

# Development of Grid Frameworks for Clinical Trials and Epidemiological Studies

Richard SINNOTT, Anthony STELL, Oluwafemi AJAYI  
*National e-Science Centre, University of Glasgow, United Kingdom*

**Abstract.** E-Health initiatives such as electronic clinical trials and epidemiological studies require access to and usage of a range of both clinical and other data sets. Such data sets are typically only available over many heterogeneous domains where a plethora of often legacy based or in-house/bespoke IT solutions exist. Considerable efforts and investments are being made across the UK to upgrade the IT infrastructures across the National Health Service (NHS) such as the National Program for IT in the NHS (NPFIT) [1]. However, it is the case that currently independent and largely non-interoperable IT solutions exist across hospitals, trusts, disease registries and GP practices – this includes security as well as more general compute and data infrastructures. Grid technology allows issues of distribution and heterogeneity to be overcome, however the clinical trials domain places special demands on security and data which hitherto the Grid community have not satisfactorily addressed. These challenges are often common across many studies and trials hence the development of a re-usable framework for creation and subsequent management of such infrastructures is highly desirable. In this paper we present the challenges in developing such a framework and outline initial scenarios and prototypes developed within the MRC funded Virtual Organisations for Trials and Epidemiological Studies (VOTES) project [2].

## 1. Introduction

Clinical trials allow for the large-scale assessment of the moderate effects of treatment on various diseases and conditions. Typically the various stages of a trial involve identifying willing participants, evaluating their eligibility for the study, obtaining their consent, beginning the course of treatment and undertaking follow-up study both during and potentially long after the treatment has completed. Statistical analysis of the impact of the trials, e.g. on the efficacy of the drugs being tested can then be undertaken. The large-scale processes involved in this can be broadly broken down into three areas: patient recruitment; data management, and study administration and co-ordination.

Until recently it was the case that clinical trials and epidemiological studies would be human intensive and paper based. Examples include, the West of Scotland Coronary Prevention Scheme (WOSCOPS) study [3] conducted at the University of Glasgow, where over 20,000 letters were sent out to eventually recruit 6595 middle-aged men (age 45-64) with a mean cholesterol of 7.0 +/- 0.6mmol. On a much larger scale the UK BioBank effort [4] will be sending many millions of letters to potential trial participants in the hope of recruiting 500,000 members of the population between 40-69 years of age. Not only are these expensive solutions, they are also highly inefficient and human intensive often with members of the population being contacted that do not meet the appropriate constraints for the given trial, e.g. their cholesterol is too high or too low, or they are on other drug treatments etc. E-health initiatives are now moving towards electronic based clinical trials which in principle offer solutions to improve how trials are set up and subsequently managed. However, establishing an electronic trial is not without its own challenges. Each individual trial will face the same kinds of challenges for recruitment, data management and study co-ordination, hence a framework supporting a multitude of trials would be extremely beneficial and is something currently being explored within the MRC funded VOTES project [2].

To establish an e-Infrastructure for clinical trials requires addressing heterogeneity and distribution of systems and data sets, and differences in general practices, e.g. how data is backed up (or not) at given sites. One of the key challenges from an IT perspective is security. The “weakest link” adage applies to security and a single site that does not take appropriate security considerations, both in terms of the technologies they have used, how they are using them and their general practices, can in principle jeopardise the security of all collaborating sites [5]. The risk of data disclosure is an ever present security risk that cannot be ignored. Ensuring that Caldicott guardians and other independent senior health professionals with strategic roles for the management of the data protection or confidentiality associated with patient data sets are involved in the decisions that influence the development of such infrastructures is crucial to their success; from their development, their acceptance, and perhaps more importantly their ethical usage.

It could be argued that the immediate hurdle in establishing an electronic clinical trial is how to recruit people. Key sources of data in Scotland include national census data sets such as the General Register Office for Scotland [6] which includes information such as the registration of births, marriages, deaths as well as being the main sources of family history records. The access to such information whilst useful does not include direct health related information which will likely impact upon the suitability of patients to a trial. Primary care and secondary health care data sets are other immediate choices, however access to and usage of these data sets will likely require ethical approval. Patients should have the opportunity to consent that their data can be accessed and used. However in running a clinical trial, it is often the case that statistical information is enough. Thus rather than disclosing information on specific patients, statistical information is sufficient. Even here however, questions on ethics are raised. At the very least, doctors and their patients need to be included in any data access decisions.

Yet the establishment and running of electronic clinical trials is a compelling one with data often being stored in some form of digital format, albeit across a multitude of databases behind firewalls. One of the key challenges is to allow secure access to these data sets to the right people for the right purpose. High levels of security should not be at the cost of usability. A good example of this is the remote control car key - a far improved and more complex technologically, security solution, but easier to access and use. Similarly, end users of e-Infrastructures should be largely unaware of the fine grained security solutions that are restricting and controlling their access and usage of the facilities. Usability of the infrastructures is of uppermost importance to their success and take-up [7].

