Initial Experiences in Developing e-Health Solutions across Scotland

Prof. R.O. SINNOTT, A.J. STELL, O. AJAYI
National e-Science Centre,
University of Glasgow,
United Kingdom
ros@dcs.gla.ac.uk

Abstract.
The MRC funded Virtual Organisations for Trials and Epidemiological Studies (VOTES) project 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 project started in September 2005 and is due to run for 3 years. The primary goal of VOTES is to develop a reusable Grid framework through which a multitude of clinical trials and epidemiological studies can be supported. The National e-Science Centre (NeSC) at the University of Glasgow are looking at developing the Scottish components of this framework. This paper presents the initial experiences in developing this framework and in accessing and using existing data sets, services and software across the NHS in Scotland.

1. Introduction

Reliable assessment of moderate effects of treatment of important diseases (such as cardiovascular disease and cancer) on major clinical outcomes requires studies that guarantee both strict control of bias (which, in general, requires proper randomization and appropriate analysis) and strict control of random error (which, in general, requires large numbers of relevant outcomes). Generating this evidence typically requires the collection of data on many thousands of people over a period of at least several years. Similarly, observational studies assessing the impact of particular exposures (such as cigarette smoking, industrial chemicals or blood cholesterol) to important clinical outcomes can require large numbers of such events during prolonged follow-up, in order to avoid misleading results caused by the play of chance. The conduct of large-scale randomized trials or observational studies usually requires a collaborative effort, in which data are collected from individuals and from existing health records at multiple investigative sites, and study progress, data quality and analysis of results are managed by one or more coordinating centres.

Grid technology provides one way in which remote, heterogeneous clinical data sets that are often managed by numerous independent bodies, can be seamlessly brought together. A Grid can be defined as a software infrastructure (including computer systems and data storage resources) that enables flexible, secure, co-ordinated resource sharing among dynamic collections of individuals, institutions and resources. For clinical trials and observational studies, the particular attractions of a Grid-based approach are the ability to create, and subsequently manage, *virtual organisations* (VOs). VOs provide frameworks through which the rules associated with the participants and resources are agreed and enforced. This may well involve agreements upon the remote resources themselves (databases and repositories, as well as the data sets contained therein) and the services that they agree to make available to one another. The domain of large-scale clinical studies provides special challenges in the level of granularity associated with the rules, agreements and policies that might be present in a given clinical virtual organisation (CVO). In particular, CVO policies must strongly adhere to local policy constraints, for example, on data sharing or confidentiality.

It is expected that improved procedures for patient identification and recruitment, data collection and study management can be achieved by the development of Grid infrastructure to create such CVOs. For example, subject to strict ethical, data protection and security constraints, information can be shared between databases established for routine clinical use (such as general practice and hospital records, disease-specific registries, and central registries) and databases established especially for a particular trial or observational study. Appropriate access to information from routine clinical systems should enable more efficient recruitment and more complete follow-up of participants in large-scale clinical studies to be achieved more economically. Furthermore, combining up-to-date data on study participants from both routine and research systems, will allow organisers of clinical trials and

observational studies to conduct effective and efficient monitoring of these studies, so that potential problems are identified early. This will be particularly valuable in helping organizations comply with their responsibilities for monitoring under the EU Clinical Trials Directive (2001/20/EC). In addition, the use of a Grid infrastructure should offer significant advantages such as availability, reliability, scalability and efficiency for these purposes, compared with existing systems.

2. Background to VOTES

The Virtual Organisations for Trials and Epidemiological Studies (VOTES) project has been funded by the MRC for three years to establish a re-usable Grid framework which will support three key stages of any clinical trial or observational study: (1) recruitment of potentially eligible participants, (2) data collection, and (3) study management. The framework itself will be comprised of specifically engineered collections of adaptable Grid services whose usage and combinations will depend upon the needs of the particular clinical study to support different *flavours* of CVO. Diagrammatically, the intention of VOTES is depicted in Figure 1 where the framework is used to generate a multitude of different CVOs allowing user-oriented, ethical access to and usage of clinical data sets.

