# Handbook of Research on Computational Grid Technologies for Life Sciences, Biomedicine, and Healthcare

Mario Cannataro
*University Magna Graecia of Catanzaro, Italy*

Volume II

Medical Information Science Reference

MEDICAL INFORMATION SCIENCE REFERENCE

Hershey · New York

Director of Editorial Content: Kristin Klinger
Senior Managing Editor: Jamie Snavely
Managing Editor: Jeff Ash
Assistant Managing Editor: Carole Coulson
Typesetter: Jeff Ash
Cover Design: Lisa Tosheff
Printed at: Yurchak Printing Inc.

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com/reference

and in the United Kingdom by
Information Science Reference (an imprint of IGI Global)
3 Henrietta Street
Covent Garden
London WC2E 8LU
Tel: 44 20 7240 0856
Fax: 44 20 7379 0609
Web site: http://www.eurospanbookstore.com



Library of Congress Cataloging-in-Publication Data

Handbook of research on computational grid technologies for life sciences, biomedicine, and healthcare / Mario Cannataro, editor.
p. cm.

Summary: "This book provides methodologies and developments of grid technologies applied in different fields of life sciences"--Provided by publisher.

Includes bibliographical references and index.
ISBN 978-1-60566-374-6 (hardcover) -- ISBN 978-1-60566-375-3 (ebook) 1. Bioinformatics. 2. Computational biology. 3. Problem solving--Data processing. I. Cannataro, Mario, 1964-

QH324.2.H3576 2009
570.285--dc22

2009001527

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

# Chapter XXV
# e-Infrastructures Fostering Multi-Centre Collaborative Research into the Intensive Care Management of Patients with Brain Injury

**Richard O. Sinnott**
*University of Glasgow, UK*

**Ian Piper**
*Southern General Hospital, Glasgow, UK*

## ABSTRACT

*Clinical research is becoming ever more collaborative with multi-centre trials now a common practice. With this in mind, never has it been more important to have secure access to data and, in so doing, tackle the challenges of inter-organisational data access and usage. This is especially the case for research conducted within the brain injury domain due to the complicated multi-trauma nature of the disease with its associated complex collation of time-series data of varying resolution and quality. It is now widely accepted that advances in treatment within this group of patients will only be delivered if the technical infrastructures underpinning the collection and validation of multi-centre research data for clinical trials is improved. In recognition of this need, IT-based multi-centre e-Infrastructures such as the Brain Monitoring with Information Technology group (BrainIT - www.brainit.org) and Cooperative Study on Brain Injury Depolarisations (COSBID - www.cosbid.de) have been formed. A serious impediment to the effective implementation of these networks is access to the know-how and experience needed to install, deploy and manage security-oriented middleware systems that provide secure access to distributed hospital based datasets and especially the linkage of these data sets across sites. The recently funded EU*

*framework VII ICT project Advanced Arterial Hypotension Adverse Event prediction through a Novel Bayesian Neural Network (AVERT-IT) is focused upon tackling these challenges. This chapter describes the problems inherent to data collection within the brain injury medical domain, the current IT-based solutions designed to address these problems and how they perform in practice. The authors outline how the authors have collaborated towards developing Grid solutions to address the major technical issues. Towards this end we describe a prototype solution which ultimately formed the basis for the AVERT-IT project. They describe the design of the underlying Grid infrastructure for AVERT-IT and how it will be used to produce novel approaches to data collection, data validation and clinical trial design.*

## 1. INTRODUCTION

Traumatic brain injury (TBI), also known as head injury, is a significant clinical problem. The incidence of severe TBI is approximately 200 patients/100,000 population with the most common causes including road traffic accidents, falls and assaults. Males are more than twice as likely to receive a severe injury than woman and currently the reported mortality rate following severe TBI ranges from less than 10% up to 50% with the most common rate quoted between 20-30%. Although the incidence of TBI is significantly less than those of other major medical diseases such as cardiovascular disease, cancer and stroke, as TBI occurs mostly in the young and the resultant morbidity is severe and long-lasting, the burden of TBI to the individual, their carers and the society that supports them is as great if not greater than the other disease domains. On a European level, fifty percent of the years individuals spend with disability are caused by brain disease of which traumatic brain injury now carries an equal burden to patients as do those of cerebrovascular and depressive illness disorders (Olesen 2003).

