TECHNICAL CHALLENGES IN LEVERAGING DISTRIBUTED CLINICAL DATA

Anthony Stell, Richard Sinnott, Oluwafemi Ajayi
National e-Science Centre
University of Glasgow
Glasgow
United Kingdom
a.stell@nesc.gla.ac.uk

ABSTRACT
As the digital age progresses the amount of clinical data that is being gathered and stored in electronic format is increasing rapidly. Much of this data is stored in isolated repositories, separated by administrative boundaries, large geographical distances and different standards of maintenance. However, there is great potential in unifying these data stores together and harnessing the collective information that they possess. One area that stands to benefit greatly from such an initiative in the unification of such data is that of clinical trials and studies.

Grid technology is a computing paradigm that attempts to find solutions to exactly this type of problem: how to federate data from wide-scale distributed sources on a time-scale useful to the typical end-user. As such, the VOTES project (Virtual Organisations for Trials and Epidemiological Studies) is a pilot-project, collaborative between seven UK institutions and funded by the UK Medical Research Council (MRC) to investigate how to implement a solution that draws distributed clinical data together in real time to enhance the primary trial processes – patient recruitment, data collection and study management.

This paper outlines the solution that has been implemented in the VOTES project, and the technical challenges that have arisen, focusing in particular on the security and usability of the back-end technologies used to perform the data federation.

KEY WORDS
Database and Information Systems, Distributed Clinical Data, Data Security

1. Introduction

Clinical trials and epidemiological studies are highly important processes. As the final measure in the suitably of all new drugs, treatments and medical interventions, their importance to society cannot be understated. However, the processes involved are also highly complex, and as the global population increases at an ever more rapid rate, the use of technology to conduct such studies is moving quickly from being a luxury, to being a necessity.

As such, trials and studies are very well placed to make use of the distributed clinical data stores that are already in existence around the world, gathered through a variety of different medical processes.

The ability to federate data in this way is one of the central paradigms of Grid technology: creating a data grid, the sum of which enhances the value of the information well beyond its individual value.

The VOTES project [1] has been commissioned to investigate the challenges involved in creating a data grid, specifically for use within the major processes of clinical trials. These have been identified as patient recruitment, data collection and study management.

Patient recruitment involves the targeting of members of the population with specific conditions or treatments necessary for the testing of a particular drug. Data collection focuses on the follow-up data relevant to the trial, and monitoring the patients involved. Study management relates to the efficiency of the conduct of the trial and also looks at the ethical review process, safe-guarding the interests of the patients.

As the project enters its third year, two main challenges have appeared to dominate the field: data heterogeneity and data security. Data heterogeneity refers to the idiosyncratic way in which data schemas evolve and develop when they are isolated from other repositories. Data security refers simply to the individual security policies that sites have, and how to marry this with an overall policy that takes account of all sites needs and interests.

In investigating these challenges, various aspects of a political as well as technical nature have arisen. Various technological solutions have been proposed, but often they must align with the political and strategic interests of the partners involved. This is the nature Virtual Organisations (VOs), which by their definition include partners that will work collaboratively towards a goal, but with limited trust between them, and with only a transient lifetime of the partnership.

The VOTES project has attempted to draw together the clinical centres at several academic institutions throughout the UK, and has built up a CVO (Clinical Virtual Organisation) that brings value to the end user of the trial system, that would not have been achieved had

they attempted to gain information through only one of the partner resources. A schematic diagram of the collaboration is shown in figure 1.

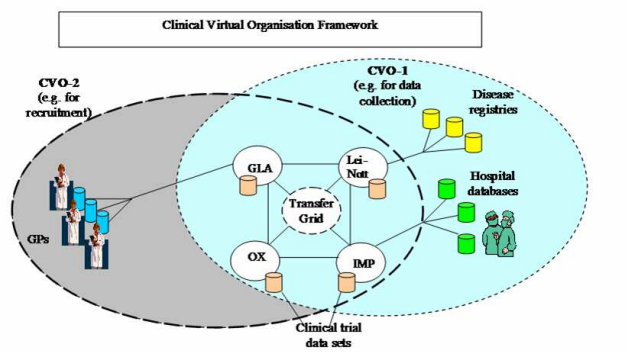

Figure 1: a CVO framework between the VOTES partners showing the various resources that can be linked through the "Transfer Grid" (denoted by the various institutions involved in the project – Glasgow, Oxford, Imperial College, Leicester and Nottingham).