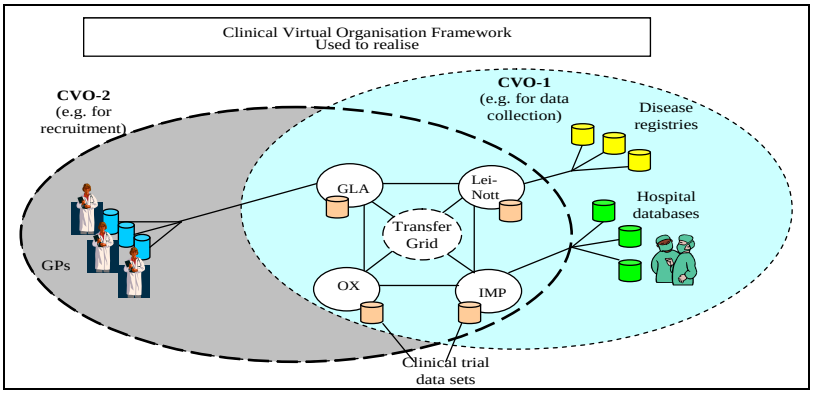

Figure 1: Clinical Virtual Organisations for Clinical Trials and Observational Studies

The *TransferGrid* identified in Figure 1 will offer generic Grid services for security, data access and management, and data movement between repositories (Peers) hosted at the partner and collaborating institutions. The TransferGrid is to be hosted by the collaborating institutions, hence complete control over the technologies used and its' design and architecture is possible. This core Grid infrastructure will then be expanded and refined to develop CVOs that include External Peers of two general classes: (a) Routine Repositories (such as those held by general practices, hospitals, disease-specific registries, device registries, or the Office for National Statistics), and (b) Study Repositories (research systems developed for a particular trial or observational study). External Peers will apply their own security policy, and may be intermittently connected to the TransferGrid. Interfacing with routine repositories will be a highly involved and potentially politically sensitive process, requiring specific project resource. We outline initial experiences in dealing with access to such resources in Scotland section 3.

Critical to the development of such a framework are languages and tools for implementing the rules and regulations (policies) relevant to clinical trials. A typical clinical trial or observational study will contain policies defining collections of entities and the relationships between them including:

- roles – e.g. clinician, administrator, participant, pharmacist, steering committee member, data monitoring committee member, software/Grid engineer, and systems/database administrator;
- level – e.g. central, regional, or national coordinating centre, and local clinical centre;
- computational resources from supercomputers, farms, PCs, laptops, mobile devices;
- specific software infrastructures to be used in the trial (such as analysis software), as well as specific Grid services (such as data access, integration, annotation, movement and replication services, Grid security authentication and authorisation services, and Grid workflows);
- data resources such as clinical databases, disease-specific registries, and study-specific databases, as well as the data sets contained therein.

Policies will specify and qualify permitted relationships between these entities, often at a fine granularity, and will provide the basis for establishing a given CVO (and hence for configuring the associated Grid infrastructure). One way that this can be supported is through XML sub-schemas for standard subsections of policies frequently used in CVOs, referencing existing standards where they are

emerging for particular classes of entity. In turn, these sub-schemas will be referenced by XML Schemas to create "Policy Templates". A given policy instance (for a particular CVO) will then be constructed from a Policy Template with specifics for a required action and the associated security considerations. The final implementation will require both informal testing and documented validation against the policy set out in the user requirement specifications for that CVO Framework.

Usage of the infrastructure developed through such policies will require that all policies are strictly adhered to. An example policy that a CVO framework would be expected to support would be: *for a specific trial, a study clinician working at the national coordinating centre, is allowed to see all trial data (except treatment allocation) for all participants in that country, but only summary data (e.g. a recruitment graph) for participants in other countries.* A query not fulfilling all the terms or agreements defined by the CVO policy will be rejected at a given resource and the information associated with the query (the sender, nature of the request, etc) will be logged to enable this to be followed up, including a review of the implementation of the policy itself.

To support usability concerns, a VOTES portal is under development using the GridSphere portal technology (www.gridsphere.org) which will host the CVOs associated with particular trials. Users accessing this portal will have a predefined role for the specific trials they are involved in. This information will be used to personalise their environment when using the portal, limiting the data sets they may see, and the services they can invoke. Ensuring the security of the location where these policies are hosted is therefore essential.

