Data Privacy by Design: Digital Infrastructures for Clinical Collaborations

Prof R.O. Sinnott[1], O. Ajayi[2], A.J. Stell[3]
National e-Science Centre
University of Glasgow
United Kingdom

[1]*r.sinnott@nesc.gla.ac.uk*
[2]*o.ajayi@nesc.gla.ac.uk*
[3]*a.stell@nesc.gla.ac.uk*

Abstract. The clinical sciences have arguably the most stringent security demands on the adoption and roll-out of collaborative e-Infrastructure solutions such as those based upon Grid-based middleware. Experiences from the Medical Research Council (MRC) funded Virtual Organisations for Trials and Epidemiological Studies (VOTES) project and numerous other *real world* security driven projects at the UK e-Science National e-Science Centre (NeSC – www.nesc.ac.uk) have shown that whilst advanced Grid security and middleware solutions now offer capabilities to address many of the distributed data and security challenges in the clinical domain, the *real* clinical world as typified by organizations such as the National Health Service (NHS) in the UK are extremely wary of adoption of such technologies: firewalls; ethics; information governance, software validation, and the actual realities of existing infrastructures need to be considered from the outset. Based on these experiences we present a novel data linkage and anonymisation infrastructure that has been developed with close co-operation of the various stakeholders in the clinical domain (including the NHS) that addresses their concerns and satisfies the needs of the academic clinical research community. We demonstrate the implementation of this infrastructure through a representative clinical study on chronic diseases in Scotland.

Keywords: Grid security, data linkage, anonymisation, virtual organizations.

1. Introduction

Clinical research and healthcare provision should be complementary. Understanding the impact of a particular treatment on individuals, or the assessment of drug therapies and their potential side-effects on large scale cohorts, or indeed moving towards the promise of personalized post-genomic e-Health, offers great potential for new and improved clinical healthcare support [1]. In this space, software and IT systems play a hugely important role in potentially enabling the myriad of data sets and software systems categorizing clinical systems today, to be harnessed to tackle a variety of clinical research questions which can in turn help improve healthcare systems and ultimately, patient care.

One way in which this vision has been addressed is through e-Science and Grid-based middleware solutions. Whilst much Grid-based middleware has been targeted to addressing computationally bound research, as exemplified by the Large Hadron Collider (LHC) [2] and the establishment of major software infrastructures such as the Enabling Grids for e-Science (EGEE) software stack [3] which can provide seamless access to extensive high performance computing (HPC) computational resources in the search for the Higgs-Boson. The Grid-model of providing seamless access to a range of digital resources, be they computers, data sets or other resources, is a compelling one and applicable to many research domains however. The post-genomic life sciences in particular have much to be gained through the seamless models of Grid-based data access and usage of biological, clinical, social, geospatial and other data sets. Examples and proof of concept systems demonstrating the Grid-vision in the clinical and related disciplines are described in [4-6]. One fundamental difference between these systems and HPC-oriented Grid projects is the focus upon finer grained security. Whilst many domains require authentication of individuals supported for example through X509-based public key infrastructures (PKI), many domains (including the clinical domain) require finer-grained access control models.

[7-9] have extensively documented the security and usability problems associated with Grid-based X509 based public key infrastructures (PKI) models [10]. Yet the PKI based model is not without its advantages, perhaps the most importance of these are in the support for single sign-on, i.e. where the user is only required to authenticate once and can subsequently access resources across multiple heterogeneous sites. Coupling PKI-based authentication with additional access control mechanisms reflecting site-specific access and usage (authorization) policies capturing what a user is allowed to do on a local resource offers considerable possibilities.

Example systems showing how the adoption of role based access control (RBAC) [11] solutions such as PERMIS [12] using federated identity providers and attribute authorities based upon the Internet2 Shibboleth technologies (shibboleth.internet2.edu) or centralized attribute authorities such as VOMS [13] have demonstrated how seamless linkage to distributed services and data sets can be supported. Furthermore, other authorization approaches including identity based access control (IBAC) models [14] and process based access control (PBAC) [15] have been demonstrated to support finer grained access control policies that can be linked to distributed resources. Whilst each of these solutions has their own particular advantages and disadvantages and all have been demonstrated to support advanced authorization capabilities, it is the case that the *real world* as represented by organizations such as the NHS are not equipped with expertise in the deployment and usage of these technologies; nor are they immersed in the world of Grid middleware.

Thus despite numerous demonstrators showing how such systems can support the vision of secure seamless access to distributed clinical services and data sets, the reality is that the real world of patient health care will simply not allow the

deployment of such technologies. There are several reasons for this. Perhaps the most importance of these is the fact that the data owners, data providers, ethical bodies amongst numerous other stakeholders in the clinical space are naturally extremely wary of any new middleware solutions which have yet to be completely proven to be robust, reliable and rigorous enough to enforce all access decisions.

Put another way, site autonomy in the clinical domain is not just a basic requirement that lip service is given to, but an absolute essential consideration which must be strictly adhered to with potentially legal consequences if not. In short, there is at present a lack of real trust in these solutions and their usage for live access to real clinical data for real patients. Yet the vision of the Grid in supporting seamless, secure access to distributed resources and the challenges facing healthcare providers are well aligned. Rather than attempting to convince healthcare providers such as the NHS to install a particular set of middleware and embrace authorization technologies and ultimately trust that the combination of these solutions will work, new access paradigms are required. These need to be aligned with the working practices and understanding of the stakeholders in the clinical domain and ultimately be fail-proof. In this paper we present a model and implementation of a system providing secure access and linkage of clinical data for research purposes that addresses many of the worries and concerns of the clinical healthcare community.

