eGEMs

Scalable Architecture for Federated Translational Inquiries Network (SAFTINet) Technology Infrastructure for a Distributed Data Network

Lisa M. Schilling, MD, MSPH;[i] Bethany M. Kwan, PhD, MSPH;[i] Charles T. Drolshagen;[i] Patrick W. Hosokawa, MS;[i] Elias Brandt;[ii] Wilson D. Pace, MD, FAAP;[i] Christopher Uhrich;[iii] Michael Kamerick;[iii] Aidan Bunting, MA;[iii] Philip R.O. Payne, PhD;[iv] William E. Stephens;[iv] Joseph M. George;[iv] Mark Vance;[iv] Kelli Giacomini;[i] Jason Braddy;[i] Mika K. Green[i] and Michael G. Kahn, MD, PhD[i]

Abstract

Introduction: Distributed Data Networks (DDNs) offer infrastructure solutions for sharing electronic health data from across disparate data sources to support comparative effectiveness research. Data sharing mechanisms must address technical and governance concerns stemming from network security and data disclosure laws and best practices, such as HIPAA.

Methods: The Scalable Architecture for Federated Translational Inquiries Network (SAFTINet) deploys TRIAD grid technology, a common data model, detailed technical documentation, and custom software for data harmonization to facilitate data sharing in collaboration with stakeholders in the care of safety net populations. Data sharing partners host TRIAD grid nodes containing harmonized clinical data within their internal or hosted network environments. Authorized users can use a central web-based query system to request analytic data sets.

Discussion: SAFTINet DDN infrastructure achieved a number of data sharing objectives, including scalable and sustainable systems for ensuring harmonized data structures and terminologies and secure distributed queries. Initial implementation challenges were resolved through iterative discussions, development and implementation of technical documentation, governance, and technology solutions.

Introduction

Comparative effectiveness research (CER) is broadly defined in the Patient Protection and Affordable Care Act[1] as "research evaluating and comparing health outcomes and the clinical effectiveness, risks, and benefits of two or more medical treatments, services, and other items. . .health care interventions, protocols for treatment, care management, and delivery, procedures, medical devices, diagnostic tools, pharmaceuticals (including drugs and biologicals), integrative health practices, and any other strategies or items being used in the treatment, management, and diagnosis of, or prevention of illness or injury in, individuals." A core benefit of observational CER is the ability to study the care of patients in day-to-day practice, allowing for consideration of the conditions that affect variability in care and health outcomes. Much of the current evidence base for health care depends on the results of randomized trials; however, those trials do not adequately account for the variability experienced in actual practice.[2,3] Other benefits include lower cost, greater generalizability, the ability to study rare events, and the ability to generate faster results.[4] An investment in CER includes not only the research itself but also the governance and technology infrastructures needed to support data sharing. Health data collected as part of routine clinical care, such as electronic health record data, payer claims data, and other administrative data, represent a potentially invaluable resource for observational studies and for building the evidence base relevant to diverse patient populations receiving care in "real world" settings, under "real world" conditions.

Even before existing electronic health data may be made available for CER,[5] it is essential to address a wide range of policy, governance, and technology challenges. One serious challenge is the absence of uniform data standards for the capture, storage, and transfer of data needed in order to ensure semantic harmonization. Data access policies and security requirements must meet the needs of the health care entities that are the guardians of the data and comply, for example, with the regulations under the Health Insurance Portability and Accountability Act (HIPAA)[6] while decreasing the burden of data access for data contributors and investigators seeking value in the data. Governance structures that address data standards, data use, data stewardship, and the monetization of data are critical for successful data sharing. Technology must be flexible enough to accommodate an evolving public understanding and emerging standards and policies. Data-sharing infrastructures must ensure that the promised benefits associated with data collected in the course of routine care is supported, by (1) ensuring that the necessary information about the characteristics of the real-world persons (e.g., eth-

[i]University of Colorado Anschutz Medical Campus; [ii]American Academy of Family Physicians; [iii]Recombinant by Deloitte; [iv]The Ohio State University

nicity, language of preference, education, income) and real-world settings (e.g., provider type, practice processes such as use of registries) is available to investigators; (2) supporting person-level identity linkage of data from disparate sources to permit a complete view of a person's interactions with the health care system; (3) and supporting mechanisms that allow access to data regarding persons with rare diseases without jeopardizing their anonymity. Finally, the resources required to establish, participate in, and use data-sharing infrastructure must not be cost prohibitive.

Distributed data networks (DDN) are one possible infrastructure solution for overcoming many of the above challenges and enabling access to data for research purposes. With a DDN, there is no centralized database; instead, each data-sharing partner stores its data locally (or, in some cases, entrusts it to a third party) and thereby controls access to its own data.[7] DDNs are typically connected on a network or grid that supports access through a data request portal and provides methods to monitor and control access. In a DDN, each organization prepares its own data for possible sharing by standardizing data storage to a common data model and an agreed-upon terminology system. De-identifying the data available for sharing depends on the purpose of the infrastructure. For example, the availability of personal identifiers supports data use for prospective clinical trials and patient recruitment but requires greater governance considerations (e.g., ensuring patient consent and institutional review board [IRB] approvals across the network) and security. On the other hand, restricting identifiers to those allowed in a HIPAA-defined limited data set may facilitate data sharing. In a DDN, each organization may maintain its own grid-enabled database and set its own policies for data access and network participation. Based on governance rules established by the network and with appropriate IRB approvals, those seeking to use the data for research purposes may submit either requests for study-specific data sets that are compiled across the network databases or more detailed analytic queries that return aggregate results.

The Scalable Architecture for Federated Translational Inquiries Network (SAFTINet)

SAFTINet is one of three national Agency for Healthcare Research and Quality (AHRQ)-funded projects charged with developing a DDN to support CER. SAFTINet's founding partners mainly include stakeholders whose priority is the care of safety-net populations. The SAFTINet DDN's technical objectives are to create and deploy:

1. A secure, trusted network environment that establishes a network of partner-specific Internet/grid-enabled databases (henceforth referred to as Grid Nodes) and a common central portal that processes queries and data extractions
2. A central query portal system, accessible via the Internet, that manages user authentication, authorization and query functions, to support approved data requests
3. A common data model and a common terminology that specifies the shared network database architecture
4. Applications to support data transformation, concept mapping, and data loading into the database
5. Applications for managing and linking patient identities between electronic health records, clinical data repositories, and administrative claims data using both clear-text and encrypted (privacy protected) record linkage methods
6. Applications for data validation and data quality reporting

The purpose of this publication is to describe the SAFTINet technical solutions for a DDN. We provide a high-level overview of the network's technical architecture and focus sharply on each component.