After ten years of pharmaceutical industry sponsored drug development and despite promising pre-clinical data, most of the clinical trials of these agents have failed to show any significant improvement in patient outcome (Narayan 2002). Many researchers feel a significant cause underlying this lack of success is the poor resolution of paper based methods for detection of adverse events and poor methods for monitoring of and controlling for protocol violations and medication errors. These limitations combine to make it difficult to detect small but clinically important treatment effects in the general noise of the brain injured patient management environment. The poor success rate of TBI clinical trials combined with the high cost to the pharmaceutical industry to conduct phase III trials in brain injury has, in recent years, caused a reluctance of the pharmaceutical industry to bring forward promising compounds to clinical trial in the field of brain injury. The high cost of conducting clinical trials is due in large part for the need to hire specially trained staff to collect and validate data. If technical solutions could be developed to reduce, even partially, the need for human resources in the data collection/validation process, potentially enormous savings could be made by these organisations. These efficiencies would ensure the organisations' longer term sustainability, and the lower running costs would reduce the cost of service delivery. Above all, this would improve the overall patient care.

This chapter focuses upon how Grid based infrastructures can help to address these issues. We focus in particular on the aspects of usability and security of Grid based e-Infrastructures and illustrate with examples from a range of projects at the National e-Science Centre at the University of Glasgow, how the vision of the Grid in providing seamless access to a range of heterogeneous resources (such as a variety of neurological data resources) can be undertaken in a secure, ethical framework where information governance and associated policy is paramount.

## 2. BACKGROUND TO BRAIN TRAUMA RESEARCH AND BRAINIT

There is increasing evidence that targeted use of IT can improve patient health care. A review article by Bates and Gawande (Bates 2003) outline several trials of IT technology that have decreased medication errors, errors of omission from poor handoffs between clinical staff and providing earlier detection of adverse events. A good example of this can be found in the work by Kupermann (Kupermann 1999) in a randomised control trial of IT for the earlier detection of adverse events which showed an 11% reduction in the time to treatment and a 29% reduction in the overall duration of dangerous adverse conditions to patients. Rosenfield and colleagues (Rosenfield 2000) reported on a study of IT based remote monitoring of a multi-bed intensive care unit and found a reduction in mortality with a reduced length of stay of nearly 30% when compared to historical controls.

Critics of this type of research will point out that better resolution of events is of no value unless their direct management influences patient outcome. Providing this type of evidence from single-centre studies with small patient numbers is not an efficient approach to answer these questions, and multi-centre studies conducting trials of new IT driven management is not readily funded nor easily justified as a research priority. Paradoxically, the patient populations which might benefit most from better IT-driven event detection and management standardisation would be the TBI population but this is also the most challenging in which to conduct such studies due to the continuing inter-centre management variation fostered to a large part by a lack of evidence for any type of effective therapy (Bulger 2002).

With this background in mind, certain individuals working within the field of TBI research met to discuss the foundation and development of a network for creating an IT based infrastructure aimed at improving the standards for multi-centre studies of monitoring and managing patients with traumatic brain injury during their acute stay in intensive care. It was agreed a different approach was needed, one which focused on using IT-based methods towards not only increasing the resolution of data capture but also the quality and validation of data captured. The pervasive nature of the internet and extensive use by clinicians of email systems fostered the creation of such a network as an internet based e-Infrastructure: the BrainIT group (www.brainit.org).

The BrainIT group works collaboratively on developing standards for collection and analyses of data from brain injured patients towards providing a more efficient infrastructure for assessing new health technology. Over a period of 12 months and four international meetings, the group have defined a core dataset designed to be collected using PC based tools and providing a common minimal dataset for all studies, regardless of the underlying research question. This data definition period was funded as part of an EC study (QLGT-2000-00454). The meetings brought together clinical and scientific experts from the domain of TBI basic research and also in the conduct of multi-centre clinical trials such as the European Brain Injury Consortium (EBIC – www.ebic.nl) as well as representatives from the medical device and pharmaceutical industries. A series of meetings and workshops spread over one year enabled the group to define a minimum set of data that could be collected from all patients with TBI, which would be useful in most research projects conducted in this population of patients. To facilitate discussion, the core dataset was sub-divided into four logical groups: a) demographic and clinical Information, b) minute by minute monitoring information, c) intensive care management information, and d) secondary insult treatment information.