We note that in developing this solution, we have worked over the last 4 years in close co-operation with the NHS to establish degrees of trust through various collaborative projects. Fundamentally the solution has been developed to address a given premise: namely, those healthcare providers do not exist to keep data private, but to provide healthcare. Through improved research practices and better understanding of the possibilities that exist in data access and usage improving healthcare and patient care more generally will be realized.

The rest of this paper is structured as follows. Section 2 provides an overview of some of the key concerns facing healthcare providers in making clinical data sets available for research purposes. Section 3 describes the architecture and key components of the virtual anonymisation Grid for access to clinical research data (Vanguard), and an overview of the typical interactions that take place in supporting secure data linkage and anonymisation. Section 4 describes the implementation of the Vanguard system and its application in a given study. Finally we draw some conclusions of the work as a whole and outline plans for the future.

2. Data Sharing Challenges in the Clinical Domain

Irrespective of the trust models or security infrastructures in place or guarantees that any middleware developers might make about the robustness and usability of their software and security solutions, it is the case that clinical data providers in the vast majority of cases, will simply not risk *direct* access to their data sets for research purposes. Instead, many clinical healthcare providers will often produce aggregated data sets and make these available through less secure and often ad-hoc approaches. CDs sent through the post containing unencrypted clinical data or unencrypted email attachments are two of the worst case scenarios for data sharing yet are not uncommon in practice.

To counter such positions of data guardians requires several considerations to be taken on board. Firstly, any solutions have to empower the stakeholders and not remove their essential roles in the access to and management of their data sets. As noted, site autonomy is not just a requirement but a fact that if violated will result in potential legal consequences. It is essential that this autonomy has to be beyond simply installing Grid middleware and security software, and managing it locally since few data providers are likely to wish to explore the currently complex offerings of Grid technology providers.

Secondly and following on from the first consideration, any solutions have to be based upon pragmatic considerations of usability and accept that any developed systems must fit in to existing clinical systems and practices. Rolling out Grid based X509-based public key infrastructures for authentication augmented with advanced authorization infrastructures has to be considered from clinical provider perspectives. It is the case that the majority of clinical software systems have not been developed with security in mind. Or more precisely, they have not been developed with providing secure, authorized access to potentially external collaborators in mind. Rather healthcare providers are typically busy working with a range of legacy systems far removed from the more exotic, Grid vision and associated middleware. These systems often have a history of decades of development and enhancements and cannot simply be replaced with a new middleware.

Thirdly, it is essential that any software solutions deployed on live clinical systems have to be designed, deployed, managed and monitored more generally with the worst case scenarios and risk assessment and threat analysis in mind. Compromises of clinical information which can lead to erroneous treatments being given are simply not tenable. Other equally damaging scenarios also exist however, e.g. where clinical data is accessed and used without express permission, or that it is linked with identifying data that results in disclosure of patient data. Such cases destroy public trust in clinical research and associated IT technologies.

Given all of this, direct access through NHS firewalls to live clinical data systems will simply not take place (other than potentially in small, closely contained scenarios where NHS and research collaborators are working on test systems for example). For the larger challenges of looking at national-level epidemiology or recruiting large patient cohorts for a clinical trial say, this model will simply not scale.

Instead, new models of data access and usage are required which meet the stringent requirements of the stakeholders in this domain yet address the needs of the clinical research community. Fundamentally, challenges facing clinical research with cautious data providers and associated stakeholders include:

- *How can it be ensured that the research is ethically driven?*
- *How can it be ensured that these data sets will not be disclosed to others?*
- *How can it be ensured that this data will not be linked with other data resources which may include identifying information?*
- *How can we guarantee that the NHS systems themselves will not be compromised through malicious attacks or accidental disclosure risks*?

The first bullet point is currently addressed in the UK through independent clinical and related expert arbiters. These are typically represented by independent ethics panels which can be local or regional/national depending upon the nature of the research proposed. These panels will typically have legal and patient information advisory group members also. Their role is typically to ensure that the research proposed is (or will be) beneficial to patients and, where any potential risks exist, that these are fully documented and ultimately used to decide whether the research should proceed. These panels are often supported by Caldicott guardians who act as independent clinical experts who can advise on the nature of the study being proposed and assess whether it is scientifically and ethically sound.

This process is inherently human driven and we believe should never be automated by a software process. However once ethical go ahead is obtained, it should be the case that systems are available to make the required data sets available in a secure environment.

The second bullet point can be addressed in different ways. Firstly, if direct trust exists between clinical data providers and the clinical researchers undertaking a given study, then the dangers of disclosure are minimized. Disclosure of data by a researcher will destroy trust and can have legal consequences and/or prohibit future access to clinical data. Alternatively, a common model is to only release anonymised data. However whilst anonymisation or pseudo-anonymisation models help to de-identify data by for example removing certain identifying information, it is notoriously difficult to truly anonymise clinical data in the most general case and make it still useful for the end user scientists. Furthermore in the post-genomic age, where there is a move towards personalized e-Health and greater emphasis on genetic information and data, further issues arise since the data by its very nature is identifying.