Methods

SAFTINet DDN Technical Infrastructure Overview

The SAFTINet distributed data network technical infrastructure includes several key components:

- A Query System composed of a Web-based Query Portal (QP) and Federated Query Processor (FQP) that is responsible for managing user access and data requests across one or more data-sharing partner Grid Nodes
- Translational Informatics and Data Management Grid (TRIAD) services that are responsible for secure and authorized communications between the Query System and the partner Grid Nodes
- Reusable OMOP-SAFTINet Interface Transformation Adaptor (ROSITA) data extraction, transformation, and loading middleware that is responsible for transforming a standardized data extract with idiosyncratic codes into the network common data model and terminology, which is the Observational Medical Outcomes Common Data Model Version 4 (OMOP CDM V4)
- Partner Grid Nodes with data formatted as a HIPAA-compliant limited data set in the OMOP CDM V4 format

In **Figure 1**, the overview schematic displays the network relationships. Partners establish Grid Nodes containing databases of standardized electronic health data conforming to the OMOP CDM V4 format. Authorized users request data from partner Grid Nodes via a Web-based QP. The QP receives and transmits requests for data (queries) to the FQP, and the FQP then returns either aggregate counts or row-level data from across the network's Grid Nodes to the QP for retrieval. Several IRBs have approved the SAFTINet DDN infrastructure.

Figure 1. Infrastructure Overview

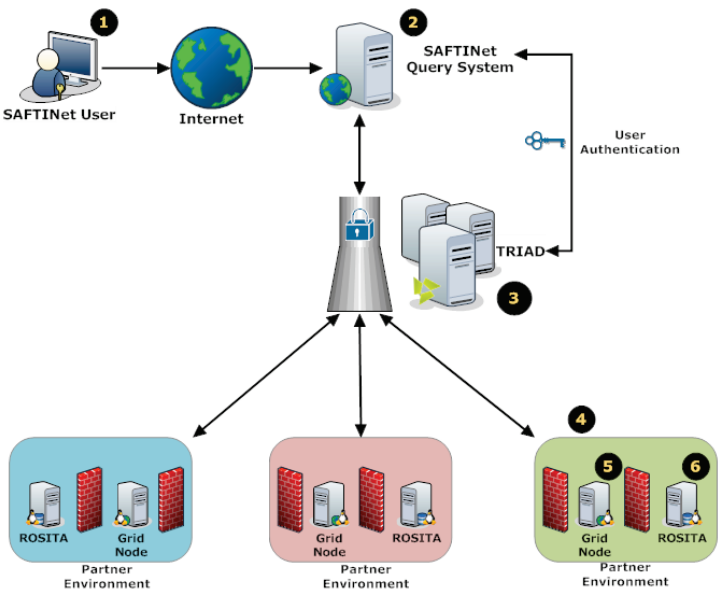
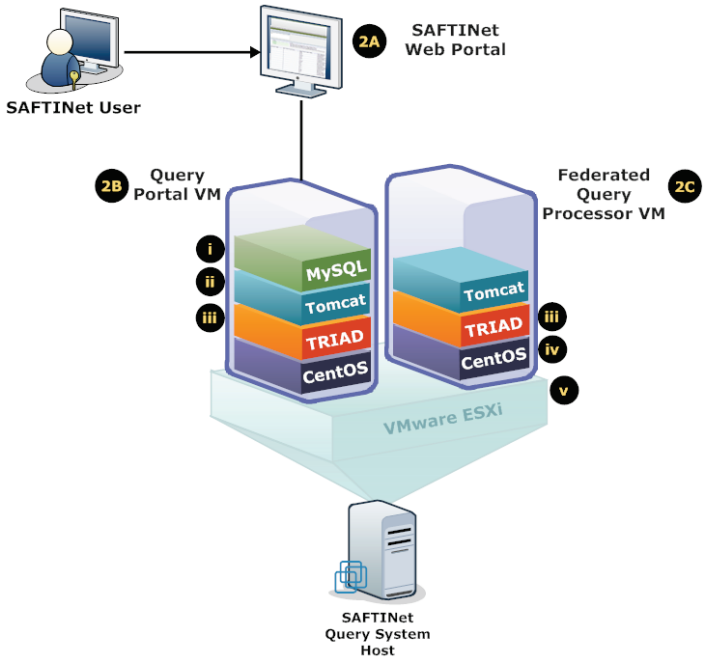


Figure 2. Query System Components



1. SAFTINet Web Portal

From an Internet browser on any computer, authorized users may request data via a secure Web-based Portal (https://saftinet.ucdenver.edu). This Web-based user interface (UI), part of the SAFTINet Query System, provides the user services (or client services) to the Query System. Users include central SAFTINet system administrators, data-sharing partners, and investigators requesting data.

2. The SAFTINet Query System

The SAFTINet Query System's key functional components **(Figure 2)** are the Query Portal (QP) and the Federated Query Processor (FQP). The Query System applications are hosted on secure servers in a VMware environment at the University of Colorado Anschutz Medical Campus (CU-AMC). A VMware ESXi virtualization environment separates the QP and the FQP. The host server contains several processors with several cores, large RAM capacity, large redundant storage capacity, and high-speed network hardware. The Query System receives and processes queries submitted by the user via the Web Portal. The QP transmits queries submitted by the user to the FQP over Secure Sockets Layer (SSL)/Transport Layer Security (TLS). The FQP contacts each Grid Node selected by the user on the QP and submits the user's query to that Grid Node. Query results are then compiled on the FQP and presented to the user securely over SSL/TLS on the Portal for the user to export. The QP and each Grid Node maintains its own lists of authorized groups and users via TRIAD. TRIAD data services authorize the FQP to send a query to a Grid Node and authorize a Grid Node execution of the query.

2A. SAFTINet Web Portal

The Web Portal and QP are the client-side and server-side QP applications, respectively. Using any Internet browser, the Web Portal GUI allows users to select the Grid Nodes they would like to query and to select the tables, variables (fields), and values (e.g., year of birth >1980) that they would like returned. Users may use the query builder interface to define cohort queries and store queries for future use. Results may be returned as aggregated counts or as row-level data sets in a .csv format. Authorized queries return results from all authorizing partner Grid Nodes and are combined into a single data set, which the user may then export for downloading to the user's computer. Query results are transmitted securely from partner Grid Nodes to the QP by using SSL/TLS cryptographic protocols.

