a high-speed Internet connection will be required to support the central portal, as well as a firewall and backup systems for redundancy. No change in hardware configurations by the individual data owners is anticipated. This hardware will support the activities and software environment shown in Figure 6 (see the balloon labeled “Network Management”). Hardware requirements for the individual data owners are also straightforward and will not require the acquisition of additional equipment.

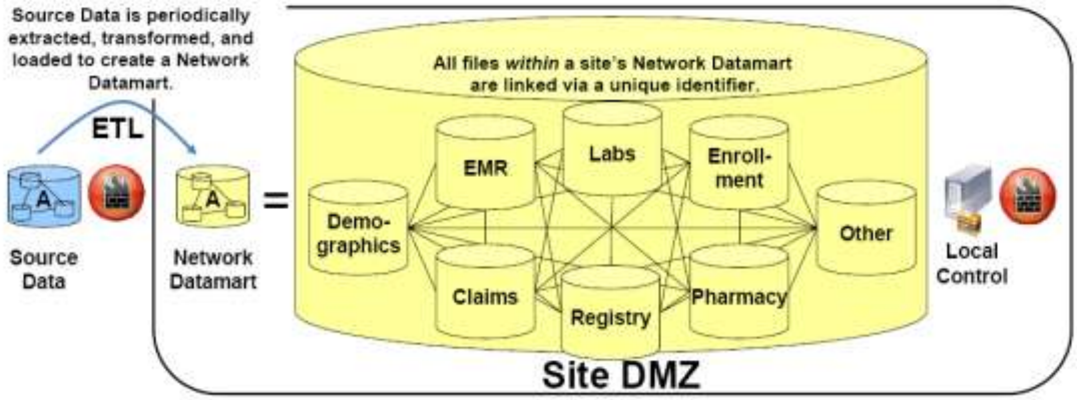
6.5. Site Database Requirements

Data owners can store their data in any standard database platform (Oracle, DB2, MySQL, and Sybase, as well as text-format files), but because the DRN design anticipates distribution of SAS programs for at least some network functions, storing data in SAS is preferred. Data owners would be required to maintain a SAS license to allow execution of distributed SAS programs.

6.6. Network Datamart

As shown in Figure 7, each data owner will support one or more Network datamarts, which will have been created from source data through ETL procedures. The datamarts will contain data from a variety of sources, such as electronic medical records, laboratories, billing, and pharmacy, and the datamarts will be positioned either physically or virtually within a "demilitarized zone," or DMZ, which is between a firewall protecting source data and second firewall protecting the datamart from the outside world. These firewalls will be implemented through usual procedures. Figure 7 illustrates the creation of a network datamart.

Figure 7. Schematic of the typical network datamart.


Source Data is periodically extracted, transformed, and loaded to create a Network Datamart.
ETL
A
Source Data
A
Network Datamart
=
All files *within* a site's Network Datamart are linked via a unique identifier.
EMR
Labs
Enroll-ment
Demo-graphics
Other
Claims
Registry
Pharmacy
Local Control
Site DMZ

6.7. Network Management

Developing the central portal and the data owner's local implementation will involve substantial software development. A goal of the DRN design is to provide each data owner with complete control over access and use of their data, and also to have a robust central portal that handles all network-wide functions. At a minimum, the DRN design has the following functional requirements: (1) authentication, (2) authorization, (3) confidentiality, (4) data integrity, (5) manageability, and (6) accountability. Authentication is the process by which an end-user identifies himself. Authorization is the process by which an end-user gains access to various resources based on role management and/or specific permissions enacted through unique policies (i.e., can person A perform action B on resource C?). Confidentiality is the protection of messages, protocols, codes, or results during transport so that only those with authorization may access files (i.e., if an unauthorized user obtained a confidential communication, encryption would prevent true access). Data integrity protects all transactions during transit from corruption. Manageability refers to a security architecture that is governed by various management policies, such as user management and authorization policy management that can be easily translated into automated functions. Accountability is the process of secure logging and auditing of all transactions on the network that can be easily monitored on a periodic basis as well as in an automated way to alert sites to circumstances that denote unusual activity. How these functional requirements are incorporated into the DRN design and architecture is described below, along with some additional network management functions. Each item detailed below is an essential facet of the DRN design and architecture. The descriptions below reflect the suggested specifications for the DRN.

Auditing. The DRN design includes detailed auditing capabilities. All network access should be audited in real-time, and a description of that event (date and time of log-on and log-off, user, sites accessed, query reference, query status, and other characteristics) will be logged on the network server. Reports derived from the audit log file will be reviewed regularly to identify inappropriate use and will be made available to network participants. These reports will characterize general activity and assist with identifying use exceptions (such as failures to log off, unexpected patterns of access, or query status patterns that may suggest inappropriate querying). The audit reports will be used to inform the continuing development of the network.

Authentication. Authentication is the process by which an end-user identifies himself and is confirmed as an approved user by the network. Examples of authentication include passwords, use of special hardware and software, and use of biometrics (i.e., retinal scan or fingerprints). The design of the authentication process is a function of the number of short-and long-term expected users and demands of the data owners. The DRN design includes flexibility for various authentication procedures as could be demanded by network participants. A minimum requirement is password-protected access coupled with authentication of the user's valid credentials from his local institution (e.g., verification that a user from Participant Site A is currently logged in to the protected network of that participant).

Authorization. Authorization is the process by which an end-user gains access to system resources based on role management and/or specific permissions enacted through unique policies (i.e., can person A perform action B on resource C). Authorization is usually performed after a user has been authenticated and will be applied to each user of the network; no user can access the network without having gone through the authentication and authorization processes mandated for the network as policy. The capability for fine-grained authorization – for example, user-level permission stratified by all network capabilities – is part of the DRN design. At a minimum the initial design should include authorization at the institutional level (e.g., all users from institution A can perform certain specified functions as determined by network governance agreements). More fine-grained authorization can be incorporated as needed.

Permissions. Although users will be authenticated prior to network use, they need to be granted access (i.e., permissions) through previously defined governance procedures and tracked in the network portal system. For example, one user may be able to submit a query that performs a detailed, stratified count-based analysis, while another user may be restricted to queries that return only aggregated, high-level counts. Initially, permissions within the system could be defined at the institutional level (e.g., all users from Institution A have the same permissions) with more fine-grained permissions (e.g., user A at Institution A has different permissions than user B from the same institution) incorporated as network use grows.

Monitoring. All activity on the network should be monitored, primarily through the use of robust auditing procedures, as described above. In addition, the DRN design will enable the network performance to be monitored to evaluate potential bottlenecks in the efficiency of query submission, execution, and results dissemination. This activity will be reported from monitoring logs that will be created automatically by the network system in the background to ensure that this monitoring process does not interfere with use of the system. Monitoring could also include complex rules for system use to prevent inappropriate system use, such as multiple sequential queries against the same data from the same user.

Workflow. Workflow refers to a system by which queries are distributed and approved within the DRN design. The DRN design includes capability for simple or complex workflow rules that could, for example, allow data owners to approve queries before being run using their data and approve results before distribution back to the central portal for aggregation. Workflow can include algorithms, based on the user authorization level and the type of query, that automatically approve some queries but require manual approval for others. Initially, the DRN should include rules that allow data owners to approve all queries and responses. Those approval rules can be adjusted, as needed, based on growing use and comfort with the system operations.

Scheduling. As the number of users and query submissions increase, it is important to address the need to establish a queue-based (first-in, first-out) approach to query scheduling. Although it is not anticipated that a need will exist for scheduling users and query submissions in the near-term, a method for scheduling these activities to ensure system usability and timeliness will be included in the DRN design.

Query synchronization, aggregation, and obfuscation. Once a query is created it will be sent to each data owner to be run within their local environment. A single query may take longer to run within some local environments than others. This poses a potential problem for query aggregation, a key function of the network portal system. The DRN design will allow rules for query aggregation to account for asynchronous responses by data owners; for example, query results could be aggregated as soon as one-half of the data owners respond (with a minimum of three) or at a set interval. The DRN design will also allow a mechanism to obfuscate results to blind the user from knowing which data owners responded to any particular query. This functionality may be needed in certain scenarios in which the user may not have network or data owner permission to see data-owner specific information. The degree of query synchronization and obfuscation of results will depend on the governance decisions.

Encryption. The DRN design requires all queries and query results, as well as all network portal system access transactions (such as entering a username and password) to be encrypted by the system using 128-bit public-key encryption. This will minimize the possibility of interception of free-text information over the Internet as well as at the portal system server and components.

Query formation and translation. Query formation and the degree to which queries will need to be translated will depend on the short- and long-term implementation and governance decisions. The DRN design can accommodate various options for query formation and translation. Requiring network data owners to transform their data to a common data model using an ETL procedure, as suggested in the DRN design, obviates the need to translate queries and greatly simplifies the mechanism for query formation. Although the DRN design allows for data owners to participate in the network without transforming their data to the common model, allowing this type of participation would substantially complicate query formation and translation mechanisms and the overall DRN design, and require complex metadata, ontology, and data mapping activities.

Metadata. All data structures, definitions, and sources will be maintained in a comprehensive data dictionary maintained on the network server as a repository for ease of access and updating, as needed. Included in the data dictionary are data owner-specific definitions of file structures, field names and descriptions, field value domains, and other data essential to the query translation process. Requiring network data owners to transform their data to a common data model using an ETL procedure, as suggested in the DRN design, greatly simplifies creation of the data dictionary and the system ontology because all data owners would conform to the same data dictionary. A key document in the metadata repository is the controlled vocabulary, which establishes the terms that can be used in a query.