Nevertheless for many research questions it is often not necessary to know detailed identifying information. Rather, for feasibility studies in a given clinical trial say, having knowledge of how many patients in a given region suffer from a particular chronic disease is sufficient.

Following on from this and addressing the third bullet point is the possibility of linking these data sets with other data sets to obtain further information and potentially, to identify individuals in a given clinical data set. Ideally it should be the case that trust between clinical researchers would mean that this danger should not arise, i.e. it is typically part of the ethical agreements that are defined in a given study protocol. However, placing such trust on clinical researchers (who are often based act universities and not in themselves clinicians or healthcare providers) is not sufficient and any information that could potentially be used for further linkage should be removed. Or, as we describe later, linking of clinical data with other data should be made in a framework where further data linkage is not possible.

A major worry facing clinical data providers and healthcare systems is in any potential threats to their systems. In the UK, the NHS has a national level firewall that has been set up. This completely separates the NHS systems from the academic research community (which are supported through the JANET network) and the wider internet. Once through the NHS firewall, the majority of NHS systems do not have advanced access and authorization systems in place. Rather, the hospitals, GP practices and their associated IT systems are primarily protected through their fragmentation and lack of coordinated, interoperable framework. As an example, clinical information in one hospital or GP practice may not be readily available to clinicians in other hospitals. Whilst data is itself protected with this model, it is also not conducive to healthcare more generally. Major investments in the NHS IT infrastructure through the Connecting for Health initiative [16] are currently under way to rectify this situation.

To address these concerns and challenges, the Medical Research Council (MRC) funded Virtual Organisations for Trials and Epidemiological Studies (VOTES – www.nesc.ac.uk/hub/projects/votes) project has developed a range of clinical frameworks for secure access to a rich variety of clinical data. The VOTES project began in October 2005 and is in the final phases of completion. Specifically VOTES aimed to develop an e-Infrastructure supporting the various stages involved in the conduct of clinical trials and epidemiological studies, namely: patient recruitment including feasibility studies of whether a trial/study has sufficient patient numbers meeting the given criteria for that particular trial; collection of data throughout the course of the clinical trial/study as well as supporting overall study management.

Throughout each of these stages it is paramount that the right information is made available to the right individuals (and only those individuals) to ensure both information governance and ethical considerations are strictly adhered to, and that the results of any trials can be independently validated according to strict and measurable criteria. One key aim of VOTES was to develop a framework that could be applied for a range of clinical trials and studies and not simply develop a single bespoke system for a particular trial say. This has been achieved through the definition and implementation of a variety of clinical virtual organizations (CVOs) offering capabilities for data access and integration. These systems were designed upon user-oriented role based access control where services and data sets were made available through portals according to user privileges. Example trials undertaken within the existing VOTES systems include brain trauma trials [17], paediatric endocrinology trials with specific focus on congenital anomalies [18], primary care and secondary care trials [19] and more recently in establishing large scale patient recruitment, e.g. as part of the UK Biobank project (www.ukbiobank.ac.uk) which aims to recruit a cohort of 500,000 individuals for a range of clinical genetic studies.

The latest incarnation of the VOTES framework is the Vanguard system which has been designed with the aforementioned bullet points at the forefront of the design process. The major achievement of the Vanguard system is that it supports clinical data privacy by design as opposed to potentially intrusive, non-validated, Grid-based technological solutions.

3. Design of Vanguard

The design of Vanguard system is based upon a range of principals that must be strictly adhered to. We have deliberately developed the system architecture and key components to be simplistic and intuitive. We are fully aware

that NHS IT systems personnel and policy makers will have difficulties endorsing complex solutions.

Needless to say, the Vanguard system has been developed to run on a variety of platforms. This has meant that proprietary software and protocols have been avoided. Vanguard has also been developed to survive network outages and hardware failure with minimum disruption to end-users.

Perhaps the most important principal that we are focused upon in the design and implementation of the Vanguard system is with regard to information governance. As noted from the bullets before, we recognize that information must be exposed to the minimum extent possible or in many circumstances not at all – this implies that strong encryption must be used whenever data is exchanged between systems or temporarily stored outside of memory, and that datasets should be trimmed at source before transmission rather than on receipt. It is essential that ultimate control of access to datasets must reside locally with their owners.

A key consideration of the Vanguard system is with regard to the acknowledgment of the natural wariness (skepticism) of data providers. Experience from several years of working with clinical data providers is that they simply will not allow direct access through their firewalls to their data. As mentioned, there are many good reasons for this including lack of robust security infrastructure across the NHS.

Instead of direct connections through healthcare provider firewalls, the Vanguard system is based upon anonymous pull models of data linkage. Thus, rather than clinical data systems being queried directly, i.e. through opening of firewalls, queries are generated based upon a knowledge of the data sets (schemas) that exist at given clinical provider sites. If a given site has registered itself for participation in a given study, it may subsequently pull the generated queries into their clinical systems. Depending upon local security policies, these queries are validated and authorized, and if valid, will result in their execution. In short, the clinical systems are completely protected from inbound internet connections (and hence do not have to open their firewalls to the outside world!) but rather are based upon a model only allowing outbound connections to be established. The Vanguard system itself is being designed based upon this pull model. However the question of security must still be explicitly satisfied, i.e. what queries are being defined by whom and what artifacts are coordinating the access to and usage of clinical data resources to users with particular privileges.