Each CVO resource (e.g. Peer on the TransferGrid, trial database or clinical repository) will have its own local policies, capturing information about who may run queries against them, and what type of queries or data sets can be extracted from them. For some external resources (i.e. resources not on the TransferGrid), such information might be implicit (e.g. only a GP is allowed to submit a query against their database), and local policy-based access control software might not exist. In such cases, it is necessary to work with the controller of the external resource to define and subsequently enforce policies for querying the associated data sets (either directly or via access to adaptors or local services). For other resources (particularly the core TransferGrid), it will be possible to deploy the Grid infrastructure and thereby use local policies to control access to and usage of resources.

Usage of Grid services often requires appropriate adaptors to facilitate access to, and usage of, specific data resources. Such Grid middleware should provide facilities to access and use a variety of database technologies, matching the heterogeneity of resources in clinical trials. The OGSA-DAI (www.ogsadai.org.uk) project provides mechanisms for managing, accessing and integrating XML relational and file data held in a variety of different databases (DB2, Oracle, MySQL, Xindice, etc) via the Grid. The OGSA-DAI technology is currently being used to support access to, and storage of, the clinical data, as well as the extensive structured metadata needed to organise data associated with clinical repositories. These metadata will typically describe how, when and by whom the data were produced, and will also describe the data structure and other salient features. The precise description of meta-data will facilitate the ability to conveniently find, create, store, access, integrate and subsequently analyse data from heterogeneous sources.

3. Background to Scottish Data Sets, Services and Infrastructure

One of the immediate key challenges that must be addressed in developing this framework is gaining access to appropriate data sets. Key sources of data in Scotland include national census data sets such as the General Register Office for Scotland (http://www.gro-scotland.gov.uk/) 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 requires ethical approval. This is arguably the greatest hurdle that has to be overcome to realize the e-Health vision and allow clinical research to be supported. This should not be orthogonal to patient care however. Rather patients should have the opportunity to consent that their data can be accessed and used. 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 and usage decisions.

The focal point of the primary and secondary care data sets being considered thus far within VOTES are based upon discussions with the NHS Information Services (www.isdscotland.org) and their associated technology providers. These include:

- The General Practice Administration System for Scotland (GPASS) (www.gpass.co.uk) which is the core IT application used by over 85% of clinicians and general practitioners involved in

primary care across Scotland. GPASS is the focus of primary care data sets being considered within VOTES. The VOTES team at NeSC Glasgow have been given a copy of the GPASS software and various training data sets.

- Scottish Morbidity Records (SMR) (http://www.show.scot.nhs.uk/indicators/SMR/Main.htm) which includes good quality (linked) records relating to a variety of patient information records including: patient records discharged from hospital between January 1981 - March 1997 (SMR1); COPPISH discharges from April 1997 onwards (SMR01); historic discharges 1981 – March 1997 (SMR4); COPPISH Admissions April 1996 onwards (SMR04); GRO Death Records January 1980 - December1995; GRO Death Records January 1996 onwards; SOCRATES (Cancer Registrations) 1980 onwards. The VOTES team have currently been given access to the schema descriptions associated with these data sets. It is planned that an anonymised version of the datasets themselves will also be made available in the near future to the NeSC team.
- Scottish Care Information (SCI) Store (http://www.show.scot.nhs.uk/sci/products/store) 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. The VOTES team at NeSC Glasgow have been given a copy of the SCI Store software and various training data sets.

It is the case that these solutions are currently undergoing a process of evolution across the NHS in Scotland. Commercial suppliers are producing software under contract by the NHS, yet it is currently the case that a fabric or overall systems architecture which will be used to integrate all of these software families and their associated data sets has not been fully defined. Thus for example, different sources of data and services use different classification schemes such as International Classification of Disease version 10 (ICD-10) and ICD version 9, as well as older Read coding exists. Similarly, systems such as the SCI Store laboratory systems often use different patient identifiers, hence SCI Store supports locally controlled cross-matching to ensure an incoming piece of information is attached to the correct patient's record. This cross-matching is in turn based on a link to one or more local patient administration systems and the national Community Health Index (CHI) number patient identifier, equivalent to England's National Health Service (NHS) number.