2B. The Query Portal

The QP resides on a Virtual Machine (VM) that hosts a set of software packages that support the functionality of the QP, including the client-side Web portal applications.

QP Components

i. The MySQL Database Management System

MySQL provides support for the user query history functionality of the portal. It supports the database management functions. The history is accessible only from with the Query Portal VM by a function available only to the Query Portal Web application. Users are not able to view other users' query history. It is installed as a service on CentOS.

ii. The Apache Tomcat Web Server

Tomcat provides the framework from which all of the UI and data transactions originate on the Query Portal. The UI is built with standards-based Web languages, such as HTML, CSS, and Extensible Markup Language (XML). The secure grid-enabled data transactions are handled by using libraries and executables programmed in Java that are a part of the TRIAD services.

iii. TRIAD Service Clients

TRIAD allows the leveraging of TRIAD functions that support the trusted network. See TRIAD Section 3 below.

iv. The Community Enterprise Operating System (CentOS)

CentOS provides the operating platform on which all other software packages run. It is a community-supported, free, open-source operating system based on Red Hat Enterprise Linux.

v. VMware ESXi

VMWare ESXi is an enterprise virtualization environment that allows several virtual machines to run on a single hardware platform. ESXi handles the resource allocation of the hardware components to the Virtual Machines hosted in the environment to allow each machine to function simultaneously. This technology significantly reduces the costs and resources needed to support SAFTINet by supporting several systems on a single hardware platform. It also provides simplified deployment of the various SAFTINet systems by allowing each system to be packaged in a virtual machine template, which may be deployed rapidly in a virtualization environment by system administrators rather than installing each software component separately.

2C. The Federated Query Processor (FQP)

The FQP resides on a virtual machine and is responsible for processing queries and their results. It provides the Query Processing, Results Compilation, and Delivery services to SAFTINet QP. The TRIAD data service authorizes data requests before sending them to the Grid Nodes.

FQP Components

ii. The Apache Tomcat Web Server

The FQP uses Tomcat. See description in QP Components Section 2B.II above.

iii. TRIAD Service Clients

TRIAD allows leveraging of TRIAD functions that support the trusted network. See TRIAD Section 3 below.

iv. The Community Enterprise Operating System

The FQP uses CentOS. See description in QP Components Section 2B.IV above.

v. VMware ESXi

The FQP uses VMWare ESXi. See description in QP Components Section 2B.V above.

Translational Informatics and Data Management Grid (TRIAD)

TRIAD **(Figure 3),** an application of the caGrid architecture developed for the National Cancer Institute's caBIG project, is an Authentication, Authorization, and Accounting (AAA) system and serves as the trusted communication and grid networking fabric for SAFTINet. The Biomedical Informatics Program at the Ohio State University (OSU) hosts the core TRIAD services and functionality on which SAFTINet relies. It is also possible for a network to implement and host TRIAD independently. TRIAD's middleware system is designed to create a loosely coupled yet highly interoperable grid service-oriented architecture (SOA). It has been adopted as the basis for the TRIAD Community grid system. The TRIAD System provides two primary classes of services: (1) Security and Indexing Services and (2) Data Services.

Security and Indexing services provide functions such as provisioning user accounts, distributing trusted certificates, temporarily delegating user identities, and managing group-based authorization. Data Services support data sharing on the TRIAD grid. These services implement a consistent query interface and publish metadata describing the structure of the data they make available. In combination, these two capabilities allow for the creation of distributed queries and rapid integration of new data-sharing partners, i.e., Grid Nodes. Administrative functionality is provided within the above two classes of services and provides functionality to allow for overall management of the TRIAD infrastructure.

Figure 3. TRIAD Index and Security Services

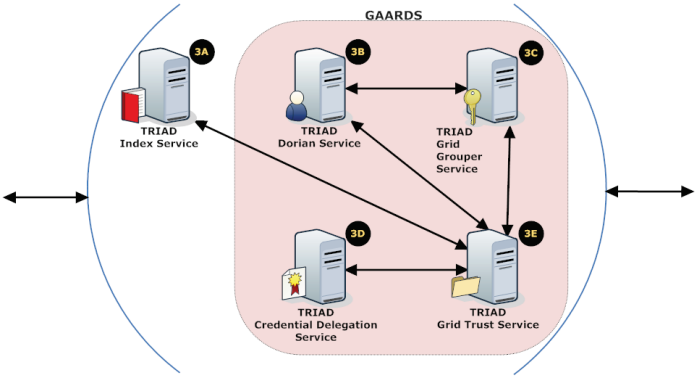
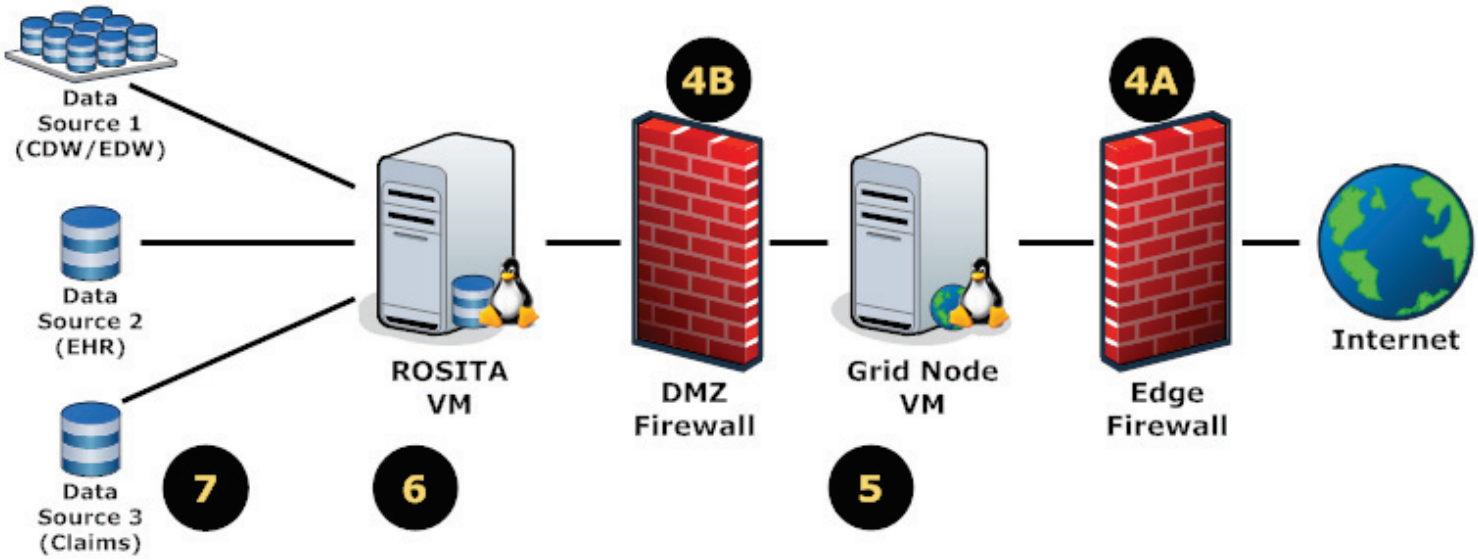