Syntactic and semantic interoperability. Understanding and addressing these concepts is crucial when considering a network design that queries disparate data systems. Syntactic interoperability arises when two or more systems use the same terms, rules, and values to describe the same concept. For example, if two systems code the term “sex” as 1=Female, 2=Male, they are syntactically interoperable on that term. Semantic interoperability results when two or more systems use the same term to describe the same concept. For example, “gender” can mean different things in different systems. If two systems are semantically interoperable, they use “gender” to describe exactly the same concept; the system would not be interoperable if one used “gender” to describe “sex” and the other used “gender” to describe the sociological concept of “gender”. Once again, requiring network data owners to transform their data to a common data model using an ETL procedure, as suggested in the DRN design, addresses both syntactic and semantic interoperability through the specifications in the common data model. That is, the common data model precisely defines all terms, rules, and values and requires the data owners to transform their data within the confines of the common model. Therefore, the term “sex” would have a common meaning across the DRN and a standard set of expected values (e.g., 1=Female, 2=Male, 9=Unknown) and rules (e.g., missing values not permitted, 1, 2, and 9 only permitted values). Conversely, allowing data owners to participate without conforming to the common data model would substantially complicate the network design and implementation steps by requiring complex metadata, ontology, and data mapping activities to address data heterogeneity.

7. Constitutional and Organizational Management of a DRN

A large-scale distributed research network as described in Section 6 will require a substantial investment in administrative infrastructure along with the investment in information technology. All of the information technology experts contacted for this project warned that that system governance would be the biggest barrier to implementation of a successful and functional network. The administrative 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. Although some of the administrative tasks can be accomplished or facilitated through information technology (e.g., automated auditing and monitoring systems), other tasks will require dedicated network personnel.

7.1. Potential Organizational Structure

An organizational model that allows input from all stakeholders and addresses the multitude of governance and organization issues that will arise as part of a broad-based network is envisioned. Without strong support from stakeholders, implementation and maintenance of a DRN will be a substantial challenge. Beyond some form of leadership/executive oversight group, it is expected that there will be the need for a series of working groups and committees in areas such as information technology and interoperability, health care data standards, membership, proposal evaluation and prioritization, data access and integration, methodology, publications, protection of human subjects, and finance. Governance and organization issues to be addressed include development of by-laws and operational guidelines and procedures.

7.2. Business Case

Developing a persuasive business case primarily involves convincing data owners that the benefits of participation outweigh the real and potential costs of participation. Benefits of participation include contributing to the advancement of scientific knowledge about the real-world costs and benefits of medical interventions, the potential direct financial benefit of participating in funded activities, and the ability to better understand their own data and conduct local analyses using network resources. Improving the understanding of costs and benefits of interventions may improve business practices for data owners in their role as providers and/or payers.

The risks of participation for data owners are substantial and must be mitigated in the development of a broad-based network. The most obvious risks relate to the protection of PHI and adherence to state and federal regulatory requirements for research involving electronic health data. Any unauthorized release of protected information would be a serious set-back for a network and could result in substantial penalties and fines. Additional concerns relate to the release of proprietary information, such as enrollment counts and trends, payment and charge information, and health insurance product coverage information (e.g., formularies).

Other industries have developed and support information exchange networks; however, these industries have a business need to share information. Although the creation of a DRN has many advantages, there is no clear business need for data owners to share information, and many reasons data owners are wary of data sharing. The business case for a DRN must be clearly articulated, and the concerns of data owners should be at the forefront of the information system design of the network. Without a clear and convincing business case, data owners will be wary of participation and a network will become impractical.

7.3. Risk Mitigation and HIPAA Considerations, PHI

The implementation of a DRN creates risks for the participants, especially the HIPAA covered entities that have a responsibility to secure and protect PHI. Most risks involve the unauthorized release of protected health information, either through fraud, abuse, or criminal activity, or via lax oversight and adherence to relevant state and federal regulations. Strong security procedures, effective administrative controls, local data autonomy, and effective systems for auditing and monitoring can substantially mitigate potential risk and help persuade data owners that there are sufficient safeguards in place to protect against the unauthorized or unlawful access and use of their data. Keeping data under the control of data owners and not creating a central data warehouse mitigates the potential for a large-scale release of PHI, either by accident or through unlawful activity. Although these DRN design approaches should mitigate risks, if some large data owners do not accept the level of risk mitigation and decline participation, development of a network will be impractical.

7.4. IRB and Human Subjects Review

Most activities envisioned by the DRN clearly fall under the purview of IRB and Human Subjects Committee review. It is expected that all data owners will maintain a close working relationship with an appropriate IRB and that all network activities will be approved by the appropriate review boards. It is expected that some network activities, such as menu-driven queries for preparatory to research tasks, may receive blanket pre-approval from all or most data owners. Anything that falls within the category of human subjects research will require IRB approval.

The network architecture can greatly assist with ensuring that network activities are consistent with the required approvals. For example, many categories of menu-driven queries could be pre-approved with a blanket IRB approval, but all other queries would require each data owner to approve a protocol and log their approval of the protocol with the network before a query could be executed. In addition, network workflow systems could require that all information leaving a data owner be checked manually, including verification of all necessary human subjects approvals.

As the DRN matures, it is expected that centralized IRB approvals would be possible following a model in which data owners cede authority to a single IRB for a study. A system of centralized IRB approvals enabled by ceding authority is currently in use within the HMORN and is planned for other networks. Once again, data owner acceptance of the validity and trustworthiness of these approaches is crucial for DRN implementation.

7.5. Legal and Proprietary Issues

DRN data owners maintain the responsibility of protecting their data, that is, the data owners should only allow network access to data that they can legally use for research purposes, and must purge from the data any information that members have requested not be used for research purposes. These restrictions and requirements are not unique to a DRN and would be undertaken for any secondary use of data. Potential additional issues to address include anti-trust policies and intellectual property concerns in the event that the network develops new intellectual property. Data owners also must agree to the adequacy of network safeguards against competitors accessing their proprietary data.

7.6. Oversight and Management—Audit Trails and Checking

Audit trails and monitoring of access and use of data will be paramount for the network. Data owners, network administrators, and the network committees and boards must maintain strong oversight of all network utilization, including who is accessing the network and their own data, the purposes of network access and data use, and identification of suspicious network use or data access. Audit trails often benefit data owners because they can use the network auditing resources to manage all access and use of their data for secondary purposes, thereby providing a central mechanism for this important monitoring activity.

8. Study Limitations and Implementation Barriers

Potential limitations of the report include biases related to our experience within an existing research network (the HMORN) and our inability, due to study period and scope constraints, to more fully assess competing strategies for developing a distributed research network. For the purpose of this report, a selection of leaders in the field was consulted; a complete environmental scan of all types of software and vendor approaches could not be conducted. However, the consistency of the responses from the range of vendors and stakeholders provides some confidence that the findings represent a broad consensus of the potential approaches to developing a distributed network. In addition, our experience within the HMORN, particularly with respect to our roles in our respective health plans and our knowledge of health plan data, may have biased our views as to the relative importance of healthcare encounter and claims data versus other types of healthcare data. To address this concern, numerous examples of other types of data (e.g., survey data, registry data, and genetic information) that could be included in a functioning network were included.

The main barriers to implementation of a network such as the ones described in Section 7 relate to governance and design decisions, implementation strategy, funding, and data owner participation. Sections 6 and 7 and Appendix B address implementation trade-offs for many of the DRN design attributes. The primary concern or barrier regarding DRN implementation is designing a network that will be acceptable to the data owners; without adequately addressing data owners' needs regarding data protection, privacy, and security, or mandating participation, a network will not be possible. In addition, a phased implementation approach is most likely to yield success as compared to a single large effort to create and implement a fully functioning network all at once. Further, design and governance decisions, such as allowing data heterogeneity instead of requiring an ETL procedure to a common data model, will have a substantial impact of the feasibility of a network and an implementation timeline; governance decisions should be made with an understanding of how the decision will affect network implementation and viability.

9. Proof-of-Principle Prototype Design

A proof-of-principle will help demonstrate some of the DRN functionality by illustrating how a software system can submit a query to multiple data sources and aggregate the results. As noted in Section 2 above, a study on the comparative effectiveness and safety of second-line anti-hypertensive agents will be implemented in parallel with this proof-of-principle to further elucidate the capabilities the network should have.

The specifications for the proof-of-principle were developed in collaboration with our project partners, with substantial input from AHRQ. Designing and implementing the proof-of-principle within the timeframe of the study required several trade-offs, including the use of synthetic data (i.e., data with artificial information) to facilitate study completion and selecting targeted system and analytic functionality for inclusion. The proof-of-principle design focuses on the key novel elements of the proposed network and does not attempt to illustrate the design features (e.g., redundant security, fine-grained permissions, secure messaging, and authentication) that are well-established within existing software and information technology systems.

The key features for the prototype are:

- **Distributed architecture.** Data remain at the data owner's site, under the data owner's control. Analysis code must be distributed to the data owner, be executed under the data owner's control, and results returned only with the data owner's approval.
- **Strong local control of data uses.** Data owners must be able to control access to, and uses of, the data they hold, and to have access to audit trails/logs of all uses of their data.
- **Federated querying.** A single network portal will be used to develop and distribute queries, aggregate and distribute results, and maintain centralized logs of network usage.

An additional feature included in the proof-of-principle specification was the specific demonstration of the ability to distribute SAS code to remote data nodes, execute the SAS code behind the firewalls of the remote nodes, and return and aggregate the results.

Once the specifications of the proof-of-principle were clear, an appropriate partner to assist with implementation of the prototype specifications was identified. The prototype will be built in collaboration with our partners at the CDC's National Center for Public Health Informatics (NCPHI). The NCPHI grid research team will be using the Globus® Toolkit to implement the prototype. The Globus® Toolkit is an open source middleware that has been used by CDC and others to implement distributed research architectures and federated queries.[31] However, the Globus® Toolkit has not been used in the manner required by the design specifications, making this a true proof-of-principle activity in which all participants are working to develop a novel solution to some of the key challenges of a DRN, namely, the distribution and execution of a query written using a proprietary statistical software package and the heterogeneity of the local software and hardware environments within our network partners.

Four demonstration sites will be guided through a Globus® grid node installation process using the Virtual Data Toolkit (VDT). The NCPHI team will lead this process.

The specifications of the proof-of-principle are described below.