The rest of this paper describes the various resources that the partners have brought to the collaboration, the technical implementation of how these resources are accessed, and the challenges and solutions that have been encountered.

2. Clinical Data Resources

To implement the VOTES project, a variety of clinical partners were brought together with the e-Science community. Through co-operative meetings and guidance, the two fields have reached a middle-ground where the needs and requirements of the clinical community have been reconciled with the abilities of the e-Science partners to provide a usable and viable solution.

As such, the clinical interests of the partners have been the first logical step in establishing how Grid technology can be applied to the clinical trial field. Broadly, some trends and discrepancies have appeared between partners depending on scale of trials that they are involved in. The larger enterprises include the UK Biobank project based out of Oxford and the Robertson Centre for Biostatistics based in Glasgow. These centres demand much more in terms of rigorous security and quality-control testing, which places a certain amount of constraints on what is achievable within the scope of the VOTES project – a pilot, research-based endeavour.

Smaller-scale trials are being conducted by centres based at Imperial College London and at the University of Nottingham. Due to their relative scale, the rigours of applying new technology here are much less, and also give application to data-sets that are very limited in scope. However, a major advantage is that it allows the technical nature of such collaborations to be explored more fully, with the minimum of political considerations impeding progress.

2.1 UK Biobank project

The UK Biobank project is a large-scale enterprise to accumulate and document genetic material from 500,000 members of the population between the ages of 40 and 69. The project is being conducted with a view to improving the prevention, diagnosis and treatment of a wide range of serious and life-threatening illnesses [2]. A Scottish arm of this project is also being conducted under the title of Generation Scotland [3].

The main tests in the project involve measuring a variety of external facets (height, weight, etc.), some basic functions (spirometry, hand-grip), then collecting some samples of blood and urine. The end result is envisioned to be an electronic catalogue of health information on a scale that will be amongst the largest in the world.

The usefulness of bringing such a resource into the VOTES project is immediately apparent. By cross-referencing such a database with other data sources in the country, focused on other more specific studies, it would greatly enhance the information at the disposal of an investigator conducting a clinical trial.

2.2 Scottish Clinical Records

A variety of data, known as the Scottish Morbidity Records (SMR) [4], has been released by the Information Systems Division (ISD) of the National Health Service (NHS) in Scotland for the purposes of the VOTES project. These data-sets include anonymised information for the following processes:

- Hospital discharges
- Psychiatric admissions and discharges
- Cancer registry
- Deaths

The data-sets are comprehensive and provide an invaluable resource of information that would provide highly relevant information to investigators conducting trials or feasibility studies for nationwide trials.

In addition to the SMR data, the SCI Store [5] data repository and GPASS [6] clinical software have been added to the sources available to the VOTES project. These applications provide the main mechanism for processing patient data in Scotland. With 85% of GPs using GPASS to enter patient details, the records are periodically uploaded to the central SCI Store repository, and made available to other clinicians through this centralised process.

Though the structure of the data is formally described as being standard across the NHS in Scotland, the schemas have developed with differing characteristics. As a result,

the mechanism, while providing a step towards the federated data model, still requires some work to be uniformly available to all participants.

2.3 Other studies

Two other studies have been made available to VOTES by virtue of the clinical partners within the project. These are a diabetes study conducted by Imperial College London, drawn from the General Practice Research Database (GPRD) [7], and a benign prostate hypoplasia (BPH) study [8], which is shortly to be conducted at the University of Nottingham.

Though these studies provide individual information that is relatively isolated on a national scale, they provide useful templates that could potentially be replicated for different conditions at different areas, and linked to each other to provide valuable insight into medical linkage.

3. Implementation

The software solution that has been developed for the VOTES project is shown in figure 2. Though it encapsulates the main facets, it should be noted that it is an intermediate representation (the latest is shown in figure 3). Representing a single node on the CVO, there are four main components:

- Portal server – allowing easy access for end users
- Grid server (implemented using the Globus Toolkit v4.0 [9]) – provides an intermediate point where the VO-wide security policy can be enforced.
- Data server (implemented using OGSA-DAI [10], a data integration software) – provides a means of connecting to the other nodes within the CVO.
- Driving database – joins the pool of local auxiliary databases that hold the various types of clinical data.