From these series of meetings a consensus dataset was formed which includes nine categories:

1. **Demographic** and one-off clinical data, e.g. pre-neurosurgical hospital data, first and worst CT scan data etc;
2. **Daily management** data, e.g. daily summary measures of the use of sedatives, analgesics, vasopressors, fluid input/output balance etc;
3. **Laboratory** data, e.g. blood gas, haematology, biochemistry data etc;
4. **Event** data, e.g. nursing manoeuvers, physiotherapy, medical procedures (line insertion), calibrations etc;
5. **Surgical** procedures;
6. **Monitoring** data summary, e.g. the type and placement location of Intra-Cranial Pressures (ICP) sensors, Blood Pressure (BP) lines, etc;
7. **Neuro-event** summary, e.g. Glasgow Coma Scores (GCS), pupil size and reactivity;
8. **Targeted therapies**, e.g. mannitol given for raised ICP, pressor given for arterial hypotension etc;
9. **Vital monitoring** data, e.g. minute by minute BP, ICP, $SaO_2$ etc collected from the bedside monitoring.

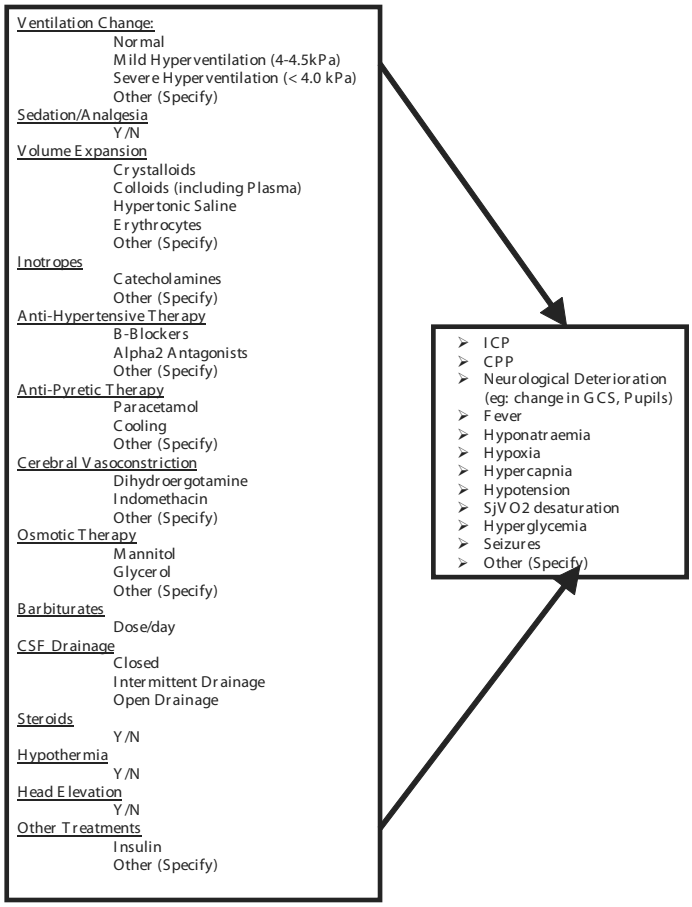
The full details of the core dataset definition and the collaboration structure of the group can be found in the BrainIT publication: BrainIT Group – core concept and data definition (Piper 2003). Unique to other dataset definitions, the BrainIT core dataset defined a special approach to quantify secondary insult management. This is medical management therapy given to patients specifically to treat secondary insults which occur to patients despite their baseline intensive care medical management. To distinguish therapy given to patients to treat secondary insults from those of baseline intensive care, we have devised a coding system which allows specific categories of therapy to be assigned a ''therapy target''. For example, arterial pressors may be given to treat systemic hypotension or to treat reduced cerebral perfusion pressure (CPP) secondary to raised ICP. Choosing an appropriate target for each secondary insult therapy will enhance the usefulness of the database on medical therapy. Each therapy must be assigned a ''Target'' chosen from a drop down list. If drugs are given then one can indicate continuous infusion if drugs are delivered by a continuous infusion pump or one can indicate boluses if it is delivered non-continuously. Figure 1 summarises the minimum choice of therapy categories and associated targets for the BrainIT core dataset. This therapy tracking model has been designed to be easily implemented in software.

It is one thing to define on paper a dataset and another to actually collect it. Although a paper based feasibility exercise established some baseline information, the acid test was still to develop a series of IT-based tools to collect the core dataset and to prospectively trial the collection of core data from a number of neuro-intensive care centres.

A three year follow up EC funded study (QLGC-2002-00160) enabled the group to develop IT methods to collect the core dataset and to assess the feasibility and accuracy for collection of this core-dataset from 22 neuro-intensive care centres. The main data collection instrument for the "episodic" non-monitoring data was a PDA based data collection tool.