9.1. Overall Research Capability Requirements

The network and system architecture will be designed to have the capability to perform the following general tasks through an internet-based distributed network that accesses multiple data sources secured inside internet firewalls and that are controlled by the individual data owners:

- Identify groups of individuals who satisfy specific combinations of clinical, demographic, or enrollment criteria via a menu-driven query interface and return aggregated counts to the requestor
- Track and store each query for later use, and generate an audit trail
- Submit ad-hoc analysis (e.g., SAS) programs created based on the networks' common data model that will run against patient-level project datasets
- Initially, all project datasets will use a common data model.

SAS has been designated as the initial analysis software of choice because most data owners have already invested substantially in SAS software, and the majority of stakeholder analysts are experienced SAS programmers. Other analysis software may be appropriate as the network evolves, and requiring that the research data be stored in a SAS dataset is not proposed, as SAS can access many different types of data stores (e.g., Oracle, DB2, MySQL, and Sybase, as well as text-format files).

It is expected that the network infrastructure will support a range of query types (e.g., menu-driven feasibility queries, ad hoc feasibility queries, ad hoc analysis queries) and that network users will have differing levels of permissions for each query type. For example, a small set of users may have access to all query functions, whereas a larger set of users may only have access to a limited view of the menu-driven query capabilities. It is also expected that data owners will be able to permit any type of query on their own data by their own employees or designees.

9.2. Federated Query Proof-of-Principle Assumptions

The four organizations participating in this prototype evaluation will be provided a synthetic dataset (i.e., data with artificial information); the proof-of principle will run against these synthetic databases. The main drawback to using synthetic data relates to the inability to fully assess how institutional IRBs will approach and assess implementation of a distributed network. The primary and secondary objectives of the proof-of-principle are described below.

The primary proof-of-principle objective is to distribute a SAS program to data owners and to send aggregated results to an authorized user. Specifically, the use-case will describe the following steps:

- An authorized user authenticates to a central portal
- A SAS program is distributed to each data owner (node); the data owner allows or denies the request for the program to run
- The SAS program is executed at each node, and a standard results set is returned
- The results are aggregated and made available to the authorized user
- A log of site activity for each node is generated

The secondary proof-of-principle objective is a demonstration of security, authentication, and authorization functions at the network management and site management level. These may include:

- Authentication of users (user credentials are verified at a given site – e.g., a user identifier and a password or a user identifier, a passphrase and an RSA secureID key)
- Authorization management (roles are created that allow certain access permissions, such as being authorized to read specific datasets, or perform specific types of query, at a given site)
- Role management (roles serve to bridge authorization and authentication; roles are assigned to authenticated users at a given site, so user access is controlled according to the authorizations associated with the roles they have been assigned)
- Activity logs and monitoring functions
- Central and local implementation of applications to support governance through authentication and roles

9.3. Portal

As described above, the DRN design includes a portal that will contain all of the network management functions, including authentication of users, authorization of user functions, query creation and/or submission, auditing and monitoring functions, data availability synchronization, query aggregation, and communications. For this proof-of-principle, the primary capabilities for the research portal will be to illustrate a user logging into the network, submitting a SAS program to each node, and aggregating the results for the user.

9.4. Databases

The source data will be based on standard file structures as provided by the Informatics team to the participating sites and NCPHI. The proof-of-principle will run against these data structures. The storage format of the databases (e.g., SAS, mySQL, DB2, Oracle, text) is not specified; however, if the underlying data are not stored as SAS datasets, an appropriate SAS/Access interface will be required.

9.5. Network Management

Network management within the DRN includes a broad array of functionality, including authentication, security, authorization, query synchronization, aggregation, auditing, and scheduling. The proof-of-principle architecture should allow queries to be distributed as messages sent to each node or through a “publish and subscribe” architecture in which queries are published and sites retrieve queries to which they have subscribed.

This proof-of-principle focuses mainly on the ability of the network to distribute a SAS program and receive results for aggregation. Therefore, only a few aspects of network management need to be addressed. These include a simple username / password log-in, query distribution, result aggregation, and auditing.

9.6. Local Management

As with network management, local control of data access is an important aspect of the DRN design. However, for this proof-of-principle, only high-level local management functionality is required. This functionality should include the ability to allow access to specific users for the use-case (i.e., the node should be able to allow access only to the user in the use case) and return program results.

9.7. Query Interface

This proof-of-principle does not require development and implementation of a menu-driven query interface. It is assumed that, when developed and tested, a menu-driven query interface would generate a parameterized SAS program (or other) that could then be distributed in the same way as the ad hoc program distribution.

10. Prototype Evaluation Plan: Metrics for Evaluating the Prototype

An overall evaluation of the network proof-of-principle, as described above, will be conducted to describe lessons learned and propose guidelines for evaluating future network activities and growth. By describing the strengths and weaknesses identified during the process, AHRQ and potential collaborators can better understand the utility of the DRN with respect to comparative effectiveness research and the key challenges to implementation.

Although the final network design is not yet known, it is likely that the proposed network will be complex, so multiple metrics will be required to comprehensively evaluate it. The ultimate test of any information system is how well it performs the tasks it was designed to support; therefore, a proof-of-principle evaluation will be implemented in four participating sites to assess: (1) the ability of the system architecture to execute a test query; (2) the challenges encountered in trying to install and operate the necessary node software at each participating site;

and (3) the challenges in software development faced by the Informatics team and NCPHI developers. In evaluating the text query, we will determine the accuracy of the results as compared to results generated using standard methods.

During the pilot query process the system's availability and its ability to recover from system failures will be assessed. The system's usability and usefulness will be evaluated using the System Usability Scale.[32]

11. References

1. Olsen L. IOM Roundtable on Evidence-Based Medicine. The Learning Healthcare System: Workshop Summary. Washington, DC; 2007.
2. Medicare Prescription Drug, Improvement, and Modernization Act (MMA) of 2003. Available at: http://www.cms.hhs.gov/MMAUpdate/. Accessed June 2008.
3. The Sentinel Initiative: A National Strategy for Monitoring Medical Product Safety. Available at: http://www.fda.gov/oc/initiatives/advance/reports/report0508.html, http://www.connectingforhealth.org. Accessed August 2008.
4. 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. Pharmacoepidemiology and Drug Safety 2008 Dec;17(12):1137-41.
5. PubMed. Available at: http://www.ncbi.nlm.nih.gov/sites/entrez/. Accessed 2008.
6. Peterson KA, Fontaine P, Speedie S. The Electronic Primary Care Research Network (ePCRN): a new era in practice-based research. J Am Board Fam Med 2006;19(1):93-7.
7. Peterson KA. The Electronic Primary Care Research Network (ePCRN). In: Clinical Research Networks: Building the Foundation for Health Care Transformation; 2008; Bethesda, MD; 2008.
8. Electronic Primary Care Research Network (ePCRN). Available at: http://www.epcrn.bham.ac.uk. Accessed June 2008.
9. Cancer Biomedical Informatics Grid. Available at: https://cabig.nci.nih.gov/. Accessed May 2008.
10. Harvard Clinical and Translational Science Center. Available at: http://catalyst.harvard.edu/shrine/. Accessed July 2008.
11. Clinical and Translational Science Awards. Available at: http://www.ncrr.nih.gov/clinical_research_resources/clinical_and_translational_science_awards/. Accessed August 2008.
12. Cancer Diagnosis Program of the National Cancer Institute. Available at: http://www.cancerdiagnosis.nci.nih.gov/spin/. Accessed August 2008.
13. i2b2 Informatics for Integrating Biology and the Bedside. Available at: https://www.i2b2.org/. Accessed July 2008.
14. BIG Health Consortium. Available at: http://www.bighealthconsortium.org. Accessed July 2008.
15. National Cancer Institute. The caBIG™ Pilot Phase Report: 2003-2007; 2007 November. Report No.: 07-6244.
16. Vaccine Safety Datalink. Available at: http://www.cdc.gov/vaccinesafety/vsd/. Accessed July 2008.
17. Institute of Medicine (U.S.). Committee on the Review of the National Immunization Program's Research Procedures and Data Sharing Program. Vaccine safety research, data access, and public trust. Washington, D.C.: National Academies Press; 2005.
18. Biomedical Informatics Research Network. Available at: http://www.nbirn.net/. Accessed August 2008.

19. FDA's Sentinel Initiative. Available at: http://www.fda.gov/oc/initiatives/advance/sentinel/. Accessed August 2008.

20. Food and Drug Administration. The Sentinel Initiative: National Strategy for Monitoring Medical Product Safety: Office of Critical Path Programs; 2008.

21. National Effort to Measure and Report on Quality and Cost-Effectiveness of Health Care Unveiled. In: Robert Wood Johnson Foundation; 2007.

22. MediGRID. Available at: http://www.d-grid.de/index.php?id=42&L=1. Accessed August 2008.

23. Data Extraction and Longitudinal Time Analysis System. Available at: http://www.dsg.harvard.edu/index.php/Main/ProjectSafety. Accessed December 2008.

24. Observational Medical Outcomes Pilot Available at: http://www.fnih.org/index.php?option=com_content&task=view&id=583&Itemid=730. Accessed December 2008.

25. eHealth Initiative. Available at: http://www.ehealthinitiative.org/. Accessed December 2008.

26. JLConnell LS. Structured Rapid Prototyping: An Evolutionary Approach to Software Development. New York: Yourdon Press; 1989.

27. Consortium. SP. Evolutionary Rapid Development. SPC document SPC-97057-CMC, version 01.00.04. Herndon, Va; June 1997.

28. HMO Research Network. Available at: http://www.hmoreasearchnetwork.org. Accessed: July 2008.

29. Slavkovic A, Nardi Y, Tibbits M. "Secure" Logistic Regression for Horizontally and Vertically Partitioned Data. [PUBLISHER? CITY?] 2007.

30. Karr A, Lin X, Sanil AP, et al. Secure regression on distributed databases. Journal of Computational and Graphical Statistics 2005;14(2):263-79.

31. The Globus® Toolkit. Available at: http://www.globus.org/toolkit/. Accessed: August 2008.

32. Brooke J. SUS: A "Quick and Dirty" Usability Scale. Usability Evaluation in Industry. London: Taylor and Francis; 1996.