The Vanguard system architecture is shown in Figure 1 and shows the following principal components:

- *Viewer* – which is used by researchers who require access to data;
- *Agent* – which is the intermediary between other components;
- *Guardian* – which manages access to and data release from local resources;
- *Banker* – which logs usage and maintains use accounts for the clinical data access and usage;

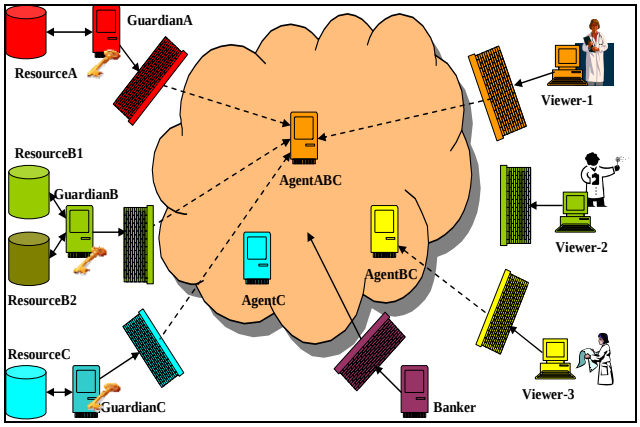

Figure 1: Vanguard Architecture

The roles of these components are defined as follows.

3.1. Viewer

The Viewer is an application run by the end-users of Vanguard. The viewer provides users with an interface to perform the following key functions:

- Display the different clinical data resources available to a particular user;
- Facilitate construction of data requests from these resources;
- Handle the datasets returned as a result of successful requests.

Currently, we are primarily focusing upon the viewers being web based browser interfaces to portals however specific client-specific applications can also be supported. These portals provide an environment where different agents can be accessed and used depending upon a given user's privileges. It is quite possible to have more than one agent involved in a given trial and study, however to begin with and to minimize the security risk we have focused on a model where each agent provides specific data linkages for a single study.

The data available to a user within a given viewer is dependent upon the privileges that they possess. To support this, we exploit digitally signed X509-based attribute certificates incorporating role based access control models. These attributes are specific to a given study or trial and allow access to one or more data resources, where data providers themselves agree how their combined data can be linked in particular ways. We emphasize that these authorization credentials are seamless and transparent to the end users and are used by the underlying framework only, i.e. the users are not required to manage or maintain their own credentials as is the case for X509-based PKIs.

Once defined, the security attributes are then used to enforce local access decisions – or more precisely, they are presented to the local data providers along with the queries that are generated. The combined roles and queries are then used to determine the authorization decision on access to the local data depending upon local authorization policy.

3.2. Agent

Agents play a pivotal part in the Vanguard system. An agent will typically perform the following roles:

- They enable communication between other system components including viewers, guardians, bankers and potentially other agents;
- They accept the generated user queries and manage the query requests themselves including their transfer, delivery and dealing with the associated security;
- They collate and manage the results of the submitted queries.

Agents are at the fulcrum of the Vanguard system in that they are responsible for ensuring both the secure communications between the different system artifacts and their coordination. A key role of an agent is in the generation of unique hash keys to support secure communications. These are passed through to the various Guardians that they help to coordinate. Through hashing of data items that should remain anonymous, data can be securely linked across sites without direct data disclosure being made. Key to this is in understanding the visibility of data and whether it can be directly accessed; used only for linking; or not accessible at all.

Agents also provide capabilities for defining the plans by which queries generated via a user through a viewer can be both formulated and subsequently enacted. Thus for example, if a particular data provider has a certain data element to be hashed or closed, then the Agent is responsible for ensuring that the appropriate hash keys are distributed and for removing the final hashed values once data has been linked across sites.

The Vanguard system supports a range of secure protocols reflecting the variety of infrastructures at clinical data provider sites. Examples of these include web security models exploiting X509 credentials for example as well as SSL/TLS for message/channel security respectively.

3.3. Guardian

Guardians act as a controlled secure gateway between the local data at clinical data provider sites and external Agents. One or more Guardian systems will be installed at every site which provides a data resource for the system. Guardians typically perform the following roles within the Vanguard system:

- Construction of site specific security policies;
- Describe the local data and security policies to known and trusted Agents;
- Enforce security policies related to incoming (pulled) queries;
- Handle incoming data requests from Agents;
- Export results of requests to Agents according to site specific data release policies.

A range of Guardians will exist within any given system – each managed and maintained by local data providers and adopting existing technologies and systems in place locally.

Guardians also allow data providers to make available their data models and importantly the way in which the data associated with these data models may be access and/or linked. Thus key to data access and linkage is knowledge of the underlying data models that are in place, i.e. the data schemas themselves. Based upon this and through negotiation with relevant ethics/oversight committees and local data providers themselves associated with a specific study, data might have be assigned three main forms of access privileges:

- *Open* – in which case the Guardian is willing to supply the actual value of the data field;
- *Hashed* – in which case the Guardian is willing to supply a hashed (and hence anonymised) value of the data;
- *Closed* – in which case the Guardian will not supply the value, but is willing to run queries for example that involve it as a selector.

We note that the clinical databases accessible within Vanguard may well have other fields which do not form part of this system. The existence of such fields is hidden entirely from the Agents. As a security precaution for data providers, Guardian `owners' are advised to create a set of read-only views of their data resources which contain the fields they are willing for a Guardian to process, and which do not contain any other fields.