In this system, clinical data is entered by bedside nursing staff on hand held PDA's which supported the BrainIT core dataset definition through a Java Struts-based tool. This allowed the core dataset to be entered by roaming research nurses using a set of PDA documents accessed via a series of buttons and tabs. With this system, indicators were present showing data documents which were fully complete, partially complete or totally incomplete. When convenient, the PDA was connected to a docking station and a client program allowed viewing and saving of patient data collected. An anonymisation routine removed patient identification elements from the collected data and labelled the patient data file with a unique BrainIT study code generated from the BrainIT

*Figure 1. Tracking therapies and targets*




web-site. A local database held in each centre linked the anonymised data to local centre patient id information which was needed during the data checking stage of the study. The multi-centre ethics approval precluded connection of the PC client system holding the data to any computer connected to the Internet. We relied on local research nurses to download the anonymised data from the PDA system client PC onto a memory stick or CD and transfer the data to an Internet connected PC for upload of the data to the BrainIT database via the website data upload page.

As this was a multi-centre study collecting data from different countries with a range of languages, a multi-language implementation was needed to foster ease of use by local nursing staff.

A training course was held for the data validation nursing staff in Glasgow on the use of this data collection instrument which also included using their medical term and language expertise to translate all PDA labels and text output into six European languages (English, French, German, Spanish, Flemish and Italian). Data could be entered in the local language, exported in an XML file format where a table lookup driven by an XSL transformation converted the data into a standard English language version.

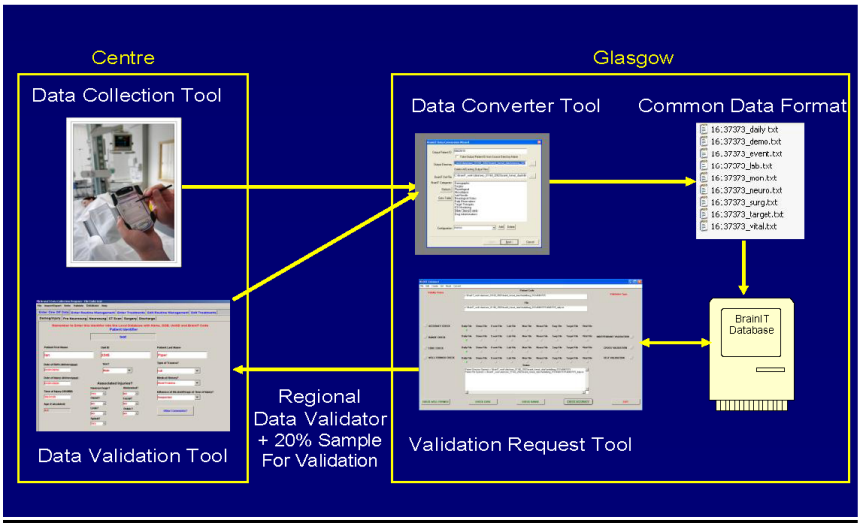
Data validation research nurse staff were hired on a country by country basis to check samples of the collected data against gold standard clinical record sources in order to quantify the accuracy for collection of the BrainIT core-dataset using

the group IT based data collection methods. Anonymised data was uploaded via the BrainIT web upload services were a server side data converter tool converted data from centre-based formats into BrainIT data format generating data category files which were imported into the BrainIT database (SQL Server 2000). A validation request tool sampled 20% of the data sent for each data category and generated a validation request file listing the timestamps and data items to be checked by local data validators. Emails were generated to the data validation staff which contained the validation data requested documents listing the data items to be checked. Data validators entered into a data validation tool the requested data items for checking from source documentation held in each local centre. Validation data was then uploaded to the BrainIT data coordinating centre via the website and using data validation checking software tools, the validated data was checked against the data items originally sent from which percentage accuracy data was calculated. Figure 2 shows the flow of data between a remote centre and the BrainIT database with the validation procedure selecting a random sample of 20% of data items uploaded per data type which are sent to data validation (DV) nurses. The DV nurses enter requested data into a PC based validation data tool, upload the data to the data manager who can then check the accuracy of data for each data category and estimate an overall error rate.

As part of this validation process, in addition to the categorical and numeric clinical data being checked for accuracy, the BrainIT system also assessed the minute by minute monitoring data too. Random samples of monitoring data channels uploaded, e.g. ICP, $SaO_2$, were selected and validation staff asked to manually enter the hourly recorded values from the nurses chart (or local gold standard data source) for the first and last 24 hour periods of bedside monitoring for a given patient for a given channel. These "validation" values could then be compared with a range of summary measures, e.g. mean, median, from the computer based monitoring data acquired from the patient.

