Supporting the Clinical Trial Recruitment Process through the Grid

Anthony Stell, Richard Sinnott, Oluwafemi Ajayi
National e-Science Centre
University of Glasgow, UK
ajstell@dcs.gla.ac.uk

Abstract

Patient recruitment for clinical trials and studies is a large-scale task. To test a given drug for example, it is desirable that as large a pool of suitable candidates is used as possible to support reliable assessment of often moderate effects of the drugs. To make such a recruitment campaign successful, it is necessary to efficiently target the petitioning of these potential subjects. Because of the necessarily large numbers involved in such campaigns, this is a problem that naturally lends itself to the paradigm of Grid technology. However the accumulation and linkage of data sets across clinical domain boundaries poses challenges due to the sensitivity of the data involved that are atypical of other Grid domains. This includes handling the privacy and integrity of data, and importantly the process by which data can be collected and used, and ensuring for example that patient involvement and consent is dealt with appropriately throughout the clinical trials process. This paper describes a Grid infrastructure developed as part of the MRC funded VOTES project (Virtual Organisations for Trials and Epidemiological Studies) at the National e-Science Centre in Glasgow that supports these processes and the different security requirements specific to this domain.

1. Introduction

To test new drugs and treatments for clinical care requires careful and long-term testing before they can be prescribed to the population in general. To facilitate such testing requires identification and recruitment of large groups of the population that fit certain criteria related to the specific condition that the drug is addressing.

Automating this process has numerous advantages including reduced cost and expediting the clinical trials process as whole, by avoiding unnecessary contacts with non-suitable members of the public. An example of this is where the recruitment criteria require candidates with a cholesterol level between certain bands. Such information is not typically known by the vast majority of the public. Weeding out patients outside of the needed bands is therefore beneficial. Over-recruiting is also advantageous since it may well be the case that potential candidates may decline to be involved in a given trial, or alternatively that certain candidates are better matches than others.

In order to co-ordinate such a trial process requires the combined effort of numerous personnel with specific roles in the whole process: clinical trials investigators and nurses wishing to recruit patients for a given trial; ethical oversight committee members and Caldicott guardians responsible for deciding on the ethical aspects of the study; clinicians and general practitioners (GPs) responsible for individual patients; and importantly the patients themselves. Before any access to identifying patient data sets is made, it is necessary to obtain patient consent for the use of a given patients' personal and private medical data. The interplay between these roles and the process by which a clinical trial is co-ordinated is crucial to the overall success, viability and legality of a trial.

Given the fact that the clinical data sets are typically scattered across many resources and institutions, including GP databases, hospitals, and disease registries amongst others, Grid technology, in principle, provides many potential advantages to deal with data federation. However, this domain also has numerous challenges, especially related to security, which must be explicitly addressed. Due to the sensitive and confidential nature of the data in this domain, strict controls are required on access and distribution, with only sufficiently privileged actors having the appropriate levels of access.

The VOTES project (Virtual Organisations for Trials and Epidemiological Studies) [1] has been funded by the Medical Research Council (MRC) to explore this problem space. One of the focuses of VOTES, and the primary focus of this paper, is to develop Grid solutions that address large-scale recruitment needs in the clinical domain. In addition VOTES is addressing two other important areas in the support of clinical trials and epidemiological

studies: data collection and study management. Furthermore, it is a requirement that the infrastructure that will be developed will be effective yet simple to use for the non-Grid personnel involved in the clinical trials process.

2. Clinical Patient Recruitment

As described previously, clinical patient recruitment is a large-scale and resource-consuming exercise. The human challenge in co-ordinating such a large effort can be immense and in some cases, such as the UK Biobank Project [2] the number of potential candidates is so large that the use of distributed technology is mandatory for the task to be completed in a meaningful time-scale. (UK Biobank expects to recruit 500,000 members of the population between 40-69 years of age).