Figure 4: Partner Environment Components



TRIAD Index and Security Services provide security and indexing for the SAFTINet network. Four sub-services collectively comprise GAARDS (Grid Authorization and Authentication through Reliably Distributed Systems): the Grid Grouper Service, the Dorian Service, the Grid Trust Service, and the Credential Delegation Service. In brief, upon receiving a query, the QP obtains a list of available Grid Nodes from the Index Service over HTTP, authenticates SAFTINet users with the Dorian Service over SSL/TLS, and authorizes SAFTINet users with the GridGrouper Service over SSL/TLS.

3A. The Index Service
The Index Service is a Web application running on a virtual machine at OSU. Each Grid Node registers with the Index Service of the SAFTINet network. The Index Service also contacts each Grid Node periodically to verify that the indexed Data Services provided to the Grid Node are still online. The Query Portal contacts the Index Service to determine which Data Nodes are currently available.

3B. The Dorian Service
A secure Web application runs on a virtual machine at OSU. The Dorian Service is contacted by the SAFTINet Query Portal, FQP, and the partner SAFTINet Data Services to authenticate SAFTINet users.

3C. The Grid Grouper Service
A secure Web application running on a virtual machine at OSU specifies a set of trusted users for a grid service. It is contacted by a Grid Node to determine which SAFTINet users are authorized to obtain data from the Grid Node. A trusted central administrator or the Grid Node's administrator may manage the service.

3D. The Credential Delegation Service (CDS)
A secure Web application running on a virtual machine at OSU maintains the SSL/TLS certificates that are used to encrypt data moving between the Query System and the Grid Nodes. The CDS makes it possible for the FQP service to perform queries as a specific user through use of the user's delegated credential.

3E. The Grid Trust Service
A secure Web application running on a virtual machine at OSU establishes a shared security fabric for all the applications, services, and users of the grid network. Through the use of distributed and synchronized certificates, it maintains associations among users and Grid Nodes belonging to a particular grid network, such as SAFTINet. It ensures that all partners abide by the same data security rules.

Partner Network Environments

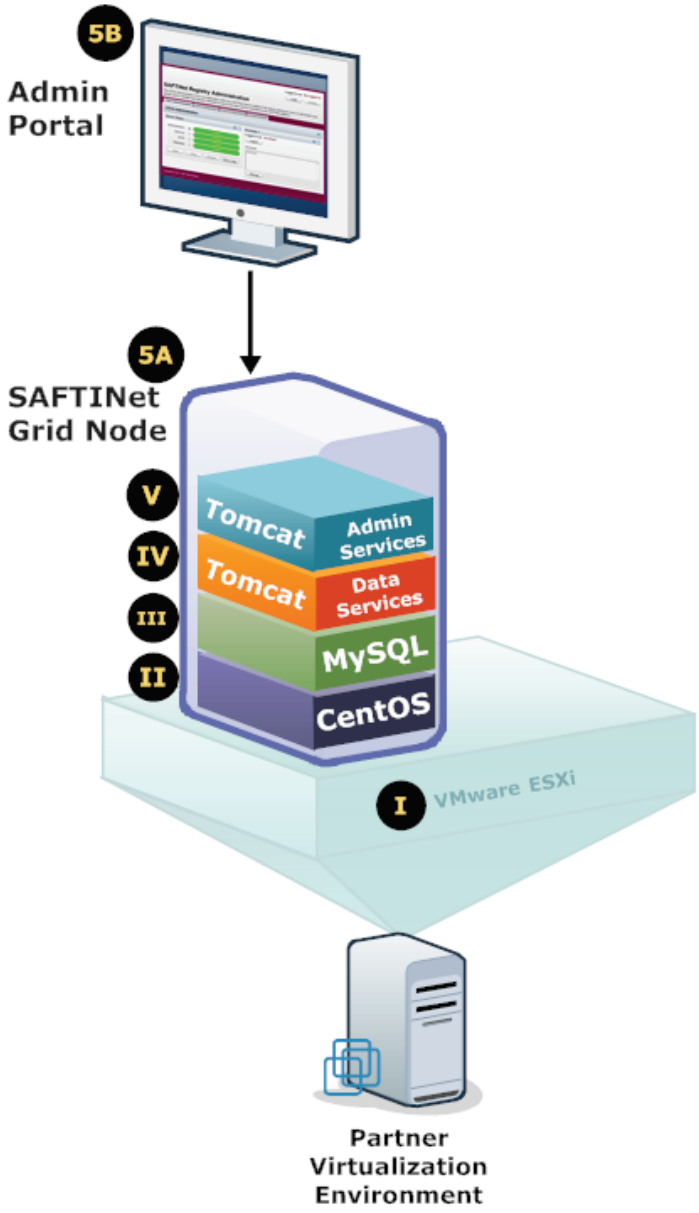
Partners host two VMs within their network environments: a Grid Node and a ROSITA system **(Figure 4)**. The Grid Node resides behind edge firewalls within network environments that are configured to communicate only with the Query System and the TRIAD index and security services, in what is referred to as a De-Militarized Zone (DMZ). Partner networks are configured according to each partner's own security policy. An external "edge" firewall protects partner Grid Nodes and allows only FQP requests to enter a protected portion of their network—the DMZ—where the Grid Node with data resides and to return data under authorized circumstances. The data Grid Node is further partitioned from a partner's internal network by another firewall that allows traffic to flow from ROSITA to the Grid Node where the data is made available for querying. Only certain communications are allowed to move from the Grid Node to ROSITA, and the ROSITA administrator must initiate those communications. To make the most efficient use of space, SAFTINet partners have elected to stand up the SAFTINet architecture within local virtualization environments. The Grid Node and ROSITA system are configured as VMs and thus may be installed on a single host, separated by firewalls, or on separate hosts, according to partner preferences.