Prior to running any queries, Guardians must supply a description of the data that it has to the Agent. This will typically contain a list of the names and types of the data resources a Guardian is managing; the version information of the Guardian and a list of the features that a Guardian can support. Initially the Vanguard system has primarily focused upon relational database resources supporting SQL-based queries hence this kind of information includes the list (names) of tables in the database; a textual description of the database contents; for each table a list of names of fields in table; text description of the table contents and the number of rows in each table; for each field in each table information might include the field type (int, string etc); the protection level (Open, Hashed, Closed); a textual description of the field contents; alternative nomenclature(s) for the field, e.g. SNOMED clinical codes; nullability; the uniqueness of the field; relationship to other fields, e.g. is-a-foreign-key; and the size of the field itself.

In short, the Guardian must provide detailed information on the data model for the databases it makes available so that this can then be used for data access and linkage by Agents. We emphasize that this is purely the data model that is being made available via the Guardian and not the data itself.

3.4. Banker

The Banker in the Vanguard system is responsible for managing resources across the whole system. Bankers have the following main roles:

- Maintain a log of actions taken across given trial systems – specifically through recording the queries generated by the viewers/agents and those sent to the guardians (including those that were denied);
- Maintain charging accounts for users – to ensure for example that a single user is not over-utilizing the federated data available through the Vanguard system.

3.5. Vanguard Component Interaction Scenario

The interactions between the previous components in supporting secure anonymous data access and linkage proceeds as follows. In the first instance, we assume that ethical approval for a given trial is applied for and granted as per typical procedures. In the UK this might for example be through applying for Multicentre Research Ethics Committee (MREC) or Local Research Ethics Committee (LREC) approval. In addition, we assume the precondition that an Agent for this particular trial has requested the data-schemas

from all Guardians it is aware of. This includes information on the visibility of the data sets themselves, e.g. whether they are open, closed or hashed. The trial coordinator will then use a Viewer to request the data-schemas available from that Agent to construct particular queries.

A trial coordinator may construct particular views of data for the different roles of individuals involved in a particular trial, e.g. a nurse may issue the following queries, or an ethical oversight committee member may see all data etc. The Vanguard system extends the existing VOTES systems to support such role-based scenarios. We outline the basic functionality here in terms of creating and executing queries and retrieving query results since they are similar irrespective of the role in the study.

3.5.1. Query Creation

To create a specific query the end-user uses the Viewer to construct a query based on the data-schema available to them. The query is transmitted from the Viewer to an Agent, where it is stored and given a unique ID.

3.5.2. Query Execution

To start executing a query the Viewer transmits a signal to the Agent requesting that a previously-stored query is executed. This is accompanied by a public-key generated by the end-user (PKU). The Agent verifies that the query is permitted and produces an action plan decomposing the query into local-queries across any Guardian system(s) required. Fields are tagged as either being required for external-joining (within the Agent) or purely for returning to the Viewer.

The Agent generates the queries and either issues them directly to the Guardian systems (where they allow direct querying), or signs and stores the queries. Each non-directly querying Guardian will then periodically pull these queries down and if valid/authorized, return the results associated with it. The query requests themselves are accompanied by PKU, a public-key for the Agent (PKA), and a unique (per-query) hashing key generated by the Agent (HA).

To execute the parts of this query each Guardian checks the local-query against its local access-schema to verify it is permitted. If satisfying the local policy, the Guardian executes the query. Fields tagged for external-joining by the Agent are hash-encoded using HA. Fields tagged for returning to the Viewer are encrypted using PKU. The whole datasets are then encrypted using the PKA and returned to the Agent along with the cost for executing the query in resource credits.

On receipt of the local-queries the Agent stores the resultant datasets until all partial queries have returned. Once all queries have returned, the Agent transmits a signal to the Banker with the action taken and the number of resource credits used, e.g. the number of result sets. In addition, the agent joins the partial queries according to any external-joining fields. It will also discard any external-joining fields that the end-user has not requested or that the end user does not have the privilege to view. Finally it sets a query flag to indicate that the query is completed.

3.5.3. Query Retrieval

Whilst a query is being executed the user is able to use the Viewer to see the state of progress of the query. Once the query-complete flag has been set on the Agent, the end-user uses the Viewer to download the results of the query from the Agent. At this point, the Agent deletes all data returned by the query and passes all logging/charging data to the Banker. Once acknowledged by the Banker, the Agent deletes this information from its cache.

3.6. Example of Vanguard System

To understand how these various components can be used for secure data linkage within the Vanguard system we consider the following example. We assume that a range of clinical datasets are distributed across clinical data provider sites *alpha*, *gamma* and *delta* as depicted in Figure 2 hosting the datasets stay and birth, linkage and disease respectively.

alpha.stay

Field	Type
hospID	Integer
mother	Integer
days	Integer
status	Integer

alpha.birth

Field	Type
nhs	String
mother	Integer
dob	Date
weight	Real
sex	Int

gamma.linkage

Field	Type
nhs	String
chi	Int
Active	Bool

delta.disease

Field	Type
chi	Int
hiv	Bool
hepatitis	Bool

Figure 2: Example Data Linkage Scenario

With the above tables we assume that the National Health Service (NHS) number in table alpha.birth, and the community health index number (CHI) number in data resource delta.disease must both be hashed (represented here through different colouring). Similarly, the HIV information in the database delta.disease is closed and hence cannot be disclosed. With this data model, in place we wish to answer the following query: *How many days did mothers with HIV stay in hospital?*