2.1 Crossing Domain Boundaries

To effectively identify suitable trial candidates on a large a scale requires knowing the structure of patient information data sets across a broad set of domains. For instance, to sample the national population of the UK would require knowing the data structures of the health services in England and Scotland, knowing how they relate to each other and knowing how to translate between the schema of both.

Ideally there should be a single electronic health record which captures all necessary health information associated with a given patient that is accessed and updated by all health care providers throughout a patients' lifetime. This should support tracking of a patients' place of residence throughout their lifetime, and allow for cross checking of records in one area to those of the other. However such a single e-Health record remains a distant wish and a variety of heterogeneous and largely non-interoperable legacy infrastructures and data sets is the norm across the NHS, with paper based patient case histories and records still commonly used.

To support the linkage of distributed data sets associated with a given patient, it is beneficial to have a common, unique identifier for patients that spans all domains and can be used to join the patient records between databases. In Scotland this unique identifier is a number known as the Community Health Index (CHI), which is currently being rolled out across the nation as part of a new Scottish parliamentary initiative. It is planned that the CHI number will be rolled out across all of Scotland by mid-2006. In England, the unique identifier is the NHS Number, which is a distinct entity from the CHI. Relating these two numbers, which have different structures in different contexts, is a major, yet unavoidable, challenge.

2.2 Recruitment Work-Flow

A patient recruitment process must ideally capture patient consent as early as possible. Whilst information on patients is stored in a variety of digital formats and locations, *a priori* consent that these data sets can be accessed and queried to decide that a given patient be recruited to a clinical trial, is needed. One of the best sources of information associated with potential trials candidates is through primary care sources, i.e. in their GPs databases. Understanding whether a given clinical trial is in the interest of a particular patient is best answered by these GPs.

Figure 1 gives a pictorial representation of the process/workflow through which a clinical trial investigator and a GP might interact to support primary care recruitment.

1. The trial co-ordinator wishes to set up a new clinical trial, with a specific description of the drug/treatment and the patients' characteristics that would potentially qualify them for entry into the clinical trial. They need to contact a set of GPs to describe this new trial and find out if these GPs have patients that fit the required criteria.
2. Assuming that the GP is interested in participating in this particular trial, they need to search their own patient records for anyone that may fit these criteria.
3. Assuming that one or more matching patients have been found, the GP decides if it is really in the patients' interest to participate in the clinical trial. If this is the case then said patient is contacted by the GP and told about the trial. If they are willing to participate having had the potential benefits and issues that might be associated with the trial fully explained, they are asked for their consent to use their personal data for this purpose. Once consent is obtained, this information is recorded.
4. Based upon the specifics of the clinical trial, various data sets are collected from the GPs database. These can be non-identifying information such as age, height, weight, medical conditions and social and demographic details. Identifying information may be anonymised with the de-anonymising keys maintained by the GPs.
5. This information along with the note of consent is communicated back to the trial

co-ordinator. The information is then validated and stored for later use in the trial.

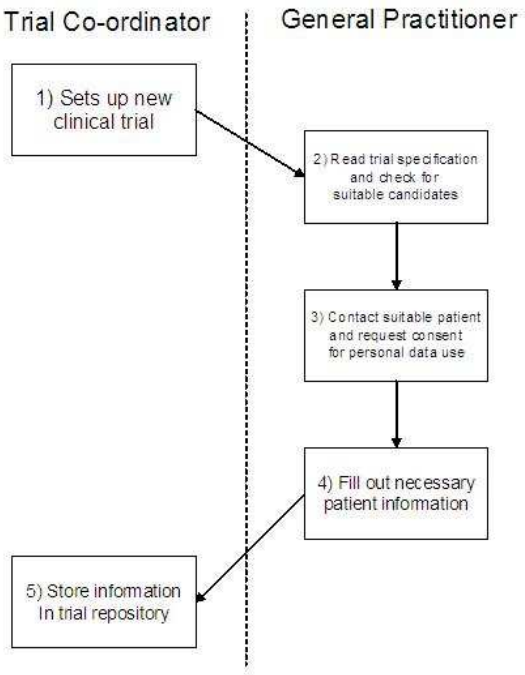

Figure 1: A diagram of the primary care patient recruitment workflow.