4A. The Edge Firewall
The Edge Firewall is the outermost security device that filters traffic between the Internet and the partner's network environment. To safeguard the systems that lie behind the firewall, rules allow only certain types of traffic to certain systems. Given that all communications involving the HIPAA Limited Dataset hosted on the Grid Node use the SSL/TLS data encryption protocols, the Edge Firewall must allow traffic in and out to the Grid Node on specified ports. To secure and restrict access to the Grid Node even further, the addresses of the other systems that will communicate with the Grid Node—such as the TRIAD and FQP systems—may be specified as the only allowable systems to communicate through the Edge Firewall.

4B. The De-Militarized Zone Firewall
The DMZ Firewall provides a second layer of security for internal private network systems to protect against access by unauthorized external systems. The DMZ Firewall restricts communication among systems within the DMZ and within the partner network according to specific access policies implemented by each partner. For administrators to monitor and manage the Grid Node, Secure Shell (SSH) and SSL/TLS traffic is allowed from the internal network to the Grid Node on specified ports. In addition, for the ROSITA Server to transmit the de-identified and translated data to the Grid Node, Java DataBase Connectivity (JDBC) traffic from the ROSITA Server to the Grid Node is allowed. To ensure that only partner-approved data transfers are allowed, only the ROSITA administrator may initiate communication with the Grid Node through the DMZ Firewall.

5. Grid Nodes and Data Services
A partner Grid Node **(Figure 5)** resides on a virtual machine hosted in the DMZ of the partner environment—a partitioned area of the network that is accessible in limited fashion by external systems. The TRIAD Data Services includes Web services that support communication between the Grid Nodes and the Query System at CU-AMC. The Grid Node receives data across the DMZ Firewall from the ROSITA system. Grid Node components include TRIAD Services Client software and Observational Medical Outcomes Partnership Common Data Model V4 (OMOP CDM V4) formatted databases that store clinical and person-level claims data. Identifiers in the databases are restricted to those allowed under the definition of HIPAA-defined Limited Datasets (i.e., visit dates, birth dates, dates of death, and five-digit-or-greater ZIP codes).

Figure 5. Grid Node Components



5A. Grid Node Software Applications

I. VMware ESXi

The Grid Node software is packaged as a virtual machine and distributed in the Open Virtualization Format (OVF) to support deployment within a virtualization environment as VMWare ESXi. All current SAFTINet partners use VMWare ESXi, but other virtualization software may be used if preferred by a partner. See description of VMWare ESXi in QP Components section 2B.V above.

II. The Community Enterprise Operating System (CentOS)

The Grid Node Server uses CentOS. See description in QP Components section 2B.IV above.

III. MySQL

MySQL is the database management system that houses the OMOP V4 CDM HIPAA-compliant Limited Dataset on each Grid Node. It also maintains the access control list to the database, allowing the TRIAD Services and the ROSITA server direct access to the database. In this way, all access to the data is limited to an application interface and does not permit direct interface by users.

IV. SAFTINet Data Service

The SAFTINet Data Service provides the Web Services interfaces that enable the query request and retrieval of data between the Grid Nodes and the Query System. Created with TRIAD middleware, the Data Service supports standard functionality of Grid Node discoverability, querying, and optional functionality of increased security using Grid Grouper. The Grid Node is deployed within a secure Tomcat server, which provides the HTTP communication capability for the SAFTINet Data Services. For Tomcat details, see Section 2B.II above.

V. SAFTINet Administration Service

The SAFTINet Administration Service allows partners to control access to their data. The Administration Service is deployed within a secure Tomcat server, which supports the Web-based front end for the Administration Portal.

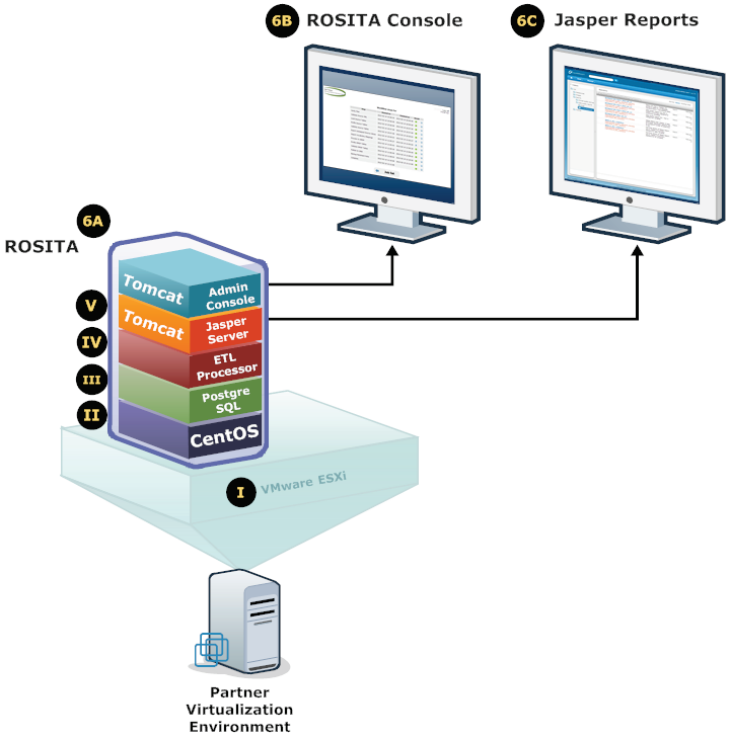
5B. Administration Portal

The Administration Portal is a secure Web-based user interface that allows local partner administrators to perform management of the SAFTINet TRIAD software deployed on the Grid Node. Administrators may start, stop, restart, and view logs for the Data Service and database as well as manage security configuration of the SAFTINet Data Service.

6. Reusable OMOP-SAFTINet Interface Transformation Adaptor System

The ROSITA system **(Figure 6)** supports the consumption, transformation, and loading of clinical and administrative data from partner electronic health records (EHR) (or surrogate EHR databases such as data warehouse extracts) and payer claims data to a Grid Node database. We use the term "consumption" because ROSITA does not actively profile or extract data from other data systems but rather processes clear-text data extracted from the partner's source system and transmitted to ROSITA in XML or delimited flat file formats. ROSITA also supports mapping to OMOP CDM V4 format and OMOP-compiled standardized terminologies and concept identifiers. A ROSITA version now in development supports patient-level record linkage of data from disparate sources (e.g., clinical and claims data), computation of descriptive data quality statistics, derived variables, and simple performance measures. The current version of ROSITA has been released under an open-source license.