33. Ozsu MTV. Principles of Distributed Database Systems. 2nd ed. Upper Saddle River, NJ: Prentice Hall; 1999.

34. Sheth AP, Larson JA. Federated database systems for managing distributed, heterogeneous, and autonomous databases. ACM Comput Surv 1990;22(3):183-236.

35. Controlled vocabulary. Wikipedia, The Free Encyclopedia. Available at: http://en.wikipedia.org/w/index.php?title=Controlled_vocabulary&oldid=251807628. Accessed September 2008.

36. Are you a covered entity? Available at: http://www.cms.hhs.gov/HIPAAGenInfo/06_AreYouaCoveredEntity.asp. Accessed August 2008.

37. Cyberinfrastructure. Wikipedia, The Free Encyclopedia. Available at: http://en.wikipedia.org/w/index.php?title=Cyberinfrastructure&oldid=253414152. Accessed August 2008.

38. DMZ (computing). Wikipedia, The Free Encyclopedia. Available at: http://en.wikipedia.org/w/index.php?title=DMZ_(computing)&oldid=255485715. Accessed August 2008.

39. Code of Federal Regulations, Title 45, Public Welfare. Department of Health and Human Servies. Available at: http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.htm#46.101. Accessed August 2008.

40. Federated search. Wikipedia, The Free Encyclopedia. Available at: http://en.wikipedia.org/w/index.php?title=Federated_search&oldid=256543682. Accessed August 2008.

41. Medical Privacy - National Standards to Protect the Privacy of Personal Health Information. Department of Health and Human Services. Available at: http://www.hhs.gov/ocr/hipaa/. Accessed August 2008.

42. Koutrika G. Heterogeneity in digital libraries: Two sides of the same coin. In: DELOS Newsletter; 2005.

43. Horizontal scalability. SearchCIO.com. Available at: http://searchcio.techtarget.com/sDefinition/0,,sid182_gci929011,00.html#. Accessed August 2008.

44. Privacy/Data Protection Project. University of Miami, Miller School of Medicine. Available at: http://privacy.med.miami.edu/glossary/xd_limited_data_set.htm. Accessed August 2008.

45. Middleware. Wikipedia, The Free Encyclopedia. Available at: http://en.wikipedia.org/w/index.php?title=Middleware&oldid=257221860. Accessed August 2008.

46. Ontology. Available: http://tomgruber.org/writing/ontology-definition-2007.htm. Accessed September 2008.

47. Protected health information. Wikipedia, The Free Encyclopedia. Available at: http://en.wikipedia.org/w/index.php?title=Protected_health_information&oldid=217204208. Accessed August 2008.

48. Publish/subscribe. Wikipedia, The Free Encyclopedia. Available at: http://en.wikipedia.org/w/index.php?title=Publish/subscribe&oldid=258319255. Accessed August 2008.

49. Bondi AB. Characteristics of scalability and their impact on performance. In: Proceedings of the 2nd international workshop on software and performance. Ottawa, Ontario, Canada: ACM; 2000.

50. Aronsky D. Factors affecting the sustainability of information technology applications in health care. AMIA Ann Symp Proc 2003.

51. Vertical scalability. SearchCIO.com. Available at: http://searchcio.techtarget.com/sDefinition/0,,sid182_gci928995,00.html. Accessed August 2008.

52. Karasavvas KA, Baldock R, Burger A. Bioinformatics integration and agent technology. J Biomed Inform 2004;37(3):205-19.

53. Defense Information Services Agency. Horizontal Fusion Developers Reference Document: Department of Defense Assistant Secretary of Defense for Networks and Information Integration/DoD CIO; 2004.

12. Resources

Web Sites of Interest:

The Agency for Health Care Research and Quality (AHRQ): http://effectivehealthcare.ahrq.gov/

BIG Health Consortium™: http://www.bighealthconsortium.org

BIRN: http://www.nbirn.net/

Cancer Biomedical Informatics Grid (caBIG™): https://cabig.nci.nih.gov/

ePCRN: http://www.epcrn.bham.ac.uk/

i2b2: https://www.i2b2.org

Globus®: http://www.globus.org/

HMORN: www.hmoresearchnetwork.org

MediGRID: http://www.d-grid.de/index.php?id=42&L=1

Medicare Prescription Drug Improvement and Modernization Act of 2003: http://www.cms.hhs.gov/MMAUpdate/

On December 8, 2003, President Bush signed into law the Medicare Prescription Drug Improvement and Modernization Act (MMA) of 2003 (Pub. L. 108-173). This legislation provides seniors and individuals with disabilities with a prescription drug benefit, more choices, and better benefits under Medicare.

As part of the Act, a new Medicare Part D was created, providing access to prescription drug insurance coverage to individuals who are entitled to Part A or enrolled in Part B. Participation in Part D is voluntary and requires an affirmative election to join. Coverage began January 1, 2006.

SPIN: http://www.cancerdiagnosis.nci.nih.gov/spin/

Vaccine Safety Datalink: http://www.cdc.gov/vaccinesafety/vsd/

13. Appendixes

Appendix A: Glossary of Terms[a]

Administrative scalability—the capacity for substantial numbers of new organizations or enterprises to address all necessary administrative challenges in joining the network.

Authentication—the process by which an end-user provides evidence of their identity and is confirmed as an approved user by an information system. Authentication is typically performed in networked environments (in which users access systems via a network connection). Examples of authentication include user identifier and password combinations, use of hardware based secure token generators (e.g., RSA SecureID), and use of biometrics (e.g., retinal scan or fingerprints).

Authorization—the process by which a user is permitted to access a specific system resource, through a defined set of relationships between a user's identity and a set of specific permissions assigned to him through explicit policies (e.g., Person A is permitted to perform action B on resource C). Authorization is only permitted after a user has been authenticated (*see Authentication*).

Autonomy[33,34]—a property of distributed database systems that refers to the degree of local control maintained over all local resources.

Client-server (hub-and-spoke) network—a network in which all nodes are connected to a central server and do not necessarily know of the existence of any other nodes. As compared to a peer-to-peer network in which all network nodes are connected to each other, and each node functions as both a client and a server.

Controlled vocabulary[35]—a carefully selected list of words and phrases, which are used to tag units of information (document or work) so that they may be more easily retrieved by a search. A controlled vocabulary is often referred to as an *authority list*. Each concept is described using only one authorized term and each authorized term in the controlled vocabulary describes only one concept.

Covered entities[36]—under HIPAA, covered entities include health plans, healthcare clearinghouses (public or private entities including a billing service, repricing company, community health management or community healthcare delivery organizations, and "value-added" networks and switches, that does either of the following functions: a) processes or facilitates the processing of health information received from another entity in a nonstandard format or containing nonstandard data content into standard data elements or a standard transaction; b) receives a standard transaction from another entity and processes or facilitates the processing of health information into nonstandard format or nonstandard data content for the receiving entity) and healthcare providers who transmit any health information in electronic form.

Cyberinfrastructure[37]—describes the new research environments that support advanced data acquisition, data storage, data management, data integration, data mining, data visualization and other computing and information processing services over the Internet.

[a] Some of the definitions above were written by the authors; others were derived from other sources, as indicated.

Defense-in-depth strategy—a comprehensive approach to network security that uses multiple layers of defense to protect the network.

Demilitarized zones (DMZs)[38]—in computing, a DMZ is a physical or logical sub-system that contains and exposes an organization's external services to another network that usually sits between an organization's firewall and another firewall that separates the DMZ from the Internet. A DMZ is sometimes referred to as a "perimeter network."

Distribution[33]—a property of distributed database systems that refers to how data is situated within the system.

Exempt Research[39]—research using data from living persons that does not require IRB approval when the research either does not involve human subjects as defined in the Code of Federal Regulations (45CFR46) Protection of Human Subjects Subpart A, or the only involvement of human subjects is in one of the six "exempt" categories listed in the Code. Exemption category 4 is most pertinent to the DRN: research involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects.

Expedited IRB Review—a procedure that may be used to review either or both of the following: (1) some or all of the research appearing on the list [of potential research activities promulgated by the Office of Human Research Protections] and found by the reviewer(s) to involve no more than minimal risk, (2) minor changes in previously approved research during the period (of one year or less) for which approval is authorized. Under an expedited review procedure, the review may be carried out by the IRB chairperson or by one or more experienced reviewers designated by the chairperson from among members of the IRB. In reviewing the research, the reviewers may exercise all of the authorities of the IRB except that the reviewers may not disapprove the research.

Extensibility—the ability of a network to allow for expansion and growth, including more analytic and system capabilities, incorporation of additional data types and sources, and inclusion of new data and software standards.

Federated Network—a collection of institutional networks, each of which functions independently, but which collectively form a larger, cooperative, interoperable networked environment through technical and administrative arrangements that allow a user from one network to gain access to specific resources on other networks. From the user's perspective, the Federated network functions as if it were a single, integrated network. (See *Virtual Organization.*)

Federated Query[40]—consists of (1) broadcasting a standardized query to a group of distributed databases with the appropriate syntax, (2) merging the results collected from the databases, and (3) presenting them in a succinct and unified format with minimal duplication. Examples of federated queries include Expedia.com searches for flights and Amazon.com searches for books or other material.

Health Insurance Portability and Accountability Act (HIPAA)[41]—a statute enacted in 1996 that requires the U.S. Department of Health and Human Services to set rules that apply to "covered entities" regarding transactions involving personal healthcare information.

Heterogeneity[42]—a property of distributed database systems that refers to any dissimilarities in the way that data are represented, maintained, or manipulated in a database.

Heterogeneity transparency—the ability of a distributed database system to "mask" the heterogeneity of the component systems from the end-user, such that the components appear to be similar.

Horizontal scalability[43]—the ability of information systems to be extensible by adding new hardware or software components and configuring them to act as a single logical system. Compare with vertical scalability, which is the addition of capacity to existing systems without considering their role in acting as a larger, single unit.