One of the central requirements of this process is the need to clearly separate the distinct duties of the two actors involved. There is necessarily a disparity between the privileges of the two roles, and the data that they will respectively be allowed to access, and at which particular times. To enforce this type of interdependent and role-based access control requires a sophisticated security system.

3. Grid Security for Clinical Trials

Grid computing depends on the collaboration and sharing of resources across domain boundaries. A loose coupling of resources and user access to achieve a specific goal over a set period using Grid technology is often termed a Virtual Organisation (VO).

Traditionally, Grid security has been expressed in terms of access control between nodes within a VO. In this VO, sites have only a limited amount of trust between each other, yet they also wish to share certain ring-fenced resources that will allow the VO to accomplish its overall goal.

However, enforcing data security in the clinical and health-care domain is a more complex problem. While the tenets of Grid security apply – that of Authentication, Authorization and Accounting ("AAA") – there are other more subtle requirements that must also be met when attempting to set up a system that is flexible enough to use Grid technologies effectively, but also maintains the high standards of privacy and integrity required in the clinical domain. A term used to describe this particular entity is a Clinical Virtual Organisation (CVO). The level of prescription of security policies and how they are enforced must be strongly adhered to within such a CVO.

3.1 Clinical Security Considerations

Clinical security can be broadly divided into the following three areas [3]:

- Sensitivity
- Consent
- HCP (HealthCare Professional) speciality

Sensitivity relates to the importance and privacy level applied to a data field within a health-care record. Considering risk analysis, it can be described in terms of the possible consequences if privacy of this data record is broken. The level of sensitivity will ultimately be determined by the HCP dealing with the particular record, which in turn relates to the speciality of that HCP (see below). Where the records are being used for statistical aggregation of data, a corollary of sensitivity is the need for "anonymisation" of that data – where any information that identifies a patient must be hidden from unprivileged users.

Consent relates to the requirement of asking a patient whether they will allow their information to be used in this clinical trial. There is a subtlety in obtaining consent as to what parts of the patient's record they will release for use and which they wish to remain private. To allow this "pick and mix" release of consent gives a more flexible structure. However, the patient must be guided in this by the GP, to provide a professional opinion on the consequences of releasing this consent on differing parts of their data records, with the possibility therein of differing levels of sensitivity.

HCP Speciality refers to the many different categories of HCP that can exist. This affects access to data records as one HCP may have rights to see highly sensitive records in their own fields but not in another. This speciality will be classified within the security policy that defines privileges within the VO.

Another consideration that is idiosyncratic to security within health-care is the scenario sometimes called the "broken glass" situation. This is where highly sensitive data is accessed by an HCP that does not have the necessary privileges, in an attempt to save a patient's life. In the immediate situation, the HCP believes it

necessary to access this record, and time is of the essence. The system here is to let the HCP have access to this information but have an irreversible record that this unprivileged access has occurred. When the immediate situation is resolved the logs of the event should be investigated by an auditing authority, to see whether the actions of the HCP were justified.

3.2 Clinical Data Security Policies

A security policy defined in one site node of a VO will not necessarily have the same structure as a different node within the same VO. This is inherently tied to the structure of object classification within a specific domain, as security can only be defined and enforced on data that itself has a well-defined structure. Put another way, sites must have their own autonomy and hence define and enforce their own security policies on access to different data sets.

There are numerous developments in standards for the description of data sets used in the clinical domain however. These are complex and evolving with numerous commercial bodies and standards groups involved in developing strategies and include major initiatives such as Health-Level 7 (HL7) [4], SNOMED-CT [5] and OpenEHR [6]. There is 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) 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. ICD-10 was introduced in 1993, but the ICD classifications themselves have evolved since the 17th Century [7].