Figure 6. ROSITA System Components



6A. ROSITA Server Hardware and Software

ROSITA Server Hardware

ROSITA requires a large resource allocation from the virtualization environment. It must run on a server with several processors with several cores, a large amount of RAM, high- capacity redundant storage, and high-speed networking hardware.

I. VMware ESXi

The ROSITA system uses VMWare ESXi. See description in QP Components Section 2B.V above.

II. The Community Enterprise Operating System

The ROSITA Server uses CentOS. See description in QP Components Section 2B.IV above.

III. PostgreSQL Database Management System

A database persists source data, data-profiling results, logging data, standardized vocabularies, and OMOP data and executes PostgreSQL functions to process the data. The database also manages the JasperServer database that holds all of the JasperReports, user accounts, and access rights for executing reports in ROSITA. PostgresSQL was selected over MySQL because of (1) concerns that MySQL may not remain freely available in an open-source environment (Oracle had recently acquired it at the start of our design process), and (2) PostgreSQL supports stored procedures, which were not supported in the then-available version of MySQL.

IV. The Extraction Translation and Loader (ETL) Processor

ROSITA accepts clear-text fully identified data that have been extracted from partner data sources, in XML or delimited file formats, into a preliminary PostgreSQL database. A combination of Java and PostgreSQL methods profiles and processes the data for loading into a PostgreSQL database that conforms to the OMOP CDM V4. All direct identifiers, as defined by HIPAA, are removed, with only prescribed indirect identifiers (birth dates, service dates, and ZIP codes) remaining per HIPAA Limited Dataset specifications. The data are pushed through the translation process, with the source values mapped to OMOP CDV V4 fields and OMOP concept-identifiers. As a part of this process, unrecognized or unmapped terms and codes values are exported and available in a Comma Separated Values (.csv) file format for manually building terminology mapping for each unmapped value and ensuring that all data are translated successfully. Once the data have been translated and validated against the terminology mappings, ROSITA loads the data to the Grid Node MySQL database via a Java Database Connectivity (JDBC) protocol.

V. The Tomcat Web Server

ROSITA uses Tomcat to provide the front end for the ROSITA Administration Console and the JasperServer Web Application. For Tomcat details, see Section 2B.II above.

6B. The ROSITA Administration Console

The ROSITA application includes a Web-based administrative console executed under the Tomcat Application Server. The Administration Console allows a user to select a single XML file or several delimited files to be processed through ROSITA and translated into the OMOP CDM V4, to export unmapped source values and import updated vocabulary mappings, and to load the translated data to the Grid Node. Both the ROSITA Administration Console and JasperServer application are accessible only by local area network connections.

6C. JasperServer

JasperServer application provides ROSITA's reporting functionality. JasperServer is a Web application that allows partner data administrators and researchers to run data quality reports as well as a variety of other internal reports on the clinical and claims data stored in the ROSITA databases.

7. Data Sources

In a broad sense, Data Sources are points of origination for patient-related clinical, claims, and administrative data that feed into the ROSITA system. Data may originate from any source, but, to be processed correctly by ROSITA, data must be formatted in alliance with the SAFTINet-specific OMOP CDM V4 ETL Specifications. Potential data sources include local partner clinical data warehouses (CDW) or enterprise data warehouses (EDW), EHRs, or claims databases.

Discussion

Three essential objectives drove the initial design of the SAFTINet DDN infrastructure. First, we needed to harmonize data from disparate data sources, both semantically and syntactically. We met the objective by selecting a common data model, the OMOP CDM V4 and its standardized vocabulary. When we initiated our project, the available OMOP CDM was Version 2. We worked alongside the Foundation for the National Institutes of Health OMOP national community during the transformation to CDM V4 (V3 was a working interim version that was not publicly released) to ensure that the model satisfied our use case for comparative effectiveness research, particularly health care delivery comparisons. Modifications included the addition of new data tables and fields to accommodate (1) insurer benefit and cost information; (2) information about organizations, practices, and providers; and (3) maintenance of population cohort identities via a dedicated cohort table.

We developed detailed ETL specifications to guide partners in profiling and mapping their source data to the OMOP CDM and to provide clarity as to how ROSITA would handle their data along the continuum from consumption to Grid Node availability. The ETL guidelines ensure that partners make uniform decisions in translating their source data to the common data model. Although a discussion of our experience with data harmonization is beyond the scope of this paper, we encountered major challenges such as ensuring the consistent interpretation and use of variables. For example, the OMOP CDM uses "type" variables to indicate important metadata properties, thereby assisting with data transformation and investigators' common interpretation. A "type" variable that describes medication data allows users to indicate whether the information source is a prescription, a fulfillment, an administration, or a claim. Simpler harmonization threats occur due to non-standard use of EHR systems such as the incorrect use of race and ethnicity categories, when, for example, "Hispanic" may be entered as a race.

The ROSITA software facilitates accurate field mapping to the OMOP CDM and mapping of source values to OMOP concept identifiers, which are maintained in the OMOP standard vocabulary. ROSITA is currently in Phase 2 development and will include clear-text record linkage capabilities for linking clinical and claims data and will compute and display data quality metrics to improve data quality transparency. ROSITA 1.01 is available via open-source APACHE 2.0 license at https://github.com/SAFTINet.

Second, we required a secure network that would support distributed data requests and retrieval of large data sets. We met the second objective by using TRIAD, an application of the caGrid architecture developed for the National Cancer Institute's caBIG project.[8-10] Grid computing technology such as TRIAD provides a strong security model by using SOA,[11, 12] allowing for better integration with current and future data and analytic technologies.[13, 14] TRIAD allowed us to leverage the existing FQP, which requires use of DCQL (Distributed caGrid Query Language). We demonstrated in earlier work that use of the FQP with DCQL supported large-scale queries across several nodes.[15, 16] Earlier work demonstrated acceptable degradation in processing time with grid deployment on virtual machines compared to physical machines, where the return of roughly 2 million person records across 32 nodes took approximately 50 minutes, which was 8.4 percent more time than with physical machine deployment.[13]

Third, we set forth a long-term objective to develop systems that would be both scalable and sustainable. Our investment in the ROSITA software development and our decision to use open-source components for ROSITA is a prime example of how we sought to achieve this objective. To minimize the expertise and resources required by data-sharing partners deploying ROSITA and the Grid Node, we package all required technologies into two preconfigured virtual machines that require only general technical knowledge about system administration and virtual system management. We provide a detailed technical systems guide to aid in the deployment of ROSITA and the Grid Node at partner sites. We employ central resources with expertise in network and system administration, ETL, the OMOP CDM, software development, record linkage, data quality reporting, and query development to assist partners and investigators with maintenance and use of SAFTINet systems. Finally, we are investing in the development and refinement of a Web-based query portal with an intuitive user interface for querying the Grid Nodes.