There are different ways of getting access to the information held in SCI Store, e.g. through a secure web browser, through a fixed set basic web services, or through direct access to the back end database (based on SQL Server). It is this latter option which is most apposite to Grid based solutions and is being explored in VOTES since we broad range of access and usage capabilities. Thus it is not known exactly what information might be required for a clinical trial, or what web service interfaces should be supported.

For the end users, the emphasis of repositories such as SCI Store has been on the providing 'look-up' access, e.g. to access laboratory test results. Ideally this information should be provided to clinicians so they are able to see results within minutes of them being analysed, as this reduces the requirement for telephone calls. Look-up access will always have its place, however looking up information in a remote repository has its limitations for example clinical information may not be integrated with the primary care record on a given GP's system. With regard to VOTES, there might be numerous ways that the remote repository data might be integrated with the GPs data however what is clear is that this data should be *ethical* and electronically verifiable. For example so that the data has the appropriate type and is within given predictable constraints, e.g. upper and lower limits for a cholesterol level, or for a blood pressure measurement, or for a patients body mass index level etc. The term ethical is italicised here since it is essential to ensure that software solutions are developed for the right reasons and that ethical control of software IT push is ensured and that clinical researchers requirements are tempered by patient care necessity. Thus for example, software solutions which allow laboratory test results to be automatically incorporated into the primary care record without the GP's knowledge could be envisaged. Such an approach could be considered to be of use however this would be potentially very dangerous and clinically unacceptable since the potential possibility for errors always exists. Hence, interfaces and associated causalities are needed which can facilitate user/clinician driven control which is in the best interest of the patient at all times.

It goes without saying that the success of e-Health infrastructures across Scotland is by no means a technology-only challenge. The hence Electronic Clinical Communication Implementation (ECCI)

(http://www.show.scot.nhs.uk/ecci/) is targeted at funding local, people-focused implementation support as well as development of common standards such as access control protocols and information standards.

4. Integrated Health Records within VOTES

As CVOs necessarily span heterogeneous domains, a pre-requisite to the construction of distributed queries and aggregation or joining of returned data is the development and use of a standard method of classification or common vocabulary more generally. The VOTES team are currently exploring the various data classification schemes and associated data models used across Scotland. Currently the primary focus of VOTES is in understanding the schemas and associated classifications of the GPASS, SMR and SCI Store data sets. Once robust and validated solutions have been engineered for the VOTES TransferGrid for access to GPASS, SCI Store and the SMR data sets, these will be explored in the wider context of a peer grid. A key element in the realisation of the live system is trust: trust of the technology and trust of the people and process. To support this it is planned that the NeSC Glasgow researchers will be given honorary contracts to work at the NHS in Glasgow. Through immersing themselves in the NHS environment, a better understanding of the daily processes for data production, data management and general software IT support will be garnered. It is clear that to have any chance of wide scale uptake, the Grid cannot subsume existing infrastructures and practices. Instead, the Grid and associated solutions must be constrained to function within the context of existing solutions and practices.

One key challenge in this is in understanding the currently used classifications and schemes and models that will have some future longevity. There currently exist a broad range of different standards and data classifications across the e-Health domain including efforts such as Health-Level 7 (HL7) (http://www.hl7.org/), SNOMED-CT (http://www.snomed.org/snomedct/), OpenEHR (http://www.openehr.org/), International Classification of Disease version 10 (ICD-10) and version 9 (ICD-9) (http://www.connectingforhealth.nhs.uk/clinicalcoding/classifications/icd_10) and Read coding (http://www.connectingforhealth.nhs.uk/clinicalcoding/faqs/) classifications are used across the NHS along with many other standards for imaging such as DICOM (http://medical.nema.org/). It is possible to develop Grid based solutions which will allow access to and usage of data sets existing in or supporting all of these classifications, however the effort required to do so is considerable since they are often much more than simply a data classification. The HL7 data model and information schema for example incorporates messaging infrastructure and in its entirety appears to be (for the VOTES Grid team learning about the e-Health/clinical arena) a particularly complex solution which offers much functionality, but at, it would seem, the cost of complexity. It is also not clear precisely how widely supported the HL7 standards are in the NHS across Scotland. Instead the focus of VOTES currently is looking at those data sets associated with SMR, GPASS and SCI Store.