To compound this problem of classification, there is the issue of the dynamic nature of virtual organisations. One of the standard characteristics of a VO is that it is not only a loose collaboration between sites but it is also a transient one, with a limited lifetime and for a specific purpose. As such, any security policy enforced will also have a limited lifetime – requiring the re-evaluation of security requests after a given time period. This must be considered when defining security policies and establishing chains of trust.

3.3 Grid Security Solutions

Security solutions in the Grid community are largely categorised by where they fit into the "AAA" scenario.

Authentication – this is almost always achieved using Public Key Infrastructures (PKIs) where public and private keys and certificates are used to verify the authenticity of a user's identity.

Authorization – as this is a more complicated requirement, in terms of establishing privilege rights based on identity, there is a wider range of possible solutions in the community (PERMIS [8], CAS [9], VOMS [10], Akenti [11]). These applications all have various advantages and disadvantages depending on the needs of the developer and the implementation idiosyncrasies, but no clear leader has yet been established in the field. In the VOTES project, authorization is provided by a simple implementation of an Access Control Matrix (see section 4.3).

Auditing – this is a security measure that has more relevance later in the production cycle of a system. Whilst not underestimating the importance of logging all user activity and being able to attach events to individuals, design and production in the VOTES project is currently focused on securing access to the system in the first instance and supporting the recruitment process.

4. VOTES Implementation

The details of the portal application that is currently under development to provide solutions to these security, usability and process challenges, are now described in this section.

4.1 Grid Technologies

The implementation of the VOTES project has built largely upon the expertise in certain Grid technologies based at the National e-Science Centre in Glasgow. These include the following applications that provide tools to develop and maintain the application: GridSphere, Globus and OGSA-DAI.

GridSphere [12] is a web portal technology that provides easy access to secure grid services. It can be presented as a group of layered portlets allowing simultaneous, and interactive, task processing. A development suite is available that provides easy-to-use tools and tutorials for building and deploying these portal services.

The Globus Toolkit [13] similarly provides a set of distinct modular tools that allow development of Grid services. Version 4.0 of the toolkit is implemented in the VOTES project, which has been developed to the Web Services Resource Framework (WS-RF) specification [14] - this is in line with a drive by Globus to align their service architecture to the more common web services standard [15].

OGSA-DAI [16] (Open Grid Services Architecture – Data Access and Integration) is a project that has developed a toolkit specifically for grid services that federate data from distributed sources. The OGSA-DAI services developed in VOTES again follow the WS-RF specification, but this is implemented in addition to direct JDBC connections to local data sources. The major advantage of using OGSA-DAI is that many different data sources, from DBM systems to XML and flat files, can be queried. The data returned is then rendered in a standard XML format, and tools are provided to allow easy analysis and presentation of this format.

4.2 GPASS and SCI Store

In the VOTES implementation thus far, the development team have been given access to relevant software and realistic (representative) data sets used to explore the functionality of the NHS services and data. Negotiations are on-going in rolling out the services described in the following sections across the wider NHS in Scotland and in gaining access to actual clinical data sets

The General Practice Administration System for Scotland (GPASS) [21] is a client-side application that is in use by a large number of general practitioners, in over 890 practices in Scotland. It provides a portable, 'on-the-spot' electronic interface for GPs to input patient data, to be synchronised to the centralised SCI-Store [22] repository at such times as a connection can be made. It also contains a local data store that houses and inter-relates information on drugs and treatments, thus providing a rudimentary ability to advise on a particular treatment, depending on patient symptoms and history [23]. The interface used to show the electronic record of patient history is shown in Figure 2.