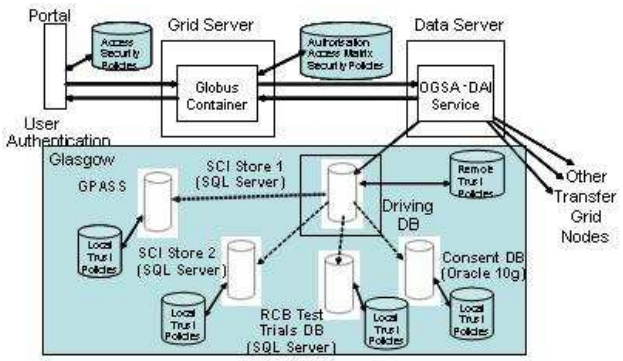

Figure 2: the architecture of a single node on the CVO within VOTES.

One of the main features to note are the security policies, of which there are two types: the CVO-wide policy enforced by the grid server, and the local security policies attached to each of the auxiliary databases. The latter allow local control to be retained by the resource owner, which is a mandatory requirement (otherwise sites would not be willing to join the CVO). The former is essentially aggregation of these local policies, which is published to the rest of the CVO.

The other major feature is the "joining" of the data which also occurs at two levels: the local auxiliary databases are joined using distributed SQL at the driving database, whilst the other nodes are connected to using the OGSA-DAI server. (The details of how nodes are connected are explained later in the paper.)

3.1 Data Indices

Key to the ability of these resources to be joined is the notion of a common interface or index value, which allows the data-sets to have a linked relationship, which is highly dependent on the data-sets in question.

However two potential values have emerged to fulfil this function: in Scotland, the Community Health Index (CHI) has been mandated as an identifier that uniquely references individual patients within the country. In England, the equivalent identifier is the National Health Service (NHS) number. These numbers do not have any corresponding relationship with each other, yet population migration between the two nations is relatively easy due to the political structure of the UK.

A future development of the VOTES project, or any similar endeavour, will be the need for probabilistic data-matching which would be able to identify patient records, to within a certain tolerance, based on other criteria than a single identifying number.

Issues like this do in turn bring up questions of privacy. The key issue is the security of the patient's data, and preventing users without the necessary privileges being able to identify patients based on the potentially sensitive medical information being processed. As such, the records available are anonymised before reaching the processing stage – ensuring that whilst the individual records do represent unique individuals, but without being able to identify those individuals.

However, the problem of statistical inference still remains, where given a set of queries which cannot identify a patient when run in isolation, would be able to given another set of innocuous data (for instance, an unusual condition in a specific postcode – on their own, these data points cannot identify someone, but potentially could when put together).

3.2 Connecting Securely

As is evident, the need for security in all aspects of this data federation is paramount. However, all security is a risk analysis of flexibility against privacy. Incorporating the various requirements of the clinical partners that have been captured throughout the lifetime of the VOTES project, it has been found that the ideal solution is to provide three mechanisms by which data resources can be accessed:

1) Direct JDBC (Java DataBase Connector) connections
2) Distributed JDBC connections
3) Limited web service connections to pre-canned queries

The schematic difference that this makes to the back-end of the VOTES architecture is shown in figure 3.

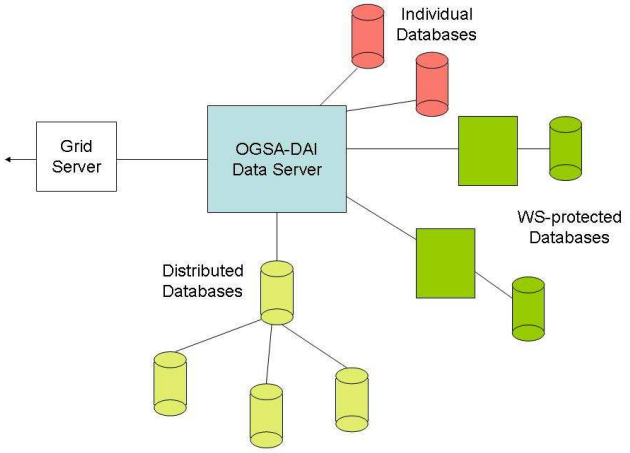

Figure 3: the three different types of methods to access the various data resources. Note that the yellow distributed database section is what is represented in figure 2.

The web service method is a service accessible using XML technologies, which exposes a limited set of methods which all run certain "pre-canned" queries over the database that they sit in front of. This allows much greater control over the access to that data from outside parties. The larger studies, such as UK Biobank, have expressed a preference for the web service access to data, precisely because of this greater element of control.