In this paper we describe our attempts to establish and support a Grid framework at the National e-Science Centre (NeSC) in Glasgow as part of the initial phase of the VOTES project. As this work is in the early stages, the solution presented is necessarily grounded in this specific use-case but is conducted with a view to scaling up and generalising as the project proceeds. Through this framework we expect to support the efficient establishment and subsequent conduct of clinical trials and studies. In the rest of this paper we present the technical and non-technical challenges facing the design and development of this framework, along with an outline of the early proof of concept prototypes currently supported. We also outline the future work of the project and challenges, still to be addressed to realise the vision of an e-Infrastructure for a range of clinical trials and studies.

## **2. Existing Infrastructures and Data Sets across Scotland**

The VOTES project [2] is a collaborative effort between e-Science, clinical and ethical research centres across the UK including the universities of Oxford, Glasgow, Imperial, Nottingham and Leicester. The primary focus of VOTES is to build an infrastructure to support a multitude of clinical virtual organisations. Virtual organisations (VOs) are a common concept in the Grid community and provide a conceptual framework through which the rules associated with the participants, their roles and the resources to be shared can be agreed and subsequently enforced across the Grid. VOs in the clinical trials domain are characterised by a much greater degree of emphasis on security, data access and data ownership. We term these Clinical Virtual Organisations (CVOs) since they place requirements not typical to other High Performance Computing-oriented VOs common to the wider Grid community. Rather than developing bespoke CVOs for each individual clinical trial, it is our intention to develop a framework supporting a multitude of CVOs. Each of these CVOs will be derived from the framework and adapted depending on the needs of the trial or study being conducted.

Common phases of many clinical trials and epidemiological studies, and the primary focus for core components that will exist in the VOTES Grid framework are:

- Patient recruitment enabling semi-automated large-scale recruitment methods for investigators conducting large-scale clinical studies in a variety of settings;
- Data collection incorporating data entry including intermittent connectivity to other resources, such as a trial-specific databases, code lists for adverse events and non-study drugs, randomization programs and support for internationalisation of case report forms;
- Study administration supporting the administration of the study, including logging details of essential documents, enabling rapid dissemination of study documentation and by co-ordinating transport of study treatment and collection of study samples.

The first step in developing a Grid framework for clinical trials is to identify the potential sources of data and services that allow access to such data. Close liaison with data providers, data owners and existing services is essential. Within the Scottish element of VOTES we are working closely with the

NHS in Scotland who have identified the following data sets and software which provide initial coverage of the sets of data needed for clinical trials and epidemiological studies<sup>1</sup>:

- The General Practice Administration System for Scotland (GPASS) [8] is the core IT application used by over 85% of clinicians and general practitioners involved in primary care across Scotland;
- Scottish Morbidity Records (SMR) [9] includes records relating to all patients discharged from non-psychiatric and non-obstetric wards in Scottish hospitals (including datasets on death, cancer, hospital admissions, etc.)
- Scottish Care Information Store (SCI Store) [10] - a batch storage system which allows hospitals to add a variety of information to be shared across the community, e.g. pathology, radiology, biochemistry lab results are just some of the data that are supported by SCI Store. Regular updates to SCI Store are provided by the commercial supplier using a web services interface. Currently there are 15 different SCI Stores across Scotland (with 3 across the Strathclyde region alone). Each of these SCI Store versions has their own data models (and schemas) based upon the regional hospital systems they are supporting. The schemas and software itself are still undergoing development.
- NHS data dictionary [11] - a one-stop shop for health and social care data definitions and standards. It contains a summary of concepts for SMR datasets including online manuals for the datasets; information on the clinical datasets in use in healthcare and social care datasets along with the data standards upon which they are based.

The Scottish component of the Grid framework under development within VOTES is being targeted to these resources. Components which allow secure and ethical access to GPASS for example will provide a highly generic reusable solution applicable to over 85% of all practices across Scotland. Contemporaneously, solutions accessing NHS resources are also being developed by the other partners.

A summary of the challenges involved includes broadly: the need for a common definition of clinical standards, the need to maintain security whilst still taking advantage of the flexibility of Grid solutions, the need for scalability, authorization and anonymisation. The following sections address these challenges in more detail.

### **3. Data Federation and Distributed Security Challenges**

As CVOs necessarily span heterogeneous domains, a pre-requisite to the construction of distributed queries and aggregation or joining of data returned is the development and use of a standard method of classification or common vocabulary more generally. This includes the naming of the data sets themselves, the people involved and their roles (privileges) in the access to and usage of these data sets amongst other things. Ideally these data and roles should be standardised so that comparisons can be drawn and queries joined together for example across a range of clinical data sets.

There are numerous developments in standards for the description of data sets used in the clinical trials domain. However, this can be an involved process depending on standards groups developing and acting on strategies put together through major initiatives such as Health-Level 7 (HL7) [12], SNOMED-CT [13] and OpenEHR (Open Electronic Health Records)[14]. There are often a wide range of legacy data sets and naming conventions which impact upon standardisation processes and their acceptance. The International Statistical Classification of Disease and Related Health Problems version 10 (ICD-10) [15] is used for the recording of diseases and health related problems and is supported by the World Health Organisation. In Scotland, ICD-10 is used within the NHS along with ICD version 9 and Read codes in the SMR data sets for example. ICD-10 was introduced in 1993, but the ICD classifications themselves have evolved since the 17<sup>th</sup> century [16].