Information system—a system composed of computer hardware and software, people, and processes that is designed, implemented, and maintained for the purpose of processing data and information in an organization.

Institutional Review Board (IRB)[39]—a committee, the constituency of which includes scientists, non-scientists, and community members, as mandated by Federal law, that is charged with approving and monitoring research involving human subjects.

Limited dataset[44]—a limited dataset contains PHI, and is available for disclosure from one covered entity to another, without approval of individuals whose PHI is being disclosed. Limited datasets exclude the following direct identifiers of the individual or of relatives, employers, or household members of the individual: names; postal address information other than town or city, state, and zip code; telephone numbers; fax numbers; electronic mail addresses; social security numbers; medical record numbers; health plan beneficiary numbers; account numbers; certificate/license numbers; vehicle identifiers and serial numbers, including license plate numbers; device identifiers and serial numbers; web Universal Resource Locations (URLS); Internet Protocol (IP) address numbers; biometric identifiers, including finger and voice prints; and full face photographic images and any comparable images.

Load scalability—the ability of information systems to be extensible by adding an increased volume of operations to their system.

Location transparency—the ability of a distributed database system to "mask" the location of the data from the end-user.

Metadata—data used to describe data, intended to specify the understanding of how the data is used and managed. Metadata can be descriptive (e.g., data dictionaries), administrative (organizational charts), or structural (data models).

Middleware[45]—is computer software that connects software components or applications. The software consists of a set of enabling services that allow multiple processes running on one or more machines to interact across a network. This technology evolved to provide for interoperability in support of the move to coherent distributed architectures, which are used most often to support and simplify complex, distributed applications. It includes web servers, application servers, and similar tools that support application development and delivery.

Ontology[46]—defines a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or sets), attributes (or properties), and relationships (or relations among class members).

Operational autonomy—a property of distributed database systems that refers to the degree of local control maintained over the technical design aspects of components resources.

Parsimony—a property of distributed database systems that refers to the minimal transfer of data necessary to perform a given operation.

Peer-to-peer distributed research network—a computer network in which all network nodes are connected to each other network node and each node functions as both a client and a server. This is opposed to a client-server network (or hub-and-spoke) in which all nodes are connected to a central server and do not necessary know of the existence of any other nodes.

Protected health information (PHI)[47]—any information about health status, provision of health care, or payment for health care that can be linked to an identifiable individual.

Publish and subscribe[48]—is an asynchronous messaging paradigm where senders (publishers) of messages are not programmed to send their messages to specific receivers (subscribers). Rather, published messages are characterized into classes, without knowledge of what (if any) subscribers there may be. Subscribers express interest in one or more classes, and only receive messages that are of interest.

Scalability[49]—the ability of technical systems to be extensible either in their functions or their capacity.

Semantic heterogeneity—occurs when a data element may represent one of two or more distinct concepts (homonymy). This typically arises when data (such as distance) are combined from sources in which it has been recorded using different scales (such as meters and feet).

Service autonomy (Association autonomy)—a property of distributed database systems that refers to the degree of local control maintained over what services are provided to the distributed database system.

Structural heterogeneity—a type of heterogeneity that occurs due to any differences in database structures or schemas.

Sustainability[50]—in an information systems context, it refers to the ability of a project or program to continue to perform certain functions without major interruptions over time.

Syntactic heterogeneity—occurs when a concept is represented by two or more data element names (synonymy) in two or more independent systems. For example, "M", "Male" and "1" all mean male sex.

System heterogeneity—a type of heterogeneity that arises at the level of choice of operating systems, programming languages, authentication systems, data formats, choice of hardware and vendor, and other low level technical aspects of complex, independent computing systems that strive to provide interoperability to users.

Transparency—(1) the ability of a distributed database system to "mask" its operations from end-users (2) the openness and accountability of the system to all participants including end-users, employees and users of the services provided such as patients.

Vertical scalability[51]—the ability of information systems to be extensible by adding new capacity through the addition of hardware or software resources to existing configurations. Compare with *horizontal scalability,* which refers to increasing capacity by making systems appear to work together as one logical unit.

Virtual organization—a collection of independent institutions that agree to organize themselves into a larger, coherent administrative and technical structure, that permits the controlled sharing of resources on each of the independent networks with users on other networks.

Appendix B: Distributed Research Model Features

At the most basic level, a distributed research model is defined as a system in which data are physically held and managed by each data owner (e.g., HIPAA covered entity), that can accept federated queries distributed through network software, run the queries against the local data, and return aggregated results to the end-user. The specific system architecture of a distributed model and the required functionality will depend on the needs and demands of the network stakeholders. The following is a description of some general features and characteristics of distributed networks; the relative importance of these features and characteristics to the design of a distributed research network for public health can only be determined by the network stakeholders.

Network Principles

a. *Scalability*.[49] Scalability is generally defined as the ability of technical systems to be extensible either in their functions or their capacity. Distributed database systems are scalable if performance is not compromised when either: (1) new nodes (i.e., sites) are added (i.e., horizontal scalability) to the system; (2) new hardware or database resources are added to a single node (i.e., vertical scalability); or (3) concurrent users (i.e., load scalability) are added to the system. The first two dimensions of scalability address increases in size or capacity, whereas the latter dimension addresses increases in the volume of operations. Horizontal scalability is further distinguished by whether the new nodes that are added represent new organizations (i.e., administrative scalability) or simply new users from organizations that are already integrated into the system. Administrative scalability is more challenging. Additionally, the vertical scalability issues involved in adding new database resources can be subdivided into whether the new resources are of a new unique form (e.g., image data or datasets using non-standardized schema) or just an increased amount of data in a well-understood form (e.g., standardized datasets).

Scalability is a crucial aspect of the architecture of a research network 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 owners and other stakeholders. Certain dimensions of scalability may be more highly prioritized within the context of a research network focused on public health activities because such a network may prioritize facilitation of the addition of new organizations. Therefore, highly complicated, expensive, or rigid designs relating to hardware or other technical infrastructure may discourage participation. Adding new organizations to a distributed research network can expand both the data and intellectual resources on which the network draws to conduct effectiveness, comparative effectiveness, health care utilization and outcome studies. Accommodating load scalability within the network architecture may be a lower near-term priority because the expected volume of concurrent users is low.

In the context of a research network using healthcare data, scalability also encompasses the ability to incorporate new clinical concepts and changes in standard coding schema. For example, a scalable healthcare data network should be able to include new diagnosis or medication coding schemes as well as entirely new data areas.

b. *Transparency*. Transparency is a concept with two understood meanings within database systems, both of which are relevant to a distributed network. One definition of transparency relates to the underlying technical processes of distributed database systems (i.e., how does it work). In this context the term "transparency" can be thought of as "invisible," as in the details of the network are transparent or invisible to the user. Database systems that are transparent have the details of the system hidden from the user in such a way that the user does not need to know system details in order to effectively use the system. The other definition relates to governance and deals with issues of openness and accountability in how the system is used and managed by the stakeholders, that is, all the rules of use are known to all participants.

i. Transparency as a Distributed Database Systems Concept.[33] In a distributed database system or network, transparency refers to the degree to which the complex operations of the system are "hidden" or masked from the users. That is, the end user is blinded to the mechanics of how and where the data are stored as well as how the data are queried. To the layman, transparency might be better understood as usability because high levels of transparency make the system easy and efficient for non-expert end-users. For example, an end-user may send a query requesting the number of patients with a diagnosis of hypertension who were also dispensed at least one anti-hypertensive medication. In a transparent network, the end-user need not know the location of the data, how the data are stored, how the query is distributed, or even how the network defines the relevant clinical concepts.

ii. Transparency as a Governance Concept. In the language of governance, transparency refers to the degree to which the appropriate stakeholders, such as data owners, can understand the nature of operations performed using networked resources; that is, for data owners to know exactly how their data are being used and by whom. This includes functions like complete audit trails and use logs that would allow any data owner immediate access to information regarding the uses of their data, and therefore the ability to protect against unwanted use patterns.

While transparency as a distributed systems concept provides a usability benefit that is desirable and convenient, it is not a necessary design feature of an operational system. On the other hand, transparency in the governance aspect is a top priority in the architecture of a distributed research network because such accountability and traceability engenders participation in the network and facilitates trust-building activities among unfamiliar organizations. Technical interlocks (e.g., alerts or lockdowns when certain data is accessed) allow sites to maintain control over the use of proprietary and/or confidential data as well as to abide by any unique organizational rules or state laws. Further, the knowledge that all transactions are recorded will likely influence the behaviors of end-users and may help to deter any inappropriate use.

c. *Autonomy*.[33,34] Autonomy is a property of distributed database systems that refers to the degree of local control over a database resource and/or the degree of independence of the local system from the network. It can range from tight integration (i.e., little or no autonomy) to complete autonomy. Distributed databases that are tightly integrated are available to any user on the system without constraints in either access to the data or the ability to perform operations. On the other extreme are fully autonomous component databases that are isolated from other databases in the network. For example, in a fully autonomous network, a database at one site does not know of the existence of, or have the ability to communicate directly with, a similar database at another site. These databases also function independently from the network – that is, they may be removed from the network at any point because they do not require network functions to operate properly. Thus, data owners maintain complete control of the use of their data at all times and are free to remove their data from the network at any time. Autonomy classifications may be further understood by the way in which local control is exercised.

i. Operational Autonomy of the Source Data. Data owners exercise control over the technical design aspects of their databases and the operations on their databases. That is, a site may choose to hold their data in a unique schema and software program, which is only accessible via a unique programming language. This concept is referred to as design autonomy and implies that sites are free to choose their own design with regard to the data being managed, the representation of that data, and the semantic meaning of that data.

ii. Execution Autonomy. This refers to the ability of the component database to execute local operations without interference from external network requests. In other words, the site may dictate the operations that are performed on their databases, including time and subject constraints such as what type of queries may be processed or what time certain commands are permitted. For example, if a local database is updating itself or performing quality checks, it may choose to reject all network requests during that time. Thus, approval of scheduling requests ultimately remains local in databases with execution autonomy even if the scheduling service is performed using network resources.

iii. Service Autonomy. The local database retains the right to decide what services it provides to the network and can decide whether or not to share its resources. The component database can disassociate itself from the network as it chooses.

d. *Heterogeneity*.[42] Heterogeneity refers to any dissimilarities in the way that data are held in a database. Heterogeneity may be classified into four categories:

i. System heterogeneity. Refers to any difference in hardware (e.g., servers, network connections) and/or operating systems.

ii. Syntactic heterogeneity. Refers to differences in access protocols (how to query the data) and programming languages.

iii. Structural heterogeneity. Refers to databases with different structures or schemas.

iv. Semantic heterogeneity. Semantic heterogeneity occurs when databases refer to the same data element using different names (i.e., synonyms) and when databases use the same name to refer to two distinct concepts (i.e., homonyms).