The SQL to run this query ***directly*** is represented in Figure 3.

```
SELECT alpha.stay.days WHERE alpha.stay.mother = alpha.birth.mother
                        AND alpha.birth.nhs = gamma.linkage.nhs
                        AND gamma.linkage.nhs = gamma.linkage.chi
                        AND gamma.linkage.chi = delta.disease.chi
                        AND delta.disease.hiv = true
```

Figure 3: Direct Querying of Clinical Data Sets

However, given that certain information associated with these data resources is not directly visible to the Agents and hence to the end users via the Viewers, the actual SQL plans that is generated by the Agents needs to link the data and remove unnecessary information or data which is flagged as being for restricted disclosure. In this case with the NHS and HIV data information limitations the SQL plan generated by the Agent in this case is shown in Figure 4.

```
SELECT alpha.stay.days,H(alpha.birth.nhs)
        WHERE alpha.stay.mother = alpha.birth.mother
    SELECT H(gamma.linkage.nhs),H(gamma.linkage.chi)
    SELECT H(delta.disease.chi) WHERE delta.disease.hiv = true;
    Join on H(*.nhs) AND H(*.chi), then remove H(*.nhs) and H(*.chi)
```

Figure 4: Anonymous Linkage of Clinical Data Sets

In this case, the data that is restricted, e.g. the NHS number in alpha.birth, is hashed (and hence anonymised). These hash values are unique since a different hash key is used each time by the Agents and it is guaranteed that the same hash key will

not be used by the Agents. Thus through the use of hashed information across the different data resources, data linkage can be made, yet direct data disclosure is avoided. Furthermore, the final line of the query above then removes the hashed values to ensure that the final resultant data set protects the required confidentiality of the data providers.

It should be emphasized that the primary benefit of this scenario, is that no identifying information is released from the data providers yet data linkage is made across different data provider sites. This model thus satisfies the data providers worries of their data potentially linked in unforeseen ways with other data resources. The result of the query will then simply be the number of days that mothers with HIV stayed in hospital, without any information identifying which mothers for example.

4. Vanguard Implementation

The implementation of the Vanguard system has largely focused upon the development of Viewers, Agents and Guardians with the development of the Banker primarily at the design phase only. This has been deliberate. Building confidence in these key components is paramount and auditing and charging "after the fact" is regarded as not as immediately essential for the various stakeholders involved.

To understand the interplay between the various components we describe a typical collaboration involving different clinical data providers. These include data providers offering hospital secondary care data through SCIstore software; primary care GP data through GPASS software, and Scottish Morbidity Records data made through the Information Services Division of the NHS. Each of these data providers have data securely stored whose schema is known and understood across the project partners.

In the first prototype of Vanguard, we implemented a Viewer for a particular Virtual Organization (VO) called VOTES2. This was primarily used to test out the basic technologies and the data sets that were used here are representative only. In the VOTES2 trial, some data items were identified that should be made available and linked across the partner sites in a secure setting. As shown in Figure 5, these included a description of the particular chronic disease of interest; the diagnosis of this disease; information related to the family name and the particular patient id. The patient information was not to be made available directly through the Viewer, however it was agreed that for the VOTES2 study, the data could be linked on a particular patient identifier as indicated in the visibility information of Figure 5.

Based upon this Viewer, a user is able to create a query which is encrypted with the Viewers public key and sent to an Agent along with the public key of the individual themselves, e.g. *PKV(Qx,PKUx)* where *PKV* is the public key of the viewer, *Qx* a particular query to be run and *PKU*x the public key of the User. The Agent verifies the Viewer key, checks the validity of the request, and subsequently defines a data linkage strategy based upon the agreements set out in the VOTES2 collaboration, i.e. a federated query needs to be generated which will be pulled down by the Guardians protecting access to the SCIstore, GPASS and SMR data sets respectively. At this point a unique hash key (*HA*) is also generated by the Agent.

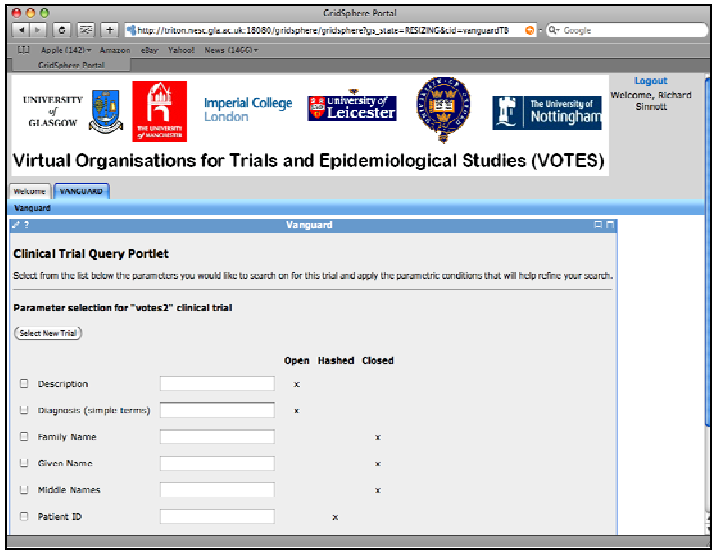

Figure 5: Viewer Interface for VOTES2 Clinical Trial

At some later time, a Guardian involved in the VOTES2 study will check to see if any queries are generated that it needs to deal with. When this is the case it pulls these queries in and checks that they are appropriately signed, i.e. from an Agent it trusts. For the SMR resource this looks like *PKA(SMRQuery,HAx,PKUx)* where *PKA* is the Agents public key, *SMRQuery* the query that is requested to be run against the SMR data set, *HAx* the unique hash key generated and *PKUx* the users public key. A typical query that a Guardian will see is shown in Figure 6.