An explicit example of the problems facing large scale (international) clinical trials is the term “neoplasia” which means “new growth for benign/malignant tumours” in Northern Europe but “cancer” in Southern Europe. Hence, the type of treatment provided depends heavily on the location of the patient. Global Grid frameworks that incorporate appropriate meta-data identifying the different local data classifications can provide capabilities to address such discrepancies.

The standardisation process itself may influence how readily any given standard is adopted. For example, standards developed to specific deadlines during the standardisation-making process, and standards bodies producing regular updates with solutions readily available for implementation are

---

<sup>1</sup> This does not imply that this data is readily available directly, but that these are the sources of data and software which we should be eventually interfacing with.

more likely to gain acceptance. This is also the case within the Grid community. Linking standardised data descriptions between domains so that entities and relationships within one organisational hierarchy can be mapped or understood within the context of another domain is fundamental to the development of the Grid applications proposed in VOTES. Once it has been established how meaningful comparisons can be made between the schemata of differing domains, this knowledge can be applied to a generic clinical trial that could run queries across heterogeneous domains, bringing back generic results, richer in scope and information than if single local sites had been independently queried.

Information stored in clinical trials is by its nature, highly sensitive – drug treatments, conditions and diseases that patients have must be kept in the strictest confidence and the exact details should only be known about by a few privileged roles in the trial. This is one of the most fundamental challenges in this work – to realise the opportunities and benefits that can be brought to this field by Grid technology but to also maintain the high security standards that must be strictly adhered to.

Within the Grid community VO security issues are generally grouped into the categories of:

- *Authentication* – the discovery of a user’s identity. This is achieved in most Grid applications by the use of the well-established Public Key Infrastructure (PKI) technology [17].
- *Authorization* – the discovery of that user’s privileges based on their identity. This is less well-established in the Grid community. Various software solutions are available for the establishment of user privilege assertions – PERMIS [18] (which implements the Global Grid Forum Authz API [19]), Community Authorization Service (CAS) [20], Virtual Organisations Management Service (VOMS) [21], Akenti [22] – with no single model having been adopted over the others.
- *Accounting* – logging the activity of users so that they can be held accountable for their actions within a system. This is also less well-established with many implementations coming from “home-grown” solutions within different projects. Though important in an overall security strategy, this area is usually addressed once the solid platform of authentication and authorization has been established.

Authentication in the Grid is achieved using PKI technology. This involves using a combination of public certificates and public and private keys to verify that a user is who they say they are. This is a well-established way of establishing user identity however it has limitations as a standalone security solution in terms of general usability, security granularity and overall scalability [23,24].

A more scalable, user-oriented solution which is being explored within the VOTES project is the Internet2 Shibboleth technology [25]. Shibboleth allows the delegation of authentication to the local sites involved. Through agreed federations where security attributes for fine grained authorisation are pre-agreed, the users are able to access and use remote Grid resources through local (home) authentication [26,27]. Typically they will log in with their own usernames/passwords at their home institution and the security attributes (which might include their roles in particular clinical trials for example) are then released and used by the target site to determine whether access to the resources being requested should be granted. As well as supporting seamless single sign-on to Grid infrastructures, this model moves the whole process of identity establishment and authentication to the home site. It also minimises the potential dangers of users writing down their PKI passwords and transparently restricts what they are able to do on the remote Grid resources. In the clinical trial domain, it is paramount that site autonomy is supported. If the home site at which a user authenticates themselves does not release all necessary attributes as agreed within the federation, then the user will not be allowed access to and usage of the remote resource. We note that the Shibboleth model is inherently more static than the true dynamic vision of the Grid where data and resources are found and used “on-the-fly”. This static oriented model is consistent with the clinical domain however where it is highly unlikely that new people, new data sets or new services are continually, dynamically added or removed from the clinical environment.

The issue in Grid security that is much less well-established than authentication is that of privilege management – what a user can actually do once their identity has been verified. The main issue is that of the heterogeneous nature of the domains across which the data is being federated. Security policies will naturally differ between local sites, which leads to several challenges when defining and implementing policies that take account of both local and remote security concerns. These include:

- Applying a generic policy that takes into account of each local policy or linking local policies together using a standard interface.
- Dynamically enforcing these policies so that, for example, restrictions applied by a site not providing pertinent information for a particular query will not impact on the sites that are involved.

- Building a trust chain that allows local sites to authenticate to the VO and therefore, by proxy, be authenticated to limited resources at other sites without compromising protected resources at those other sites.
- Prevention of inference (statistical disclosure) that arises when data is aggregated from numerous sources.
- Maintaining data ownership and enforcing ownership policies regardless of where the data might be moved to or stored or used.

In addition to authentication and authorization, another artefact of security that is essential in this domain is that of “anonymisation”. This process involves allowing less-privileged users to gather statistical data for the purposes of studies or trials, but without revealing the associated identifying data – this only being available to users with greater privileges.