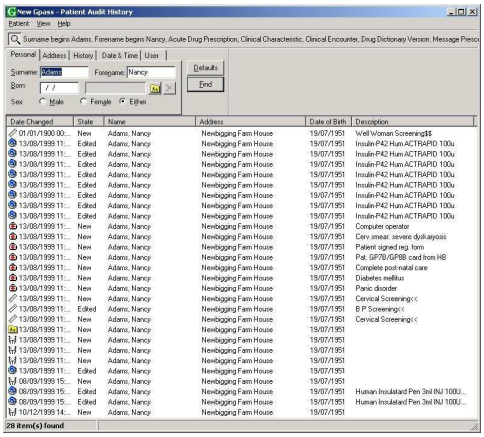

Figure 2: A screenshot of the GPASS patient history function.

Because of its proximity to the 'front-line' of primary care, GPASS is the most pertinent technology when looking at patient recruitment for clinical trials. The distributed and asynchronous structure of the application makes the process of integrating this securely into a wider system a challenging but necessary one.

The Scottish Care Information (SCI) Store [22] is 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.

SCI Store serves as a repository for this data across regional and national domains. As individual practitioners update their GPASS databases, the information is automatically uploaded to this central repository. The VOTES infrastructure makes use of this by developing grid services that allow interrogation of the centralised SCI Store repository when collecting study data, and also allow interrogation of the individual practitioner databases for patient recruitment. An interface between the grid services already developed in VOTES and a test GPASS database is shown in Figure 3.

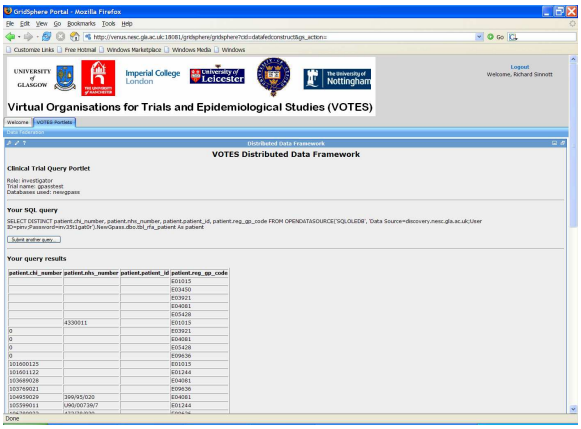

Figure 3: Results returned from a query to a local installation of GPASS.

The security implications in being able to query the individual databases used by GPASS are discussed in section 4.4.

4.3 VOTES Portal

The central theme of the VOTES project is to set up a Clinical Virtual Organisation (CVO) that will implement the three-fold vision of patient recruitment, data collection and study management.

In the context of primary care patient recruitment, the VOTES portal provides web access based on the role of either "Trial Co-ordinator" or "General Practitioner". (For data collection and study management a wider range if roles is supported). The grid and data services behind the portal provide distributed methods of:

- retrieving the necessary trial, patient and GP information.
- retrieving and storing the trial forms
- allowing asynchronous communication between the co-ordinator and the GP in the work-flow (see section 2.2).

An outline of how these scenarios are supported is depicted in the VOTES portal infrastructure shown in Figure 5.

The envisaged future development of the VOTES portal includes the addition of repeated modules in the overall architecture, including extra portal and data servers. (To see an outline of the current architecture, see [1].) This will allow a more "Grid-like" structure to be developed, providing features necessary in any production system, such as redundant failover, and also features specific to Grid technology such as intelligent load-balancing and distributed server functionality based on the resources available at specific nodes.

4.4 VOTES Security

NeSC-Glasgow has extensive experience in a range of fine-grained authorisation infrastructures across a range of application domains [17-19]. Whilst it is expected that the existing prototype will be moved to a more robust authorisation solution, the following authorization infrastructure has been developed, based on an access matrix as shown in Figure 4. This allows for rapid prototyping, which allows the problem space to be explored and user feedback to be obtained as early as possible.