Understanding the data models (schemas) of the data sets and software solutions that have been provided initially by the NHS and associated partners in VOTES is also a non-trivial process. In early explorations of the SCI Store database, data was provided and software however there was no explicit data model that was given. Of the ~100 tables that were provided, only ~15 of these contained training data sets. Extracting further information on these data models, e.g. the data schema was not immediately successful due to the commercial software providers asking for consultancy fees. Instead the VOTES team established the overall SCI Store data schema themselves.

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, e.g. via Public Key Infrastructure (PKI) technology; *authorization* – the discovery of that user's privileges based on their identity; *accounting* – logging the activity of users so that they can be held accountable for their actions within a system. Authentication is a well understood process in the Grid community however there are a variety of ways in which authorisation and accounting can be done. The NeSC at Glasgow have extensive experience of authorisation technologies and their application which are helping to guide the security solutions being put forward for the VOTES Grid framework. It would appear that there is no uniform security infrastructure in existence across the NHS. Instead various home grown solutions are the norm. Thus how firewalls are configured or how anonymisation is done and/or data encrypted is currently largely non-uniform.

The definition of policies for security and their enforcement is currently being explored in initial prototypes of VOTES. One challenge is this area is that the clinical data sets we are dealing with are

not always likely to be linked. Developing data Grids where single queries can be federated across multiple remote repositories and the results joined, requires that a common term is used for the subsequent join. Ideally all data should have for example the CHI number associated with it. At present this is not the case and different regions have supported this more than others, e.g. the CHI number is well advanced across Tayside, but less so across the other Scottish regions. Instead, record linkage is largely a process done internally by the NHS by institutions such as Information Services division of the NHS (www.isdscotland.org) however it should ideally be the case that other non-NHS bodies should be allowed to link data sets together. Thus it might well be the case that a clinical trial might want to link a multitude of NHS data sets such as SCI Store, with GPASS with disease registry data sets or other social classification data sets. Relying on internal NHS bodies to perform such work is not always practical, since they are primarily there to support IT facilities for patient care and not to support the academically oriented clinical research community.

A further challenge in developing a reusable security infrastructure is that a common security model for the NHS does not yet exist. Thus each site, hospital, registry, trust will likely have their' own in-house ways of setting up and enforcing security. It is unrealistic (it will be impossible) to try to enforce a new security model upon them, e.g. one based on advanced authorisation via Grid technologies. Until detailed evaluations by independent security focused IT staff from the NHS have worked with these technologies and are satisfied that they meet their stringent security requirements then these technologies will remain research prototypes only. Through VOTES, we are pushing these issues forward however and attempting to explore how we might define policies which can subsequently be enforced in restricting access to certain data sets by certain privileged personnel.

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 the unique CHI number associated with all patients. 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 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.

Many of the data access and security challenges depend upon the notion of consent. If a patient is willing that their data sets can be made available for research purposes then this greatly simplifies the ethical considerations which constrain the design space of the Grid framework. Within initial implementations in VOTES we have explored a variety of patient consent models.

5. 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 Grid framework is currently under production and is using a variety of Grid middleware. The basic architecture which supports federated queries in a user oriented but secure manner is depicted in

Figure 2. This infrastructure corresponds to one node of the Transfer Grid outlined above and is hosted on a trial test bed at the NeSC at the University of Glasgow.