The NHS in Scotland currently achieves this by encrypting a unique number associated with all patients across Scotland: the Community Health Index (CHI) number. Once an anonymised patient has been matched for a clinical trial, this encrypted value can in principle be sent to the Practitioners Service group (<http://www.psd.scot.nhs.uk/>) of the NHS who will as one of the many services that they provide, decrypt it and contact the patients directly (assuming ethical permission has been granted for so doing) to ask if they wish to join the clinical trial. Several challenges must be overcome to support this including ensuring that only privileged users are able access and use data sets including this encrypted CHI number. A further challenge is that there are currently many independent solutions across the NHS for how they manage their infrastructures. Thus for example, there is no standardised way in which encryption is undertaken. Hence it is often difficult or impossible to ask Practitioners Services Division (PSD) to de-anonymise an encrypted CHI number if it is generated by arbitrary NHS trusts. Pragmatic solutions overcoming the nuances of NHS systems are thus necessary.

Throughout the VOTES project, continuous ethical and legal overview of the solutions being put forward and the data sets being accessed are being made. This includes the perceived benefits of the research for the public, and is undertaken by independent ethical oversight committees. To support this, superior security roles for oversight committee members which allow access to all data sets and reports for given clinical trials will be made available.

#### 4. Initial VOTES Scenarios, Architecture and Implementation

In designing a reusable Grid framework for clinical trials immediate restrictions are imposed on the possible architectural solutions. Thus it is unlikely that *direct* access to and usage of “live” NHS data sets and resources will be achieved, where *direct* here implies that the Grid infrastructure can issue queries to a remote NHS controlled resource containing un-anonymised patient information, i.e. to a resources behind the NHS firewall. Nevertheless, it is possible to design solutions capturing sufficient information needed for a clinical trial without over-riding existing security solutions or assuming ethical permissions where none have been granted. Possible solutions being explored here include a push model (where anonymised NHS data sets are exported) to the academic Grid community (or to an NHS server in a demilitarised zone of the NHS). Another model is to allow the GPs and clinicians to drive the recruitment process, provided they consider that this is in the best interests of the patients. The exploration of these solutions may provide a basis for follow-up projects in this field.

The following scenario presents a representative sequence of interactions demonstrating how primary care identification and recruitment of patients can be *ethically* achieved with patient and doctor consent. The scenario in Figure 1 is based on discussions with Scottish clinicians, NHS IT personnel and GPASS developers and is currently being prototyped in VOTES.

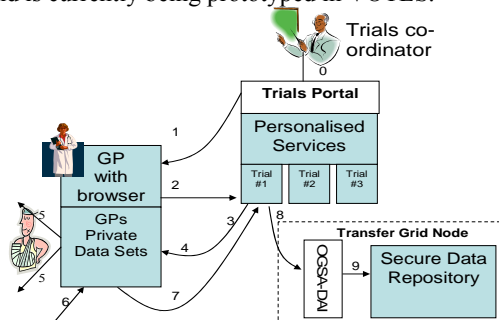


Figure 1: Example use of patient recruitment Grid application

0. A trials coordinator logs into a portal hosting various CVOs associated with a variety of clinical trials<sup>2</sup>. At this point, a personalised environment is established based upon the specific role (in this case, that of the trials coordinator) in the CVO and the location from where they are accessing the portal. Thus they should only see the Grid services pertinent to the appropriate trial applicable to them, and hence the data sets associated with those services.

1. The trial coordinator wishes to recruit patients for a particular trial. These patient details are only available in GPs local (and secure) databases – extensions to this scenario dealing with access to and usage of hospital databases are also possible. Emails are sent to the GPs/hospitals with information describing the particular trial to be conducted, the general criteria applicable to matching patients and other information, e.g. financial information about partaking in the trial. The email contains a link to a Grid service (trial #1). The GPs themselves are described in policies associated with the tentative set up of a CVO for patient identification and recruitment.

2. We assume that the GP is interested in entering into the trial, i.e. they know that they have matching patients and they follow the attached link. Depending upon whether a PKI has been rolled out to this GP and a suitable certificate (e.g. using the X509 standard) is already in the browser or a username and password combination is used instead, the GP securely accesses the Grid service. In this scenario we assume trusted certificates are being used.

3. After extracting more information about the trial from the portal, the GP decides to download a signed XML pro-forma pre-designed for this specific trial. This is a *mostly* complete document describing the main information relevant to this trial as documented in the trial protocol, where the empty fields need to be filled through a query to the GPs database.

4. The signature of the signed pro-forma document is checked to ensure its authenticity and that it has not been corrupted. If these are both true, the document is used as the basis for an XML query against the GP's database (GPASS supports such an interface). This query might in turn result in further information being extracted from other resources.

5. At this point, letters describing the trial to matching patients can be automatically produced. These are used to obtain patient consent before continuing further with the trial.

6. The matching patients may then consent to entering into the trial. Note that these letters of consent may be sent directly to the trial coordinator instead of the GP as depicted here.

7. The forms are automatically completed based on the results of the queries to the GP database, digitally signed and returned to the Grid service for that particular trial (trial #1).