	R_1	R_2	R_3	R_4
		h_1	h_2	h_3	h_4
U_1	U_1	0	0	1	0
U_2	U_2	0	0	0	1
U_3	U_3	1	1	1	1
U_4	U_4	0	1	0	0

$U_1(R_1 \Delta h_3) = 1$, $U_2(R_1 \Delta h_2) = 0$, $U_3(R_3 \Delta h_1) = 1$, $U_4(R_2 \Delta R_3 \Delta h_4) = 0$,
where Δ is a combination function, 0, 1 are bit-wise privileges, R_X, h_X are resources and U_x is a subject

Figure 4: Access Matrix Model

The authorisation mechanism implements an Access Matrix model [20] that specifies bit-wise privileges of users and their associations to data objects within the CVO. Data objects are defined as fields, tables, views, databases and sites, for the purposes of fine-grained authorisation. The access matrix is designed to enforce discretionary and role based access control policies. It is also scaleable to facilitate ease of growth parallel to the predicted growth of the infrastructure as a whole.

The NeSC at Glasgow have already shown in numerous other works [27,28] how Grid services can be protected through technologies such as GSI and PERMIS, however the effort in supporting these infrastructures is considerable and not conducive to rapid prototyping necessary to capture the *basic* functionality needed in clinical trials. Once the access matrix model has allowed for the detailed expression and enforcement of policy which the clinicians and all people involved in the clinical trials process are satisfied with, a move to a full RBAC model may well be considered depending upon the strategic direction of the project.

Security on the data sources 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 the DBMS level. 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 [9].

Considering security in GPASS, it is probable that due to the distributed nature of the application, a modification to the security model adopted so far in VOTES will need to be made in future development. The technology used in VOTES so far, in particular the portal's ability to query and return results from the back-end GPASS database, shows that it is possible to implement a web/grid service interface that provides a handle for third parties to securely interrogate GPASS.

However, until significant progress is made between the participating agencies in VOTES, it is unlikely that this interface will be adopted by the users and developers of GPASS. This is part of the human and political factor that must be overcome before the technology in VOTES will be taken up. In particular it is intended that the prototypes will be explored with sets of GP practices across the Greater Glasgow region through the SPPIRe network [26].

5. Conclusion

The prototype application developed in the VOTES project is currently a work in progress. It does not yet provide all the answers to the issues posed in this paper, but it does provide a starting place, and is being designed with the larger scheme in mind.

Using the current implementation it is possible to envisage a system that will identify potential trial candidates quickly, securely and efficiently. It is to be hoped that with the use of such electronic methods, the scope for error, such as mis-identification of patients or release of confidential information, will be very much reduced. A key aspect of the work is to support the capture of patient consent as early as possible in the clinical trial recruitment process. Scenarios where statistical information from GPASS is retrieved where, once consent is given through a combination of patient and GP interactions, more detailed information is returned are under development.

As with most research that is addressed using Grid Computing, only some of the cross-domain issues are to do with the technology. A lot depends on the human and political factors between participating bodies. These issues take time and the establishment of trust to be overcome. However, it is hoped that systems such as the one described in this paper will allow a closer and more "joined-up" network of clinicians and technologists to promote that trust and encourage closer collaborative work. To support this, it is planned that the researchers working on the VOTES project will also be given honorary contracts to work part time in the NHS in Glasgow.

6. References

[1] Virtual Organisations for Trials and Epidemiological Studies (VOTES) – http://www.nesc.ac.uk/hub/projects/votes

[2] UK Biobank project – http://www.biobank.ac.uk

[3] Roger Quin, NHS Greater Glasgow – Personal Communication

[4] Health Level 7 (HL7) – http://www.hl7.org

[5] SNOMED-CT – http://www.snomed.org/snomedct

[6] OpenEHR – http://www.openehr.org

[7] ICD background, http://www.connectingforhealth.nhs.uk/clinicalcoding/faqs

[8] PERMIS – http://sec.isi.salford.ac.uk/permis

[9] CAS – http://www.globus.org/toolkit/docs/4.0/security/cas

[10] VOMS – http://hep-project-Grid-scg.web.cern.ch/hep-project-Grid-scg/voms.html