However, the control that the web service method affords is at the cost of the flexibility of queries that can be run over the data source. As such, JDBC connections are still a viable option. In terms of security, they can also be locked down to prevent unauthorized access, but the responsibility of that access gets passed on to the querying client, as they necessarily have to provide username/password access to the database in some form.

The VOTES infrastructure incorporates all three (web service, JDBC and distributed JDBC) of these access methods and using Grid technologies such as Globus and OGSA-DAI provides a seamless infrastructure that appears to the end user to be one unified data source. A screenshot of the results of a typical trial is shown in figure 4 (at the end of the paper).

4. Web Service Challenges

In the course of development however, there have been a number of technical and political challenges, mainly with regard to security. Outlined below are the main issues that have been experienced.

4.1 Interoperability

The theory of web services is to provide a uniform method of transmitting data between programmable modules residing in different domains in a predictable and secure manner. As such, the development of XML-based protocols such as SOAP and WSDL have provided a standard framework, by which different components in different domains can communicate in an interface that each understands [11]. A variety of web service implementations exist, and the relevant standards that have been developed, are all supported by the major software vendors, such as IBM, Microsoft and SAP. Most of the mainstream programming languages have methods of implementation (e.g. Java WS, .NET, etc.).

However, using these implementations is a non-trivial task, which does not translate to widespread use with other proprietary solutions. Tutorials and frameworks often provide environments in which the construction of web service applications is opaque to the developer. While this hiding of complexity is a central feature of "good" software engineering, the result is that it is often difficult to export the use of web services into the context of other environments (such as the use of a portal server, which comes with it's own complex environment).

4.2 "Stateful" interactions

Further issues include the need for statefulness between client and server within a SOAP communication. Being built upon HTTP protocols, this means that the extended communication between parties is inherently stateless. To achieve a conversation between parties with more than one exchange requires a session to be established using some mechanism. The favoured current mechanism is to use cookies within browsers, single-value text files that allow the server to "remember" and identify the calling party as being the same one as before.

Whilst cookies are a useful measure for maintaining state between parties, they have a number of deficiencies that warrant the continued search for a better method. A cookie is essentially a place-holding text value that is passed between client and server. Its easy accessibility and modification make it less than secure method of

identification, and would require re-authentication with every transaction. A possible alternative would be the use of digital certificates, which unequivocally and securely identify the user within each transaction (see section 4.4).

4.3 Data Thresholds

The information that is communicated in a SOAP transaction, is "wrapped" in several layers of XML. This is how the standardisation of communication is achieved between services. However, because of the large file-size of the "wrapping" material, the volume of data that can be transmitted between parties quickly becomes an issue.

In the VOTES project, the limits have been explored with the Scottish Morbidity Records, which contain over 3 million patient records. Once the threshold is reached the communication times out and the connection is lost.

One solution that has been presented to this problem is the use of more sophisticated communication constructs which have been developed by the OGSA-DAI team at the University of Edinburgh [11]. Their solutions involve the use of "chunked" data streams, which divide the XML packages into manageable packets, which are transmitted in separate communications, but marked as one conversation.

This provides a useful and applicable technical solution to the problem. However, the follow-up issue is to be able to convince the clinical partners to adopt this technology. The next step is to provide working scenarios to demonstrate the application's security and usefulness in this context.

4.4 Encryption

Finally, as has been emphasised throughout, the security of any solution is of paramount importance, and this must be applied to the use of XML communications. Because of its textual nature, the use of web services results in clear-text information being passed across networked infrastructures between the communicating parties. In order to effectively secure this, there must be some form of encryption and data integrity applied, which is done using digital certificates.

Though such encryption methods are an established technology, with a mathematically proven ability to encrypt data, the effort in setting up such security is again non-trivial. It is again also the case that the methods of setting this encryption up vary depending on which development platform is being used.

5. Direct Connection Challenges

The use of direct connections to databases provide the most effective way of finding data to join or enhance any clinical search. However, there are two major issues with this technology: security of access to the data and the lack of support for the distributed version of direct access.

5.1 Security

The main issue encountered when attempting to connect to a remote data source is that of security. JDBC connections can be encrypted using a variety of digital certificate technologies that can be obtained from Java cryptography packages, so the integrity and privacy of the data on the wire is not the central issue.

The question of responsibility and accountability is the one requiring addressed in this instance. A direct database connection provides a flexible method of access, and can be restricted depending as the database administrator sees fit. However, such a connection requires some kind of direct access through the site and database's firewall, and responsibility for the security of this connection lies with the remote party running the query. This has the potential to be a weakness in the security of the local site, which is a consequence unpalatable to some system administrators.