8. The returned signed XML document is authenticated and checks on the sender (the GP) being authorised to upload this document are made, e.g. through checking that they were one of the GPs contacted initially. The document is validated to ensure its correctness, e.g. by ensuring it satisfies the associated schema and the relevant data fields are meaningfully completed (and match the desired constraints associated with participation in the trial). At this point, the responding GP is formally added to the CVO. Further follow up information may subsequently be sought, e.g. monitoring information related to the matching patients.

9. The completed XML document and the associated meta-data describing the history of how this information was established, by whom, when, for which trial etc are uploaded and securely added to the CVO repository for this particular trial.

It is important to note in this scenario that patient consent is given (step 6) before patient data is returned to the clinical trials team. Another important aspect here is that the GP can decide whether this might be in the patients' interest. The patient may ultimately say no and hence is always involved in the process. We note also that software solutions also exist for several parts of this scenario, e.g. automatic production of letters inviting patients to join the trial. Similar scenarios covering user-resource interactions are being developed and implemented within VOTES supporting secondary care patient recruitment as well as for general data collection and study management.

In this scenario we include a secure repository accessible via the Open Grid Service Architecture Data Access and Integration (OGSA-DAI) middleware [28]. This repository forms part of what we term the "Transfer Grid" as indicated in Figure 2. The Transfer Grid infrastructure provides the core of the Grid infrastructure that will underpin future CVOs, i.e. it is the platform, upon which the Grid solutions developed for security, data access and management, and data movement between repositories hosted at the partner and collaborating institutions can be supported. Since the Transfer Grid exists in the academic domain and not behind the NHS firewall, a variety of solutions for accessing and using the clinical trial data sets can be explored. The Grid applications pertinent to the

---

<sup>2</sup> Of course there are scenarios which predate this one, e.g. how CVO is established in the first instance and the policies by which the VO will be organised, managed, enforced.

clinical trials domain are constructed over this layer providing the deliverable trial services. This infrastructure will be expanded to include external peer sites of two classes:

- Routine repositories such as those held by general practices, hospitals, disease-specific registries, device registries or the Office for National Statistics (ONS).
- Study repositories such as research systems developed for a particular trial or observational study.

These external peers will supply their own security policies, and may be intermittently connected to the Transfer Grid. As such, interfacing with routine repositories will be a highly involved and politically sensitive process. This motivates the need for the initial solution to be scalable.

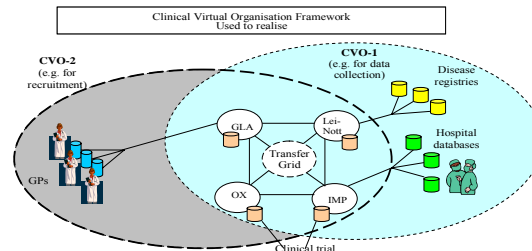


Figure 2: CVO Framework, Transfer Grid and Key Sources of Data

#### 4.1 Current Software Architecture

The basic architecture of this Grid framework, which supports federated queries in a user oriented but secure manner, is depicted in Figure 3. This infrastructure corresponds to one node of the Transfer Grid outlined above and is hosted on a trial test bed at the National e-Science Centre (NeSC) at the University of Glasgow.

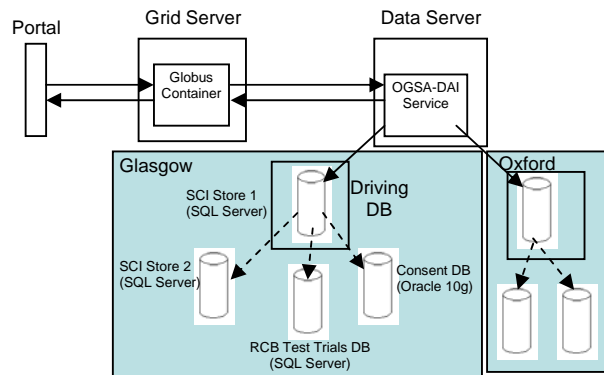


Figure 3: Software architecture schematic. The “Oxford” box indicates how other institutions will be added to the current design – the current implementation only incorporates the test databases running in Glasgow.

A GridSphere [29] portal front-end communicates to a Globus Toolkit [30] (v4.0) grid service, which in turn provides access to an OGSA-DAI [28] data service. This runs queries from the “driving database” using standard Simple Object Access Protocol (SOAP) message-passing, but also in turn runs queries from the subsidiary databases available from the pool for which it is responsible, using direct Java Database Connectivity (JDBC) connections.

The technology used in this implementation places strong emphasis on the use of grid services – essentially web services with the additional notion of permanent state. Within the Grid community this paradigm has been largely seen as the most effective solution to implementing transient and dynamic virtual organisations. An example of this is the Web Services Resource Framework (WS-RF) [31] as implemented in version 4.0 of the Globus Toolkit. Issues of access control are integrated within this framework by means of a Security Assertion Markup Language (SAML), which allows a standard exchange of security assertions and attributes. A popular implementation of this standard has been the OpenSAML project [32], which is now following the latest release of SAML, v1.1, and is currently developing an implementation of v2.0 [33].