Governance of the Network

Before implementation of the SAFTINet technology, we needed to establish a governance structure and policies for participation and data use. Ensuring appropriate data use is the highest priority for all SAFTINet partners. Partners also required assurances that network participation would not render their internal information systems vulnerable. The SAFTINet technology embodies several features designed to protect the data and internal networks. ROSITA is responsible for removing direct person identifiers, such as names, Social Security numbers, and medical record numbers, with the exception of dates (birth, death, visit) and geographic information such as county, city, state, and ZIP code, before publishing data to the grid node. TRIAD encrypts data during transfer, ensures that only authorized users may access the QP to make data requests, and monitors partners use within a given network. However, technology cannot enforce the appropriate use of data once obtained, and SAFTINet partners requested the development and execution of several written agreements, policies, and procedures, including:

1. A Master Consortium Agreement (MCA) stipulates partners' rights and responsibilities, including policies regarding data requests, publications, data-sharing responsibilities, and membership termination. The agreement also stipulates that investigators requesting data will have a written protocol available to all data- sharing partners, will obtain IRB approval as needed, and will sign a data use agreement (DUA) outlining the appropriate use, storage, and destruction of data. All partners and the SAFTINet central development and support team at CU-AMC signed the MCA.

2. A Service Level Objective (SLO) agreement between the SAFTINet central team and each partner outlines the roles and responsibilities of each for installation, maintenance, and access control for SAFTINet technology and provides network configuration guidelines for creating a secure environment to host ROSITA and the Grid Node.

3. A SAFTINet security framework document describes a wide range of potential vulnerabilities to data and network security and corresponding mitigation strategies.

4. Given the need for patient identifiers to support de-duplication and record-linkage, SAFTINet created a version of OMOP CDM V4 that includes additional fields for identifiers used only for the ROSITA ETL and record linkage processes. It also informs partners of the data transformations that occur in processing, ensuring that the final step in publishing their data to their Grid Node produces a fully compliant OMOP CDM V4 limited data set. Given the weighty responsibilities and regulations to which partners must abide to ensure appropriate data use, the SAFTINet ETL specifications document is valuable for its transparency.

Implementation of SAFTINet Technology

To date, we have fully deployed all components of the network. The query system, hosted at CU-AMC, is installed, configured, and ready for investigator use. Three active Grid Nodes are populated with clinical data transformed by ROSITA to the OMOP CDM V4. Three more nodes are expected to be online by September 2013. Despite detailed installation documentation, we experienced the typical glitches that are associated with new software installation and that require person-to-person communication. Although we attempted to standardize deployment processes, flexibility is crucial for scalability. For example, networking best practices would place ROSITA and the Grid Node on different servers separated by a firewall, but partitioning a single server to create two separate VM environments is also acceptable; we supported either arrangement.

Limitations of the Technology

For the first phase of ROSITA (ROSITA 1.0), we selected XML as the data format for sending data to ROSITA for consumption. Data coded in XML is easy to understand and easily processed by computers. XML, a W3C standard, is readily extensible, and the use of tags, attributes, and element structures allows for complete representation of the meaning of the data. XML also accommodates hierarchical data structures, which, we felt, would improve the correctness of the data transformation by enforcing hierarchies such as Organization→ Care Site→ Provider. We provided robust XML schema definitions (XSD) to allow errors to be identified in advance of ROSITA processing. However, XML proved not to be the best choice because of the verbosity of the code format. The attributes that allow for complete representation of the data syntax also lead to redundancy. We experienced a five- to six-fold increase in data load size from source flat file data to source XML-formatted data. Large XML files placed substantial processing time burden on partners. Importantly, partners had decided to perform complete data refreshes instead of incremental data feeds based data-change-capture routines. Hence, large data loads were expected with each data transfer, not just with the initial data load. In addition, some partners had limited expertise with XML. ROSITA V1.0.1 will allow data to be consumed in either XML or flat file formats, thereby providing an easier and more efficient data format alternative.

The largest burden for data-sharing partners is the initial data profiling, mapping and extraction routines that are required to identify and locate the SAFTINet-requested data in their Data Sources. For partners unwilling or unable to perform this initial task, SAFTINet typically recommends an outside company with the experience and skills to perform work. For this initial work, SAFTINet requested all diagnoses, medications, procedures, and encounter information, along with select demographic and observational (e.g., systolic blood pressure, weight) data types. Partners indicated that it would easier to provide ROSITA with all these data types rather than to filter on specific variables. ROSITA automates the mapping to OMOP concept-IDs where every source data uses a standard terminology (e.g., ICD-9, SNOMED, RxNorm) that is part of the OMOP standardized vocabulary. Idiosyncratic codes requiring customized, manual mapping are then uploaded and amended to the partner-specific rules that already exist in ROSITA. Customized mapping occurs on a need-to-use basis, as mapping all fields is time-consuming, and partners or investigators may never use many fields (e.g. serum globulin, urine pH). We feel that this solution is acceptable, and ROSITA provides a way for partners to amend source to concept-ID mapping tables. However, future enhancements to SAFTINet will reduce the burden associated with the initial extraction and streamline field mapping steps and terminology mapping.

Next Steps

The SAFTINet team and partners are currently completing a series of data validation steps to ensure accurate profiling and handling of source data during the course of OMOP CDM V4 transformation. Upgrades to the FQP and QP are underway to allow for an easier-to-use query-building interface that does not require knowledge of DCQL and that removes limits on the size of the data set returned via the FQP. Planned improvements to ROSITA will permit several data sources to load into a single ROSITA, thereby allowing smaller partners to share a single ROSITA instance; mechanism to support patient-level record linkage of data from disparate sources (e.g., clinical and claims data) by using clear-text identifiers; and more advanced computation of descriptive data quality statistics, derived variables, and simple performance measures.