[11] Akenti – http://dsd.lbl.gov/Akenti

[12] GridSphere – http://www.Gridsphere.org

[13] Globus – http://www.globus.org

[14] Web Services Resource Framework – http://www.globus.org/wsrf

[15] WS-RF specification v1.2 - http://www.oasis-open.org/specs/index.php#wsrfv1.2

[16] OGSA-DAI – http://www.ogsadai.org.uk

[17] 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 the 1st International Workshop on Bioinformatics and Security (BIOS '06), Vienna, April, 2006

[18] 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

[19] 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

[20] R. S. Sandhu and P. Samarati, "Access control: principles and practice" IEEE Communications Magazine, vol. 32, no. 9, pp. 40-48, 1994.

[21] GPASS – http://www.show.scot.nhs.uk/gpass

[22] SCI Store – http://www.show.scot.nhs.uk/sci/products/store/SCI Store_Product_Description.htm

[23] Paul Woolman, NHS Scotland Information Services – Personal Communication

[24] SMR – http://www.show.scot.nhs.uk/indicators/SMR/Main.html

[25] NHS Data Dictionary – http://www.isdscotland.org

[26] Scottish Practices and Professionals Involved in Research (SPPIRe) network, http://www.nes.scot.nhs.uk/SSPC/SPPIRe/

[27] 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, June 2005, Amsterdam, Holland.

[28] 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.

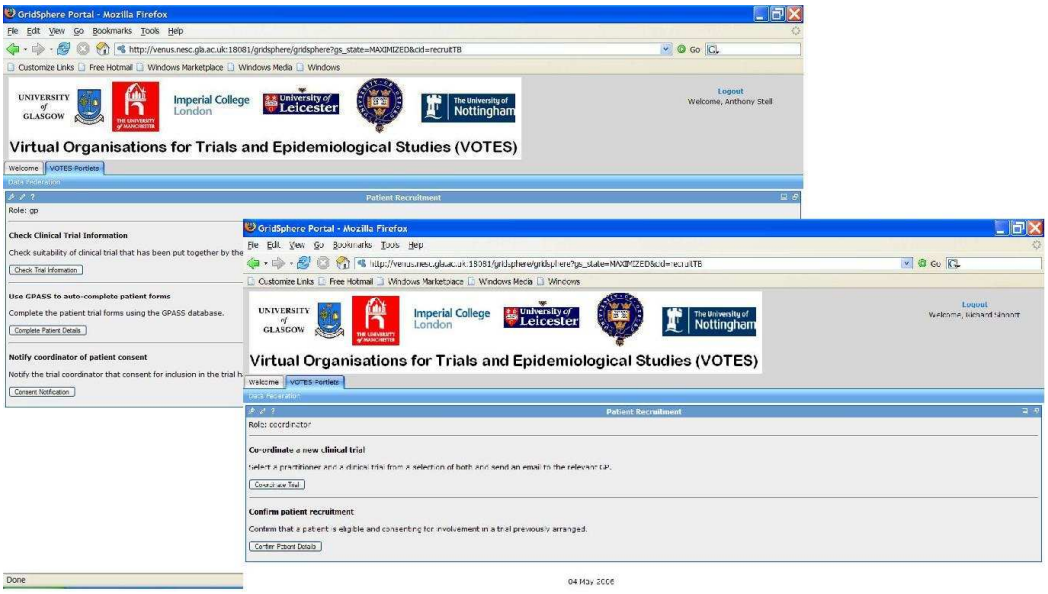

Figure 5: Two views of the recruitment portal. The left image shows the options available in the portal if the role is that of the GP. These options are to check the information on the clinical trial, to use GPASS to auto-complete the patient information and to notify the trial coordinator that patient information and consent has been obtained. The right image shows the coordinator role, which has two options, one to initiate the organisation of the trial and one to upload the final data for the patient.