The user accesses this infrastructure through a Gridsphere portal at [2]. With the appropriate privileges, users can currently bring back data from the database back-ends implemented in multiple

test repositories of SCI Store and GPASS. Unprivileged users can retrieve limited data-sets, with the identifying patient data anonymised and other restrictions applied. Through the use of this application, the end user is able to seamlessly access a set of resources, pertinent to clinical trials, in a dynamic, secure and pervasive fashion. Depending on the user's privileges, the results returned have varying degrees of verbosity thereby allowing limited statistical analysis without compromising the privacy restrictions necessarily applied in such sensitive data.

In the current version of the system to explore the problem space and gain familiarity with the clinical data sets used across Scotland, several "canned queries" representing valid clinical trial queries can be run which seamlessly access and use distributed back-end test databases as depicted in Figure 4.

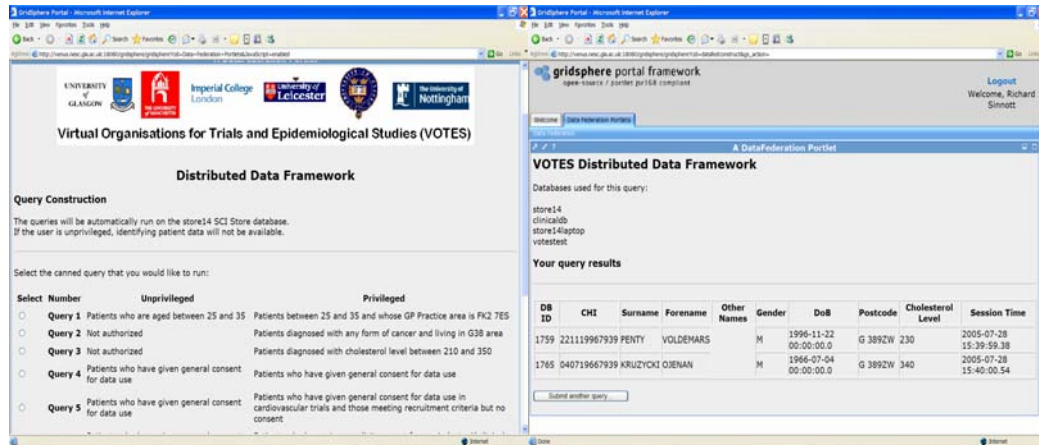


Figure 4: Screen-shot of VOTES portal welcome screen (left) showing several "canned queries" with the type of result returned based on whether the user is privileged or not (right).

Users with insufficient privileges may still be able to run queries but may not be able to see all of the associated identifying data sets (see Figure 5). It is important to note that all of this is completely transparent to the end users of the system.

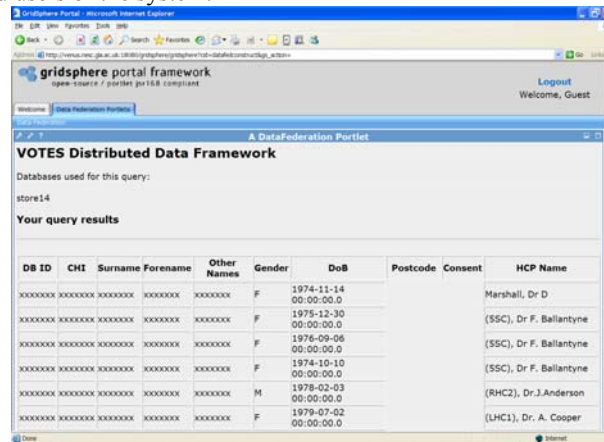


Figure 5: Results from an unprivileged user running a canned query. Identifying data is blanked out whilst statistically relevant data is available. Also the number of databases across which the query has been run is reduced because of lack of privileges.

Another key aspect of this infrastructure is how patient consent is handled. Currently the system supports a variety of models which are allowing exploration of the potential solution space for patient consent across Scotland. For example solutions have been prototyped which allow patients to consent to their data being used for a specific clinical trial, for a particular disease area or consent for their data being used generally. In addition, the system also allows for patients to opt out, i.e. their data sets may not be used for any purposes. Numerous variations on this are also being explored, e.g. the patients' data may only be used provided they are contacted in advance. To support this, a consent database has been established and is used when joining of the federated queries is undertaken to decide whether the data should be displayed, displayed but anonymised, or not displayed at all.

The NeSC at Glasgow have extensive experiences in a range of fine grained authorisation infrastructures across a range of application domains [34-36]. Whilst we expect to move the existing

prototype to a more robust authorisation solution, for rapid prototyping purposes to explore the problem space and get user feedback as early as possible, we have developed an authorization infrastructure based on an access matrix as shown in Figure 6.