It is possible that the ability to directly connect to a database is the most useful method, rather than the web service solution. If this is the case, then the security concerns can be mitigated at a higher level – usually with some form of legal agreement between the acting parties that delimits explicitly the various responsibilities in the event of a security breach.

5.2 Support

The distributed JDBC connection is a more efficient and load-balanced version of the single database connection, as it draws the data together from several auxiliary databases using the Microsoft-supported Transact-SQL language.

However, because of the commercially competitive nature of the products involved, this does restrict the number of data sources available. Whilst most clinical databases are implemented using the Microsoft-based SQL Server and Access products, others do exist based on open source implementations, such as MySQL, or using Oracle enterprise solutions. As the aim of the project is to provide as general an infrastructure as possible, the architecture has been re-engineered in an attempt to gain broader appeal.

The solution to this problem has been to allow the OGSA-DAI software to join the data at the data server level with data sources that are not Microsoft-supported. This allows a more efficient hybrid system to be used which makes the most of the load-balancing efficiency of distributed connections where possible, whilst still being able to join with other database implementations.

6. Conclusions

The methods outlined in this paper are representative of the mainstream technologies available for federating data in the way required by clinical trials and studies. As has been discussed, there are a variety of advantages and disadvantages for each method. As the project progresses, and the number of iterations that streamline and enhance the solution provided, the most acceptable and useful method will become apparent.

What is clear is that the technical challenges are only part of the story with regard to development of such infrastructures. The issues described in this paper reflect the state of maturity of many of the technologies, and this in turn has a knock-on effect in the uptake of such solutions. Through persistent demonstration of these applications, along with their usefulness and security in the context of trials and studies, widespread adoption will hopefully be encouraged and the benefits of the technologies will become apparent.

A variety of other works are being conducted in this field, including the CaBIG project being conducted at the National Cancer Institute [12], though projects such as this are focusing on bringing Grid technology to slightly different aspects of clinical science than VOTES. However, the data federation paradigm is almost universal in its appeal for building enhanced clinical infrastructures. In collaborative meetings with the project a discussion of the challenges has elicited that most of the issues highlighted in this paper are common to the entire field.

The work conducted in VOTES has already attracted interest from other clinical parties keen to exploit this federation of related data sources. The National e-Science Centre is currently constructing a portal for the Scottish Bioinformatics Forum [13] which will access their databases containing a wide range of sample data on breast cancer tissue. This will be added as another catalogue data resource that will help with trials and studies around the country.

In the same project, the portal will be linked to the GEMEPS project [14], in a search engine has been developed which matches data results to within a specified degree of probability to similar experiments, in the context of micro-array experiments. With these facilities at the disposal of trial investigators, it is hoped that a global vision of a single infrastructure, that leverages the clinical data repositories around the world, will become a reality.

Acknowledgements

The authors would like to acknowledge the UK Medical Research Council, which provides the core funding for the VOTES project.

References

[1] Virtual Organisations for Trials and Epidemiological Studies (VOTES) – http://www.nesc.ac.uk/hub/projects/votes
[2] Ollier, Sprosen and Peakman, UK Biobank: from concept to reality, Special Report Pharmacogenomics 2005, Futuremedicine.com
[3] Generation Scotland Scottish Family Health Study, http://www.innogen.ac.uk/Research/The-Scottish-Family-Health-Study
[4] Scottish Morbidity Records - http://www.statistics.gov.uk/STATBASE/Source.asp?vlnk=1106&More=Y
[5] SCI Store - http://www.show.scot.nhs.uk/sci/products/store
[6] GPASS - http://www.gpass.co.uk
[7] GPRD – http://www.gprd.com
[8] Foster, Kesselman – Globus: a metacomputing infrastructure toolkit, International Journal of High Performance Computing Applications, vol. 11, no. 2, 115-128 (1997)
[9] Karasavvas et al. – Introduction to OGSA-DAI services, Lecture Notes in Computer Science, vol 3458/2005
[10] Web services - http://en.wikipedia.org/wiki/Web_services
[11] OGSA-DAI software functionality - http://www.ogsadai.org.uk/about/
[12] Cancer Biomedical Informatics Grid (caBIG): pioneering an expansive network of information and tools for collaborative cancer research – Kakazu, Cheung, Lynne – NCBI, Pubmed (PMID: 15540527).
[13] Scottish Bioinformatics Forum - http://www.sbforum.org/
[14] GEMEPS - http://www.nesc.ac.uk/hub/projects/gemeps