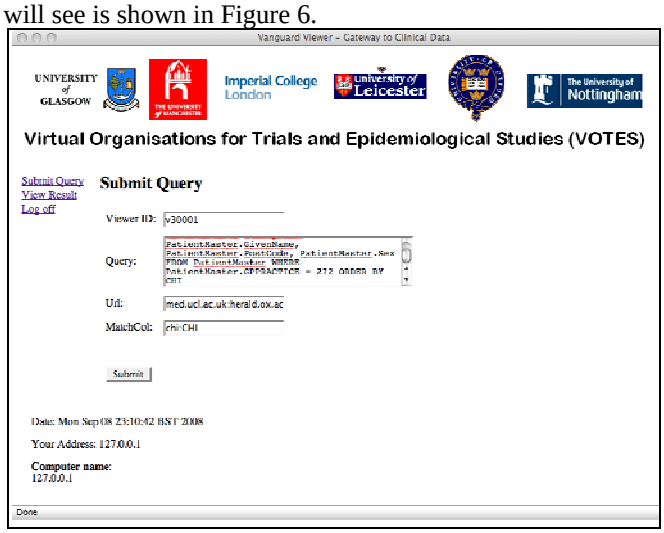

Figure 6: Agent Generated Query for a Guardian

Similar queries are pulled in to and verified by other Guardians involved in the VOTES2 study, i.e. for SMR and GPASS. Each data provider will assess the query (either through automated RBAC or similar approaches) or through non-automated mechanisms, e.g. discussions with organizational representatives. Assuming that the organization is satisfied with the request, the query is run. The data that can be linked only is hashed with the unique hash key from the Agent. The other contents of the message, i.e. the releasable data are encrypted using the public key of the individual user and the message as a whole encrypted and signed using the Agent's public key. For the SMR resource this looks like *PKA(PKUx(SMRres),HAxSMRres)* where *HAxSMRres* in this case will be the hashed patient id which

as defined by the Viewer in Figure 5 cannot be seen directly, but can be linked upon.

After receiving similar encrypted, hashed and encrypted results from all of the Guardians, the Agent can subsequently: decrypt the data using its own private key; join the resultant hashed data sets using the unique hash values that were generated previously, i.e. $PKA^{-1}(Join(HA_x^{-1}(HA_{xSMRres})\ldots),HAx^{-1}(HA_{xGPASSres})\ldots),HAx^{-1}(HA_{xSCIres})\ldots))$ where the "…" represent the other data sets that themselves are encrypted using the users public key. Once joined on the hashed keys, these other data sets are then themselves encrypted using the Viewer public key and released to the end users, i.e. *PKV(PKUx(Joined Linked Anonymised Data)).* The user, i.e. the holder of the private key then is thus able to decrypt the joined, linked and anonymised data from the Viewer. A typical Vanguard query result is shown in Figure 7.

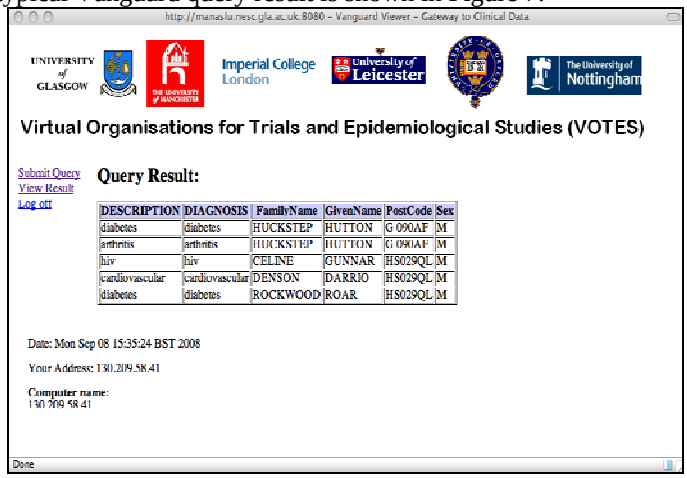

Figure 7: Viewer Results Interface

Once the results are obtained, the associated query result sets and are deleted from the Agent, however associated information of which queries were run, by which Viewer on behalf of which user are sent to the Banker for future auditing purposes.

It should be emphasized that no component in the Vanguard system is able to capture and link information directly. The Agents generate unique hash keys which are used only for data linkage of agreed fields (here patient id). The actual data associated with these patients is encrypted using the public key of the end users themselves. This linking and anonymisation based upon a pull-oriented model offers the possibility for new mechanisms of interacting with clinical providers which underpin a range of clinical-research collaborations.