To date (mid-2008), 384 TBI patient's core data have been collected from 22 European neurointensive centres. The first 200 patient's data has been cleaned and validation analyses conducted. In total, 19,461 comparisons were made between collected data elements and source documentation data (Shaw 2008). The number of comparisons made per data category was in proportion to the

*Figure 2. BrainIT data validation flow of information*



size of the data received for that category with the largest number checked in laboratory data (5,667) and the least in the surgery data (567). Table 1 summarises error rates by data class. Error rates were generally less than or equal to 6%, the exception being the surgery data class where an unacceptably high error rate of 34% was found.

With regard to the proportion of surgery errors, this was primarily due to the classification system used to simplify and thereby reduce the burden of data entry. As an example of this, through discussions with local nursing and data validation staff it was found that there was particular confusion over when to record ICP sensor placement and the presence of skull fractures as the primary surgical procedure. Typically, these procedures occur during the same operative procedure. As such, confusion over coding these two procedures by both the local research nurse and the data validation nurse accounted for the majority of errors in this data category.

To our knowledge, this study conducted by the BrainIT group is one of only a few projects to attempt to prospectively assess the data capture error rate within an academic environment. We have shown that it is feasible to collect the BrainIT dataset from multiple centres in an international setting with human-intensive (research nurse) IT based methods and the accuracy of the data collected is greater than or equal to 94%, with the exception of the surgery data type which must be revised. We have also shown that computer collected minute by minute vital signs data, summarised as end hour averages, correlate well with nursing chart end hour recordings. This allows the end hour averaged computer records to be used in database analyses assessing nurses chart recorded detection of events with computer based sampling. These validation results calculated on a subset of patients provides an estimate of the data quality for future analyses on the full patient cohort of 350 patients collected as part of the EC funded study which was conducted over the same time period by the same staff using the same data methods. Clearly though, future data collection projects will generate datasets under differing data collection conditions and will require a separate validation stage if we wish to maintain our confidence in the level of data accuracy. However, the costs of maintaining such a data validation network are prohibitively high. To maintain a full time data validation nurse within each participating country costs in excess of 1 Million Euro's per year. Such large running costs for an academic network are not sustainable in the long term and a more cost-effective solution for data validation must be found.

*Table 1. percentage error rate by data type class with common error types*

| Data Class | Error Rate (%) | Common Errors |
|---|---|---|
| *Laboratory* | 2 | *pCO2, FiO2 value* |
| *Demographic* | 4 | *Monitoring on arrival at neurosurgery, intubation on arrival at neurosurgery* |
| *Neuro Observations* | 5 | *Pupil Size, GCSv (code 1 Vs Unknown)* |
| *Monitoring Summary* | 5 | *ICP type, ICP Location* |
| *Daily Management Summary* | 5 | *Infusion type (bolus vs infusion or both), drug number (1, > 1)* |
| *Targeted Therapy* | 6 | *Non-standard target, no Target specified* |
| *Surgeries* | 34 | *ICP placement, Skull #, mass lesion* |

## 3. GRID-BASED SOLUTIONS TO BRAIN TRAUMA RESEARCH

Grid technology provides middleware that can allow distributed federation of data to occur. Different domains have their own requirements on how this federation can be achieved. The clinical domain in particular demands that strict adherence to ethics and information governance is achieved which in turn demands fine grained security.

It is the case that many IT and data storage solutions already exist crossing primary care, secondary care and a range of specialised resources such as disease registries in the clinical domain. Many of these solutions have been developed largely in isolation and as a result have widely differing data descriptions and associated security policies – or in many cases, no security policies other than protection at the firewall level provided through bodies such as the NHS in the UK. This situation is greatly magnified when crossing national boundaries. Dealing with such heterogeneity from the data perspective at least has been one of the drivers behind Grid technologies.

Due to the sensitivity of data, the establishment of security policies and their rigorous enforcement is of paramount importance in the clinical domain. It is clear that a single static system accessing a closed/fixed set of data will not meet the needs of healthcare providers nor researchers using the clinical data sets. Systems and data sets evolve. Different researchers may be allowed access to a given clinical data set for a given time after applying specifically for ethical approval to do so. Instead, software and tools are required to build infrastructures where a variety of data can be made accessible to different individuals for different times for different reasons. A cornerstone of these solutions is in ensuring fine grained security. However it is an unavoidable fact that the specific privileges required in a particular trial or study will not be known when the system is first created. Similarly a doctor in one hospital may have privileges to access various systems in that hospital, but these do not transfer directly when this doctor attends a different hospital for example. Therefore systems capable of adding and removing resources or privileges *"on the fly"* are necessary, where the corresponding allocation of privileges can be added or removed depending on the needs of different trials or healthcare systems. In Grid parlance, the framework by which such rules and regulations on the resources and the users that may access them and under what conditions is given by the concept of a *Virtual Organisation* (VO).