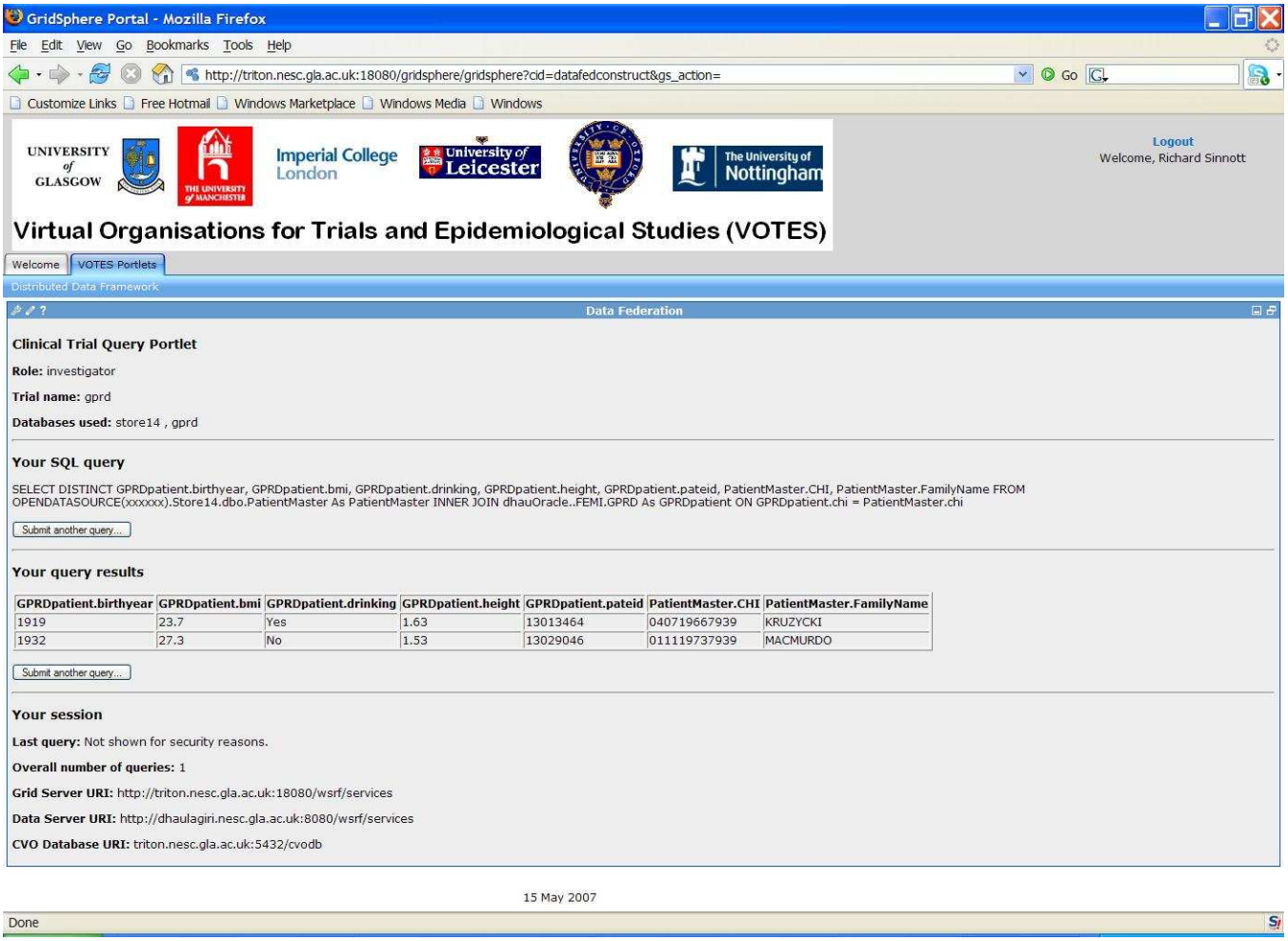

Figure 4: a screenshot as seen by the end user of the VOTES portal. A variety of records have been returned that relate records that have been found in the GPRD study in London, with the SCI Store registry in Scotland. The query, role and databases used have been outlined along with a cache of the infrastructure components that have been used. In a typical scenario this type of query could be used to identify the location of patients with a specific condition, thereby allowing the efficient targeting of a patient recruitment campaign.