	R <sub>1</sub>	R <sub>2</sub>	R <sub>3</sub>	R <sub>4</sub>
U <sub>1</sub>	h <sub>1</sub>	h <sub>2</sub>	h <sub>3</sub>	h <sub>4</sub>
U <sub>2</sub>	0	0	1	0
U <sub>3</sub>	0	0	0	1
U <sub>4</sub>	1	1	1	1
U <sub>5</sub>	0	1	0	0

$$U_1(R_1 \Delta h_3) = 1 \quad U_2(R_1 \Delta h_2) = 0 \quad U_3(R_3 \Delta h_1) = 1 \quad U_4(R_2 \Delta R_3 \Delta h_4) = 0$$

where  $\Delta$  is a combination function, 0, 1 are bit-wise privileges,  $R_x$ ,  $h_x$  are resources and  $U_x$  is a subject

Figure 6: Access Matrix Model

The authorisation mechanism implements an access matrix model [37] that specifies bit-wise privileges of users and their associations to data objects in the CVO. The access matrix is designed to enforce discretionary and role based access control policies and has been constructed to be scalable for ease of growth parallel to the growth of the infrastructure as a whole. Comparison of this approach with other solutions such as Role Based Access Control solutions such as PERMIS will be undertaken, where user views of data sets will be mapped to CVO roles.

The federated data system [38] is currently composed of four autonomous test sites, each providing a clinical data source using either SQL Server [39] or Oracle [40]. The data sources exposed by these sites are configured as data resources on an OGSA-DAI data service. The OGSA-DAI data service implements a head node model to drive the data federation. The head node is selected based on rules or request requirements and is responsible for decomposing queries, distributing sub-queries and gathering and joining query results.

In the current implementation, data federation security is achieved at both local and remote level. The local level security, managed by each test site, filters and validates requests based on local policies at Database Management System (DBMS) levels. The remote level security is achieved by the exchange of access tokens between the designated Source of Authority (SOA) of each site. These access tokens are used to establish remote database connections between the sites in the federation. In principle local sites authorise their users based on delegated remote policies. This is along the lines of the CAS model [20].

## 5. Conclusions and Future work

The VOTES prototype software is very much a work in progress. Yet the experiences in developing this prototype are helping to gain a better understanding of the clinical domain problem space and shaping the planned Grid framework. The vision of a Grid framework eventually supporting a myriad of clinical trials and epidemiological studies is a compelling one, but can only be achieved once experiences have been gained in accessing and using a wide variety of clinical data sets. In achieving this, it is immediately apparent that there are a number of political and ethical issues that must be addressed when dealing with data-sharing between domains and these are inherently more difficult to deal with than the technological challenges. Whilst the NHS in Scotland and the UK more widely are taking steps to standardise the data-sets that they have, these are still far from being fully implemented (and accepted) by clinical practitioners. For instance, the unique index reference number the Community Health Index (CHI) has only been implemented across some regions of Scotland and therefore leaves certain areas with incomplete references. Those records that do not have the CHI number are referenced using a different Patient Identification (PID) number that will be idiosyncratic to the region in question. There is also a need to build up a trust relationship with the end-user institutions that we are working with to provide this clinical infrastructure. This necessarily takes time and will be furthered by engaging in an exchange program where employees from NeSC work with and understand the processes in the NHS IT departments and vice-versa.

The current Grid infrastructure described here has allowed the investigation of automatically implementing combinations of patient consent policies. Ideally such a consent register would be maintained nationally, however this does not exist yet but is planned with the electronic patient record under discussions across the NHS in Scotland. Demonstrations of working solutions showing the trade-offs in consent or assent with opt in versus opt out possibilities allows the policy makers to see first hand what the impact of their ultimate decisions might have. We believe that it is easier to convince

policy makers when they see actual working solutions rather than theoretical discussions of what might be achieved once the infrastructures are in place.

The applications in this project are being developed with a view to being rolled out to the NHS Scotland in the first instance, moving from test data to “live” data with fully audited and standards-compliant security, upon establishment of reliability and production value. The eventual vision is that this infrastructure will one day be available on a global scale allowing health information to be exchanged across heterogeneous domains in a seamless, robust and secure manner. In this regard, we are currently exploring international collaborative possibilities with the caBIG project in the US [41] and closer to home in genetics and healthcare projects across Scotland [42].