System and syntactic heterogeneity together (often combined under the heading of technical heterogeneity) is the least challenging aspect of heterogeneity to resolve.[52] Commonly-used web services architectures, common querying languages, and accompanying web technologies can provide interoperability to address these heterogeneity issues. The need to perform complex statistical operations (e.g., linear regression) within a distributed research network presents a challenge as these statistical operations are not easily translated into commonly used web-based querying languages.

Substantial network design challenges arise when: (1) the underlying structure of databases are different (structural heterogeneity), for example, when a record (row) in one system is unique to a patient, a date, and a diagnosis, and a record in a different system is unique to a patient and a date; and (2) the terminology used in different systems are not consistent and/or can mean different things (semantic heterogeneity), for example, sex and gender may be used in different systems to refer to the same concept (biological patient sex, male or female) or two different concepts (biological sex and sociological gender). Database schema mapping to a common data model, query translation, metadata (data that describes data), and controlled ontologies or vocabularies (i.e., approved naming conventions for database concepts), and transformation to a common data model are typical solutions to this type of heterogeneity problem.

e. *Distribution*.[33] While autonomy refers to the level of control over one's data, distribution refers to actual location of data and how the network databases relate to each other. Distribution options can be described as fully distributed (e.g., peer-to-peer networks in which all data resources are connected to each other resource), partially distributed (e.g., client-server networks or a hub-and-spoke design), or centralized. A distributed approach for a research network using healthcare data is preferred for several reasons. First, many data owners simply would not participate in the network if they were required to centralize their sensitive, protected, and proprietary data in a central warehouse. Second, centralization creates a single point of failure because any single attack on the system could disable or compromise all the data in the entire network. Third, requests or queries run less efficiently on large centralized databases.

f. *Security*.[53] Security is a top priority in the design of a research network using healthcare data. The unique legal and regulatory aspects of using PHI necessitate designing an extremely secure system. Data owners are particularly concerned about any potential unplanned disclosures of data that may violate HIPAA, or with the unintended traceability of de-identified data to an individual. Like many critical information systems, a research network using healthcare data will require a defense-in-depth strategy designed to protect against vulnerabilities created by personnel, operations, and technology. A defense-in-depth strategy – defined as a comprehensive approach to security that uses multiple layers of defense – is designed to thwart inappropriate use or attacks by providing warning signals that will close access or alert site administrators to potentially suspicious use. Security layers may include any of the following: physical security (i.e., the use of locks on servers that house data or network functions); the use of biometrics (e.g., retinal scans or fingerprints) for authentication; the use of passwords or hardware devices for authentication; the use of public key infrastructure or security certificates for authorization; the use of demilitarized zones (DMZs), firewalls, virtual private networks (VPNs) or proxy servers to provide perimeter security; the use of encryption technology and secure sockets layer (SSL) for message security; the use of HTTPS or IPsec for network security; the use of auditing and logs for accountability; the use of anti-virus software for protection against attacks; and the use of timed-access control systems for fine-grained control.

g. *Sustainability*.[50] Sustainability in an information systems context generally refers to the ability of a project or program to continue to perform without major interruptions over time. Design factors that are measured in an assessment of sustainability should include (1) effectiveness, (2) financial viability, (3) reproducibility, and (4) portability.

i. Effectiveness. Effectiveness refers to the ability to meet certain measureable goals. Designing for sustainability thus requires clear elicitation of goals as well as appropriate metrics to measure achievement of those goals.

ii. Financial Viability. Financial viability is a measure of network cost-effectiveness. That is, does the investment in networked resources provide a benefit that merits continued investment? Thus, designing for sustainability requires a process to measure the costs and benefits of the resource in light of alternative ways to reach the stated goals of the network. Since the network will require continuous funding, it is important to define cost and benefit measurements at the outset, particularly for benefits that do not easily translate into dollars (e.g., the contribution to the field as a result of publication of research studies).

iii. Reproducibility. Reproducibility is a sustainability measure that refers to the ability to extend the network to new settings. It is closely related to the concept of administrative scalability. Reproducibility enhances sustainability because broader use of a resource more firmly entrenches its staying power. Rapid and uncontrolled growth may diminish value by decreasing service quality or impeding achievement of goals. Therefore, an important design feature of the network requires setting reasonable expectations for rates of growth, and designing a system that does not preclude growth.

iv. Portability. Portability refers to the ease of adapting to new technical (e.g., software or hardware) environments or system needs. Portability differs from reproducibility in that the former implies some impetus to adapt the system while the latter implies the ability to replicate the original system in new settings without adaptation. Portability requires designs flexible enough to adapt to emerging needs. Because anticipation of these new needs is difficult, the design should take measures to reduce the chances of technological lock-in that may limit flexibility.

h. Parsimony. Parsimony refers to the minimal transfer of data necessary to perform a given operation. A distributed research network design should enhance the concept of parsimony by using all available technology to ensure that data stays on-site whenever possible.

Design Trade-Offs Among the Principles

It is important to recognize that it is not always possible to accommodate all desired requirements when designing a complex system. Certain design choices have implications within the system, and it is important to consider how the prioritization of one design principle may adversely affect another.

Autonomy and Heterogeneity—By prioritizing and preserving autonomy the challenges of heterogeneity become more complex. For inclusion in the network each data resource must either be translated to the network common data model (an extract-transform, and load process), or the network must develop software mapping approaches to accommodate different data schemas. The work of transforming a new data resource to the common model is non-trivial and requires substantial ongoing effort by the data owner to complete the transformations and ongoing centralized effort and oversight by the network to ensure that the data owner transformations are performed as expected. Allowing heterogeneity places more burden on software development because new mechanisms would be needed to either translate network queries to each unique data source or transform each data source on an ad hoc basis for each query. Transformation of data to the common model is generally preferred; however, the approach to network incorporation is likely influenced by the size of the organization, the amount and uniqueness of available data, the way data are held, and funding considerations. These considerations are also important depending on whether the perspective is the data owner's or the network's. It may be impracticable from a cost and time perspective to insist on full design autonomy; some degree of data standardization will be desirable.

Sustainability and Scalability—Expanding a network (scalability) beyond the initial network participants, especially to organizations with different business needs and data resources, complicates network maintenance and governance and, therefore, its sustainability. Network resources must be divided between maintaining the present state of the network and investing in the future state. Further, as a network grows, sustainability becomes more challenging because "ownership" of the system becomes more widely distributed. To what extent should the network make resources available for new organizations to become part of the network? Further, data owners will not have equal motivation to become part of the network; a data owner with a large number of patients to contribute to the network may not have sufficient incentive to bear the full burden of joining the network because it will derive less benefit than smaller data owners. However, the network users may highly value a large increase in available data and make specific accommodations to ensure network growth.

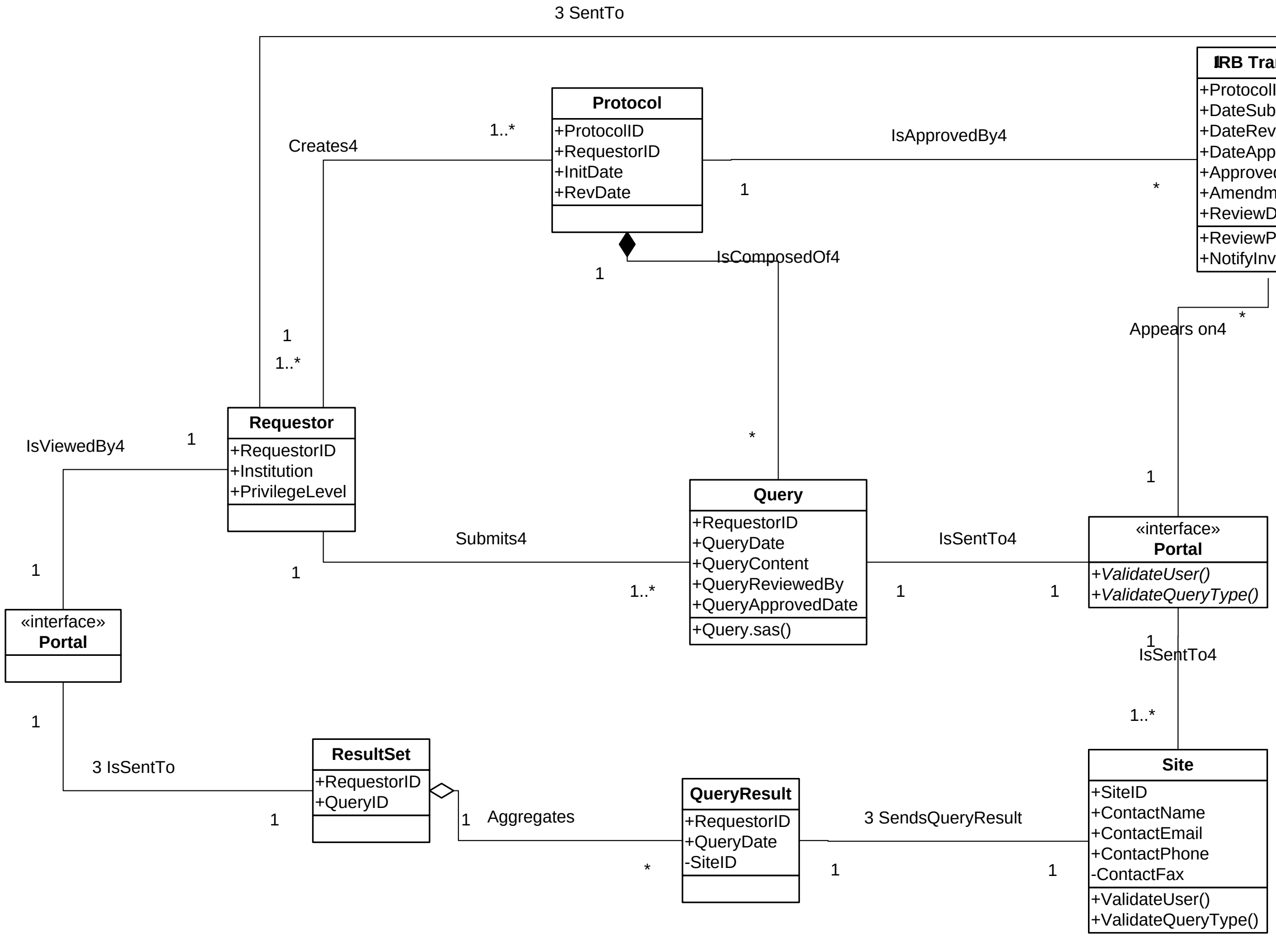
Security and Usability—Strong but onerous security measures may degrade system usability and discourage use. A single point of network access for authentication provides convenience, and it may not be as robust or represent a true defense-in-depth strategy that requires multiple layers of authentication. A balance arises between how many different checks are needed to create a secure system and how difficult the system becomes to use. If the system becomes overly burdensome to access, it may become unsustainable as a result of lack of use.