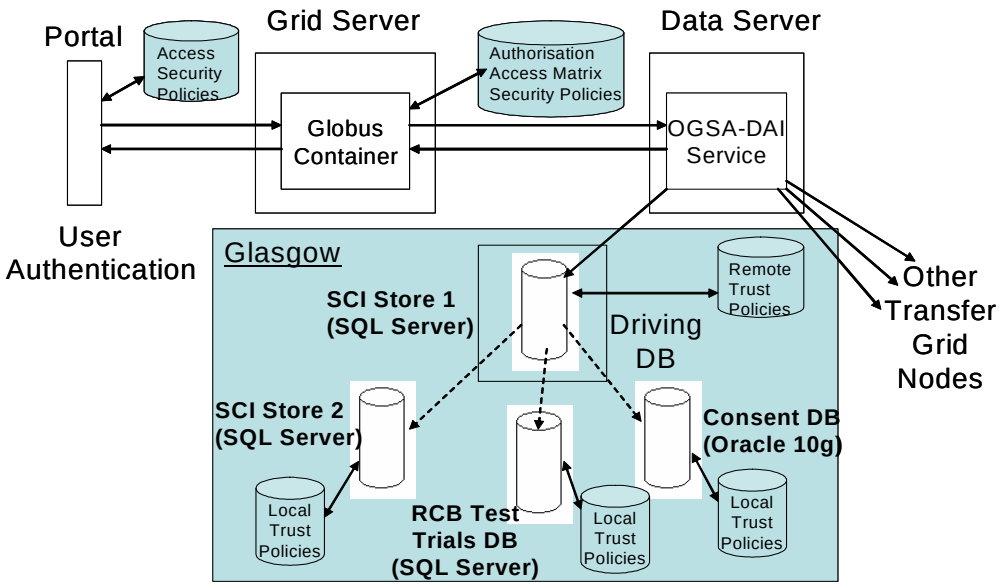

Figure 2: NeSC Transfer Grid Node Architecture

A GridSphere portal front-end communicates to a Globus Toolkit v4.0 (www.globus.org/toolkit) grid service, which in turn provides access to an OGSA-DAI data service. This runs queries from a "driving database" using standard SOAP message-passing, but also in turn runs queries from the subsidiary databases available from the pool for which it is responsible, using direct JDBC connections.

The user accesses this infrastructure through the portal, and provided they have the appropriate privileges, they can bring back a range of data from the various remote databases appropriate to their role. Currently the test infrastructure at NeSC Glasgow consists of multiple SCI Store repositories, a GPASS repository, a clinical trial data base containing representative clinical trial data from the Robertson Centre for Biostatistics at the University of Glasgow, and a consent database. To show a proof of concept, we have established canned queries that allow unprivileged users to retrieve limited data-sets, with the identifying patient data anonymised and other restrictions applied, whilst other privileged users are able to access and see a richer set of non-anonymised data as shown in Figure 3. 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 clinical trial queries are supported which seamlessly access and use distributed back-end test databases as shown in Figure 3.

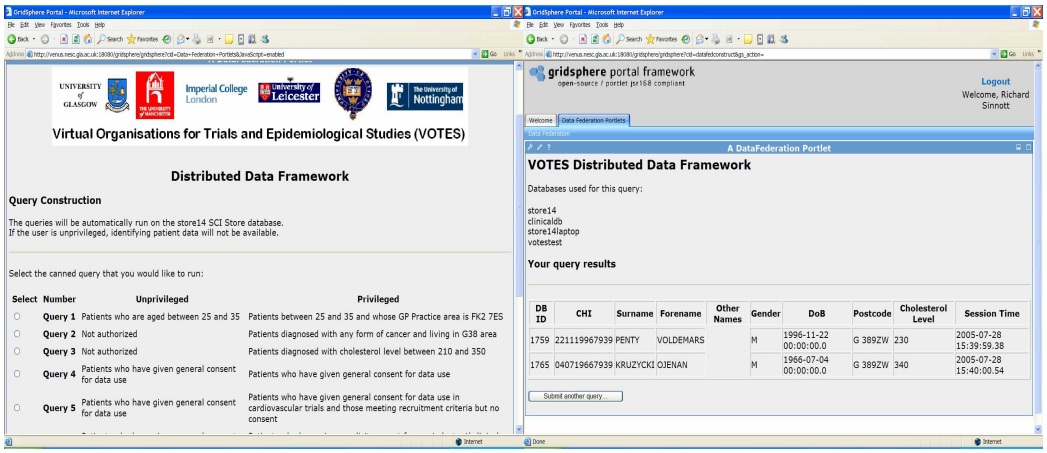

Figure 3: Screenshot of VOTES portal and privileged users result returned

Currently the prototype 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
- 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.

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 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.

6. Conclusions and Future work

The Grid is not a panacea for data access and integration and security, but helps to bring the issues on data access, integration and security to the fore. As such, 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 developing the current prototype, it is 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 CHI number 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. The Information Services group of the NHS in Scotland are working on record linkage and providing CHI numbers where the patient identification information can be confidently ascertained based on combinations of information such as names, addresses and dates of birth.

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 (https://cabig.nci.nih.gov/) and closer to home in genetics and healthcare projects across Scotland (http://www.innogen.ac.uk/Research/The-Scottish-Family-Health-Study).