Conclusions

We developed a robust technical DDN infrastructure and have deployed it successfully with three partners. We have also successfully implemented the TRIAD grid technology for distributed data sharing and the ROSITA software to facilitate harmonization of disparate data sources to a common data model. A primary goal of the technical infrastructure design has been to decrease the burden on partners by limiting the technical expertise and resources required for participation.

Appendix A.

Acronyms and Terms	
AHRQ	Agency for Healthcare Research and Quality
caBIG	cancer Biomedical Informatics Grid
caGrid	The computer network and software that support caBIG
CDM	Common Data Model
CDS	Credential Delegation Service
CDW	Clinical Data Warehouse
CentOS	Community Enterprise Operating System
CER	Comparative Effectiveness Research
CSS	Cascading Style Sheets
CSV	Comma Separated Value
CU-AMC	University of Colorado Anschutz Medical Campus
DCQL	Distributed caGrid Query Language
DDN	Distributed Data Network
DMZ	De-Militarized Zone
DUA	Data-Use Agreement
EDW	Enterprise Data Warehouse
EHR	Electronic Health Records
ETL	Extraction, Transformation, and Loading
FQP	Federated Query Processor
GAARDS	Grid Authorization and Authentication through Reliably Distributed Systems
HIPAA	Health Insurance Portability and Accountability Act
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
IRB	Institutional Review Board
Java	A general-purpose, concurrent, class-based, object-oriented computer programming language
JDBC	Java Database Connectivity
LDS	Limited Dataset
MCA	Master Consortium Agreement
OMOP	Observational Medical Outcomes Partnership
OSU	Ohio State University
OVF	Open Virtualization Format
PHI	Protected Health Information
QP	Query Portal
RAM	Random Access Memory
ROSITA	Reusable OMOP-SAFTINet Interface Transformation Adaptor
SAFTINet	Scalable Architecture for Federated Translational Inquiries Network
SLO	Service-Level Objectives
SOA	Service-Oriented Architecture
SQL	Structured Query Language
SSH	Secure Shell
SSL	Secure Sockets Layer
TLS	Transport Layer Security
TRIAD	Translational Informatics and Data Management Grid. An application of the framework implemented by caGrid
UI	User Interface
VM	Virtual Machine
XML	Extensible Markup Language
XSD	XML Schema Definitions
W3C	World Wide Web Consortium

Software Packages	
PostgreSQL	http://www.postgresql.org An open-source object-relational database system
VMware ESXi	http://www.vmware.com/ Vmware vSphere Hypervisor (free license available)
JasperReports Server (community edition)	http://community.jaspersoft.com/ An open-source reporting and analytics server from Jaspersoft
MySQL	http://www.mysql.com/ An open-source database, owned by Oracle
ROSITA	https://github.com/SAFTINet SAFTINet-developed open-source software for transformation of data to the OMOP CDM V4, creation of HIPAA-defined LDS, and loading of TRIAD node databases
Logos and Symbols	
	TRIAD Community Logo
	Linux Logo
	VMware logo
	Internet symbol
	Database symbol

References

1. Larsson ME, Kreuter M, Nordholm L. Is patient responsibility for managing musculoskeletal disorders related to self-reported better outcome of physiotherapy treatment? Physiotherapy theory and practice. 2010;26(5):308-17. Epub 2010/06/19.
2. Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. JAMA : the journal of the American Medical Association. 2003;290(12):1624-32. Epub 2003/09/25.
3. Slutsky JR, Clancy CM. Patient-centered comparative effectiveness research: essential for high-quality care. Arch Intern Med. 2010;170(5):403-4. Epub 2010/03/10.
4. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. The New England journal of medicine. 2000;342(25):1878-86. Epub 2000/06/22.
5. Peddicord D, Waldo AB, Boutin M, Grande T, Gutierrez L, Jr. A proposal to protect privacy of health information while accelerating comparative effectiveness research. Health affairs. 2010;29(11):2082-90. Epub 2010/11/03.
6. United S. Health Insurance Portability and Accountability Act of 1996. Public Law 104-191. United States statutes at large. 1996;110:1936-2103. Epub 1996/08/21.
7. Maro JC, Platt R, Holmes JH, Strom BL, Hennessy S, Lazarus R, et al. Design of a national distributed health data network. Annals of internal medicine. 2009;151(5):341-4. Epub 2009/07/30.
8. Payne P, Ervin D, Dhaval R, Borlawsky T, Lai A. TRIAD: The Translational Research Informatics and Data Management Grid. Applied clinical informatics. 2011;2(3):331-44. Epub 2011/01/01.
9. Oster S, Langella S, Hastings S, Ervin D, Madduri R, Phillips J, et al. caGrid 1.0: an enterprise Grid infrastructure for biomedical research. J Am Med Inform Assoc. 2008;15(2):138-49. Epub 2007/12/22.
10. Saltz J, Oster S, Hastings S, Langella S, Kurc T, Sanchez W, et al. caGrid: design and implementation of the core architecture of the cancer biomedical informatics grid. Bioinformatics. 2006;22(15):1910-6. Epub 2006/06/13.
11. Truong TN, Nayak M, Huynh HH, Cook T, Mahajan P, Tran LT, et al. Computational Science and Engineering Online (CSE-Online): a cyber-infrastructure for scientific computing. Journal of chemical information and modeling. 2006;46(3):971-84. Epub 2006/05/23.
12. Foster I. Service-oriented science. Science. 2005;308(5723):814-7. Epub 2005/05/10.
13. McGough AS, Cohen J, Darlington J, Katsiri E, Lee W, Panagiotidi S, et al. An End-to-end Workflow Pipeline for Large-scale Grid Computing. J Grid Computing. 2005;3(3-4):259-81.
14. Yu J, Buyya R. A Taxonomy of Workflow Management Systems for Grid Computing. J Grid Computing. 2005;3(3-4):171-200.
15. Kalra S, Megallaa MH, Jawad F. Patient-centered care in diabetology: From eminence-based, to evidence-based, to end user-based medicine. Indian J Endocrinol Metab. 2012;16(6):871-2. Epub 2012/12/12.
16. Price RC, Huth D, Smith J, Harper S, Pace W, Pulver G, et al. Federated queries for comparative effectiveness research: performance analysis. Studies in health technology and informatics. 2012;175:9-18. Epub 2012/09/04.

Acknowledgement

Funding was provided by R01-HS019908 from the Agency for Healthcare Research and Quality; Title: Scalable Architecture for Federated Translational Inquiries Network; PI: Lisa M. Schilling.