5. Conclusions

The work on the Vanguard system is on-going however we have defined the components and the interactions between them that will overcome the immediate concerns from the clinical data providers as has arisen throughout the course of the VOTES project. Nevertheless the work we have described here is not complete and numerous challenges remain to be addressed. Some of the most critical challenges that we envisage include scalability. The scenarios outlined in this paper are at a small and understandable scale. In real clinical systems in Scotland such as SCIstore, GPASS and the Scottish Morbidity Records amongst others however we are often dealing with database models comprising several hundred tables with complex field and primary/foreign structures. Furthermore it is often the case that few clinical data centers are well positioned to disclose what data tables/fields should be made available. In the UK numerous solutions exist and are outsourced to commercial software providers who are often unwilling to disclose their detailed data models. As such, knowing what data can be disclosed and linked is often not a trivial exercise. This is often further complicated through fields that can be used for textual information on a given patient for example and including identifying patient information for example. To address such scenarios, we believe the only way is to develop the systems in close liaison with the clinical providers and only after they are completely satisfied that the systems meet their rigorous information governance policies can they be used in a truly live setting.

A second challenge that we expect to face in the roll-out of the Vanguard system is with regards to data disclosure risks arising due to global data models. Thus it is only when the various data providers have agreed to release their data sets that the issues of identifying data sets can arise. As one example, it would be quite possible to extend the *alpha.birth* data provider given above with the patient name and data of birth and allow open access to this information, but this may well be opposed to the data disclosure policy of *delta.disease* for example. It is only when considering the joining or union of these data sets on the CHI and NHS numbers in the example above that such policy conflicts can be identified.

We also note that many of the challenges faced in obtaining access to clinical data stem from researchers being considered as external to the clinical and administrative bodies such as the NHS. To overcome these issues, the NeSC team is in the process of being allocated NHS honorary contracts through the work on the breast cancer tissue bank for example.

5.1. Acknowledgements

The work described here was suported by a grant from Medical Research Council in the UK to support the efforts of the Virtual Organisations for Trials and Epidemiological Studies (VOTES) project. The authors thank the partners involved in the project for their inputs.

6. References

[1] R.O. Sinnott, M. Bayer, Controlling the Chaos: Developing Post-Genomic Grid Infrastructures, Life Science Grid Conference (LSGrid2005), May 2005, Singapore.
[2] Large Hadron Collider (LHC),
[3] Enabling Grids for E-sciencE (EGEE) project, http://public.eu-egee.org
[4] R.O. Sinnott, From Data Access and Integration to Mining of Secure Genomic Data Sets, International Journal of Grid Computing: Theory, Methods and Applications, Special Issue on Life Science Grids for Biomedicine and Bioinformatics, pp 447-456, Vol. 23, Issue 3, March 2007.
[5] R.O. Sinnott, O. Ajayi, J. Jiang, A. J. Stell, J. Watt, User-oriented Security Supporting Inter-disciplinary Life Science Research across the Grid, New Generation Computing, Special Edition on Life Science Grids, editors A. Konagaya, P. Arzberger, T. W. Tan, R. Sinnott, D. Angulo, pp 339-354, Vol. 25 No. 4, 2007.
[6] R.O. Sinnott, O. Ajayi, A.J. Stell, Supporting Grid Based Clinical Trials in Scotland, Health Informatics Journal Special Issue on Integrated Health Records, Vol. 14 (2), June 2008.
[7] R. O. Sinnott, M. M. Bayer, J. Koetsier, A. J. Stell, Advanced Security on Grid-Enabled Biomedical Services, Proceedings of UK e-Science All Hands Meeting, September 2005, Nottingham, England.
[8] R. O. Sinnott, M. M. Bayer, J. Koetsier, A. J. Stell, Grid Infrastructures for Secure Access to and Use of Bioinformatics Data: Experiences from the BRIDGES Project, 1st International Conference on Availability, Reliability and Security, (ARES'06), Vienna, Austria, April, 2006.

[9] R.O. Sinnott, J. Watt, O. Ajayi, J. Jiang, Shibboleth-based Access to and Usage of Grid Resources, IEEE International Conference on Grid Computing, Barcelona, Spain, September 2006.
[10] R. Housley, T. Polk, *Planning for PKI: Best Practices Guide for Deploying Public Key Infrastructures*, Wiley Publishing, 2001.
[11] Sandhu R.S., Coyne E.J., Feinstein H.L., Youman C.E., "Role-Based Access Control Models". IEEE Computer. 1996; 29:38-47.
[12] D.W.Chadwick, A. Otenko, E.Ball, Role-based Access Control with X.509 Attribute Certificates, IEEE Internet Computing, March-April 2003, pp. 62-69.
[13] Alfieri R, et al. VOMS: an authorization system for virtual organizations, 1st European across Grids conference, Santiago de Compostela.
[14] IBAC
[15] PBAC
[16] National Program for IT in the NHS (NPFIT) - http://www.connectingforhealth.nhs.uk
[17] A.J. Stell, R.O. Sinnott, J. Jiang, I. Piper, R. Donald, Federating Distributed Clinical Data for the Prediction of Adverse Hypotensive Events, to appear in Journal of the Philosophical Transactions of the Royal Society A, January 2009.
[18] R.O. Sinnott, O. Ajayi, A.J. Stell, Grid Infrastructures Supporting Paediatric Endocrinology across Europe, UK e-Science All Hands Meeting, Nottingham, UK, September 2007.
[19] R.O. Sinnott, A.J. Stell, O. Ajayi, Development of Grid Frameworks for Clinical Trials and Epidemiological Studies, HealthGrid 2006 conference, Valencia, Spain, June 2006.