## 6. References

- [1] National Program for IT in the NHS (NPFIT) - <http://www.connectingforhealth.nhs.uk>
- [2] Virtual Organisations for Trials and Epidemiological Studies (VOTES) - <http://www.nesc.ac.uk/hub/projects/votes/>
- [3] West Of Scotland Coronary Prevention Scheme (WOSCOPS)  
<http://www.gla.ac.uk/departments/pathologicalbiochemistry/lipids/woscops.html>
- [4] UK BioBank project - <http://www.ukbiobank.ac.uk>
- [5] R. O. Sinnott, Grid Security: Middleware, Practices and Outlook, prepared for the Joint Information Services Council (JISC), [www.nesc.ac.uk/hub/projects/GridSecurityReport](http://www.nesc.ac.uk/hub/projects/GridSecurityReport)
- [6] General Register Office for Scotland, <http://www.gro-scotland.gov.uk/>
- [7] R.O. Sinnott, Development of Usable Grid Services for the Biomedical Community, Workshop on Designing for Usability in e-Science, Edinburgh, January 2006, <http://www.nesc.ac.uk/action/esi/contribution.cfm?Title=613>.
- [8] General Practitioners Administration System for Scotland (GPASS), <http://www.show.scot.nhs.uk/gpass/>
- [9] Scottish Morbidity Records (SMR), <http://www.show.scot.nhs.uk/indicators/SMR/Main.htm>
- [10] Scottish Care Information (SCI) Store - [http://www.show.scot.nhs.uk/sci/products/store/SCIStore\\_Product\\_Description.htm](http://www.show.scot.nhs.uk/sci/products/store/SCIStore_Product_Description.htm)
- [11] NHS Data Dictionary – [www.isdscotland.org](http://www.isdscotland.org)
- [12] Health-Level 7 (HL7) - <http://www.hl7.org/>
- [13] SNOMED-CT - <http://www.snomed.org/snomedct/>
- [14] OpenEHR - <http://www.openehr.org/>
- [15] International Statistical Classification of Disease and Related Health Problems (ICD-10),  
[http://www.connectingforhealth.nhs.uk/clinicalcoding/classifications/icd\\_10](http://www.connectingforhealth.nhs.uk/clinicalcoding/classifications/icd_10)
- [16] ICD background, <http://www.connectingforhealth.nhs.uk/clinicalcoding/faqs/>
- [17] R. Housley, T. Polk, Planning for PKI: Best Practices Guide for Deploying Public Key Infrastructures, Wiley Computer Publishing, 2001.
- [18] PERMIS - <http://sec.isi.salford.ac.uk/permis/>
- [19] R.O. Sinnott, D.W. Chadwick, Experiences of Using the GGF SAML AuthZ Interface, Proceedings of UK e-Science All Hands Meeting, September 2004, Nottingham, England.
- [20] CAS - <http://www.globus.org/toolkit/docs/4.0/security/cas/>
- [21] VOMS - <http://hep-project-grid-scg.web.cern.ch/hep-project-grid-scg/voms.html>
- [22] Akenti - <http://dsd.lbl.gov/Akenti/>
- [23] R.O. Sinnott, A.J. Stell, D.W. Chadwick, O.Otenko, Experiences of Applying Advanced Grid Authorisation Infrastructures, Proceedings of European Grid Conference (EGC), LNCS 3470, pages 265-275, Volume editors: P.M.A. Sloot, A.G. Hoekstra, T. Priol, A. Reinefeld, M. Bubak, June 2005, Amsterdam, Holland.
- [24] A.J. Stell, R.O. Sinnott, J. Watt, Comparison of Advanced Authorisation Infrastructures for Grid Computing, Proceedings of International Conference on High Performance Computing Systems and Applications, May 2005, Guelph, Canada.
- [25] Shibboleth Project - <http://shibboleth.internet2.edu/>
- [26] R.O. Sinnott, J. Watt, O. Ajayi, J. Jiang, J. Koetsier, A Shibboleth-Protected Privilege Management Infrastructure for e-Science Education, submitted to CLAG+Grid Edu Conference, May 2006, Singapore.
- [27] R.O. Sinnott, J. Watt, O. Ajayi, J. Jiang, Shibboleth-based Access to and Usage of Grid Resources, submitted to International Conference on Emerging Trends in Information and Communication Security, Freiburg, Germany, June 2006.
- [28] OGSADAI – <http://www.ogsadai.org.uk>
- [29] GridSphere – <http://www.gridisphere.org>
- [30] Globus Toolkit – <http://www.globus.org/toolkit>
- [31] Web Services Resource Framework (WS-RF) – <http://www.globus.org/wsrf>
- [32] OpenSAML Project – <http://www.opensaml.org>
- [33] OpenSAML Development Wiki - <https://authdev.it.ohio-state.edu/twiki/bin/view/Shibboleth/OpenSAML>
- [34] 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, submitted to 1st International Workshop on Bioinformatics and Security (BIOS'06), Vienna, Austria, April, 2006.
- [35] R.O. Sinnott, M. Bayer, D. Berry, M. Atkinson, M. Ferrier, D. Gilbert, E. Hunt, N. Hanlon, Grid Services Supporting the Usage of Secure Federated, Distributed Biomedical Data, Proceedings of UK e-Science All Hands Meeting, September 2004, Nottingham, England.
- [36] R.O. Sinnott, A.J. Stell, J. Watt, Experiences in Teaching Grid Computing to Advanced Level Students, Proceedings of CLAG+Grid Edu Conference, May 2005, Cardiff, Wales.
- [37] R. S. Sandhu and P. Samarati, “Access control: Principles and practice,” IEEE Communications Magazine, vol. 32, no. 9, pp. 40-48, 1994.
- [38] A. P. Sheth and J. A. Larson, “Federated database systems for managing distributed, heterogeneous, and autonomous databases,” ACM Comput. Surv., vol. 22, no. 3, pp. 183-236, 1990.
- [39] SQL Server – <http://www.microsoft.com/sql>
- [40] Oracle – <http://www.oracle.com>
- [41] National Cancer Institute, cancer Biomedical Informatics Grid, <https://cabig.nci.nih.gov/>
- [42] Generation Scotland Scottish Family Health Study, <http://www.innogen.ac.uk/Research/The-Scottish-Family-Health-Study>