Association Autonomy and Sustainability—Allowing data owners to withdraw participation at any time and for any reason (i.e., association autonomy) raises challenges to sustainability. Thus, tremendous emphasis in design should be put on making the network accessible, useful, efficient, and inexpensive so as to discourage withdrawal. An incentive to maintain participation used in existing networks is to reimburse data owners for participation whenever their data is accessed.

Appendix C: Unified Modeling Language (UML) Models

The following models include: (1) a high-level class diagram that elucidates the types of objects and relationships between them that need to be considered in the overall DRN architecture; (2) a data model, showing the proposed data stores that will comprise the basic database architecture that underpins the DRN; (3) a data flow diagram that demonstrates how data flow through the network; and (4) state charts that illustrate the various states queries pass through from conception by users to the return of results. Separate state charts are provided for the two main types of queries (user-generated and menu-generated) that are supported by the network. Together, these models represent the overall technical design of the network as a system, and these models are perhaps most useful to system architects who will be responsible for implementation.

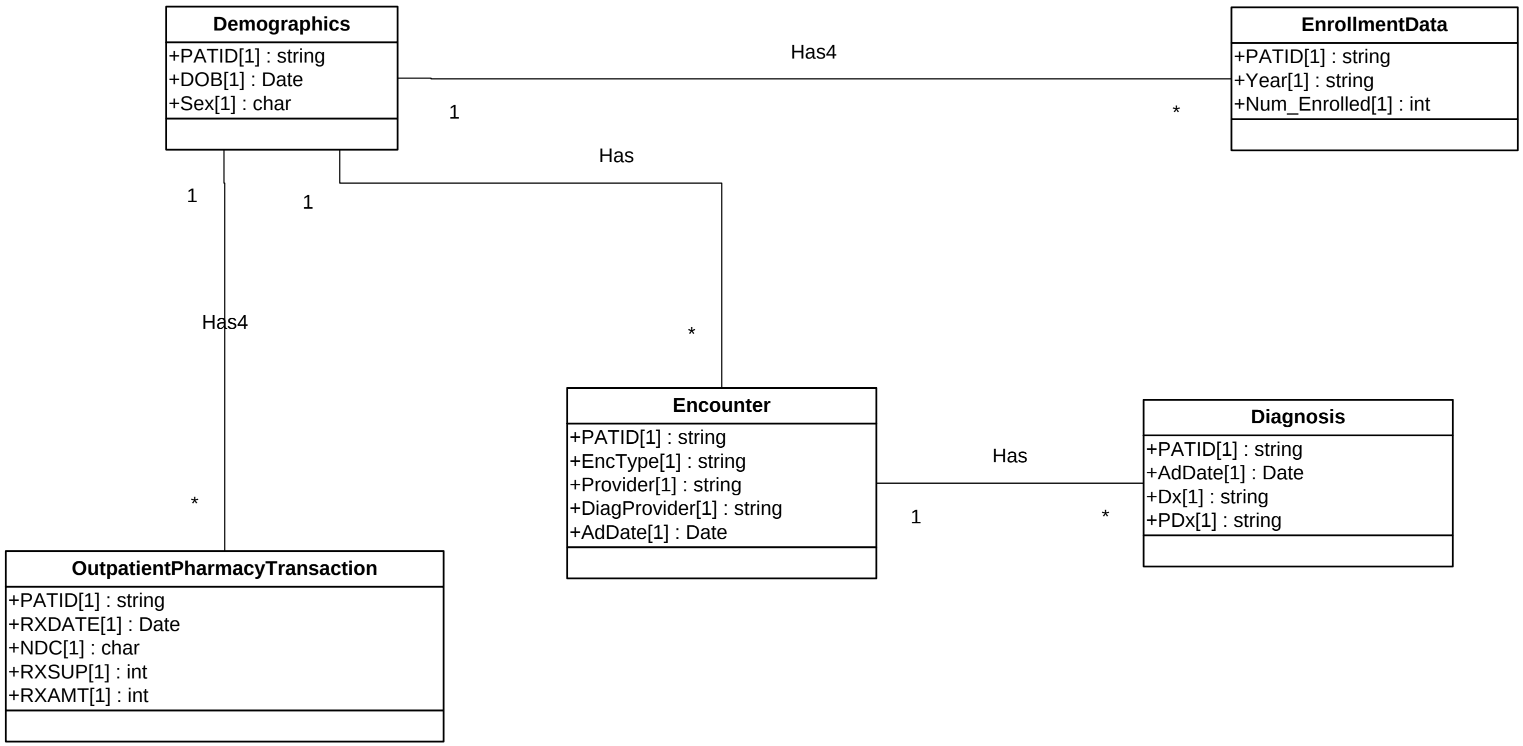
Diagram 1. Class diagram



Note that ValidateUser and ValidateQueryType are polymorphic- they are similar procedures at the portal and at the sites, but the site version differs in that it can override the validation

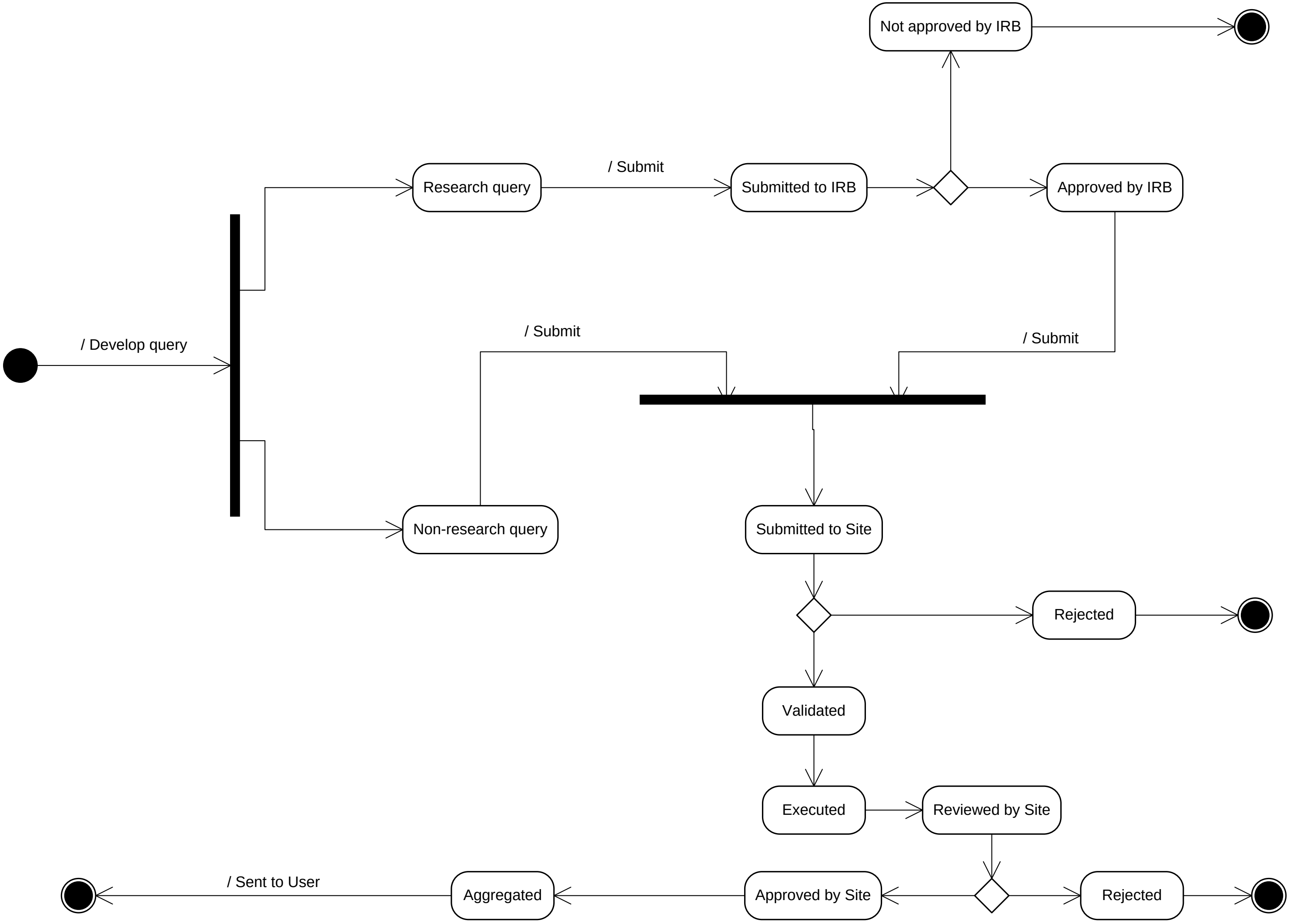
Read diagram from left to right. The boxes represent the abstract classes- these will be instantiated as objects in the system (viz. difference between “a query” and “the query”). Note that logical transitions are not shown here- these are on the State-Transition Diagram. Each class is named (top section), attributes are listed (middle section), and procedures identified (bottom section). Procedures may be software or manual, or combinations. These will be defined in subsequent documents. Lines represent relationships between classes, and are labeled to indicate cardinality (1:1, 1:many, many:many). Open diamonds indicate relationships where one class’ instances are aggregated into another. Closed diamonds indicate relationships where a class instance is composed of multiple instances of another. Note that this model does not distinguish between research and preparatory-to-research activities.

Diagram 2. Data model



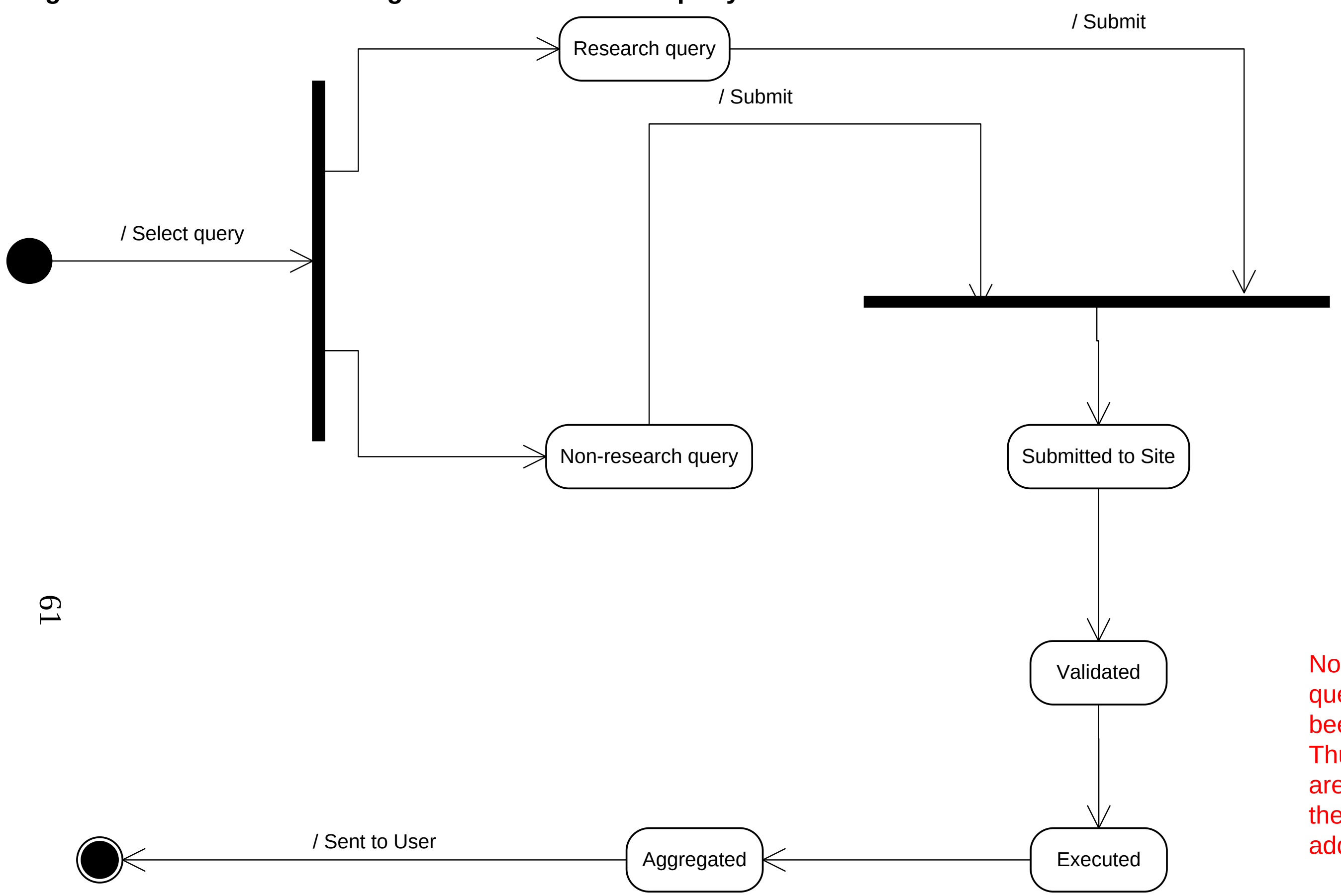
The boxes represent the entities- these will become tables in the database, as specified in the Proof-of-Principle document. Each entity has a name and list of attributes with their type and cardinality. For the prototype, all sites will adhere to this data model, and will ETL into the structure specified here. However, non-standard sites may arise later and these will be dealt with on a case-by-case basis.

Diagram 3. State-transition diagram for user-generated query



Read diagram from left to right, top to bottom, following the arrows. The top circle represents the starting state, the bottom (bulls-eye) the end-state. Each bubble represents a distinct state in which a single query can exist. Thin lines represent transitions from one state to the next. The thick lines are forked transitions (the query could go to one state or another at that point). Diamonds represent decision nodes.

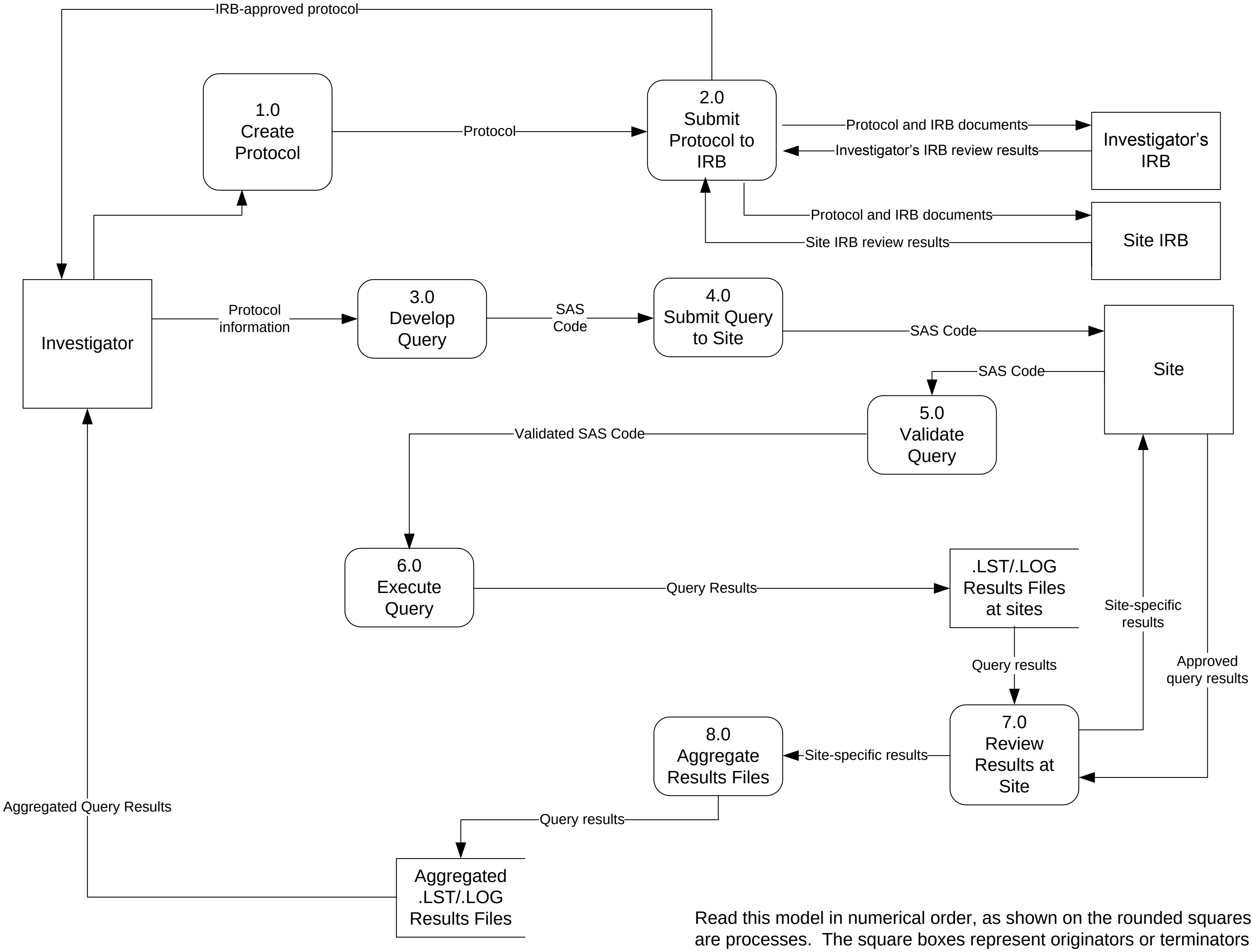
Diagram 4. State-transition diagram for menu-driven query



Note that in a menu-generated query, approvals have already been obtained from sites and IRBs. Thus, execution and aggregation are automatically performed after the query is submitted, without additional review by the sites.

Read diagram from left to right, top to bottom, following the arrows. The top circle represents the starting state, the bottom (bulls-eye) the end-state. Each bubble represents a distinct state in which a single query can exist. Thin lines represent transitions from one state to the next. The thick lines are forked transitions (the query could go to one state or another at that point). Diamonds represent decision nodes.

Diagram 5. Data flow diagram for query



Read this model in numerical order, as shown on the rounded squares- these are processes. The square boxes represent originators or terminators of data. The lines represent data flows and their directionality, and the open-ended boxes are data stores, typically tables in a database. This model does not distinguish between research and preparatory-to-research activities.