The basic models put forward for the majority of Grid-based security systems can be broadly broken down into the "AAA" categories:

- *Authentication* – establishing the identity of the person requesting access to a resource.
- *Authorization* – having established identity, establishing and enforcing what that person is allowed to do on a given resource.
- *Accountability* – being able to establish the activities, and time of activities, of a particular person on that resource (or resources) so that they can not subsequently deny potential misuse later on (non-repudiation).

Of course there are other important aspects when considering the wider challenges of building secure systems. For example in the clinical domain, confidentiality and data integrity are essential, but in this chapter we restrict ourselves to authentication and authorisation as these are arguably the most important things to get right in the first instance. Put another way if authentication and authorisation are not adequately addressed, then other aspects of security are largely redundant

Methods of authentication with grid technology tend to favour two methods, either username/password combinations or public key infrastructures (PKIs) (Housley 2001). The latter are used to set up and use safer, encrypted communication

channels through trusting a third party root of trust Certification Authority (CA) such as the UK e-Science CA (www.grid-support.ac.uk/ca). Authentication is focused upon identity management and a process exists through which a user establishes their identity to obtain an X.509 digital certificate (ITU-T X509, 2001). Typically this is through showing some form of physical identification to a local registration authority at their institution. However there are issues with this process, not least of which is the complexity of converting certificates to formats suitable for usage on the Grid and with the lack of local identity management. Furthermore, a user might be expelled from their institution but still have access to a valid Grid certificate. These issues are described in more detail in (Sinnott, 2006; Stell, 2006; Watt, 2006; Ajayi, 2006).

One approach to overcoming the issues with PKI based identity management is through exploitation of local identity management systems. The Internet2 Shibboleth technology has been developed to fulfil this need (Internet2 ShibArch, 2006, Internet2 ShibProt, 2006). In essence, Shibboleth provides a method of securely transferring attributes between institutions subscribing to an over-arching federation. The basic model of Shibboleth is that institutions enter into federations, or more precisely federated access management federations. Users at those institutions attempting to access remote resources across the federation (typically service providers) are redirected to their home institution (typically through a Where Are You From service) where they log in locally. A digitally signed Security Assertion Markup Language (SAML) (OASIS SAML, 2003) assertion showing that the user has authenticated is then delivered to the target resource which may then decide whether access is granted or not. Often further information such as attributes for authorisation need to be returned. This whole process however is transparent to the end users who only log in to their local system with the normal usernames and passwords.

The key benefit of Shibboleth is that end users have simple ways to access resources. Furthermore depending upon local policies and trust relationships across the federation, users are able to access a range of distributed resources thereby supporting *single sign-on*. Once a secure session between a user and a resource is established, the user can access further resources (in the same federation) without the need to further authenticate – based upon browser-based information. Additionally since their authentication is tied to their local institution, they will have their federation privileges revoked, if they are revoked locally (which is the most likely scenario).

Authentication is only a starting point with regard to security however. Knowing that someone has authenticated at the University of Glasgow is not likely (indeed very unlikely) to allow access to a remote clinical data set. Instead authorisation is required. Unlike authentication which has a reasonably mature body of software and approaches in supporting the process of its application, authorisation is a much more fluid area. Numerous competing software solutions and standards exist with their own advantages and disadvantages. A comparison of these and overview of some of their advantages and disadvantages is given in (Sinnott, 2005; Stell, 2005). It is essential in the clinical domain that any solution put forward for security is simple both for the end users, but equally importantly for the clinical data provider IT-administrators. These will typically be unlikely to have any experience with Grid-based systems. As such, solutions are required which provide them with capabilities to easily manage secure access to their local data according to agreed information governance policies.

One of the more mature authorisation-based approaches is based upon role based access control (RBAC) (Sandhu, 1996). In this model, roles are created which offer a given privilege and subsequently assigned to trusted individuals. Several approaches for creation of these roles and their assignment are possible including federated

models, centralised models or hybrid approaches of the two. The pros and cons of such role based creation and assignment is described in (Sinnott, 2008).

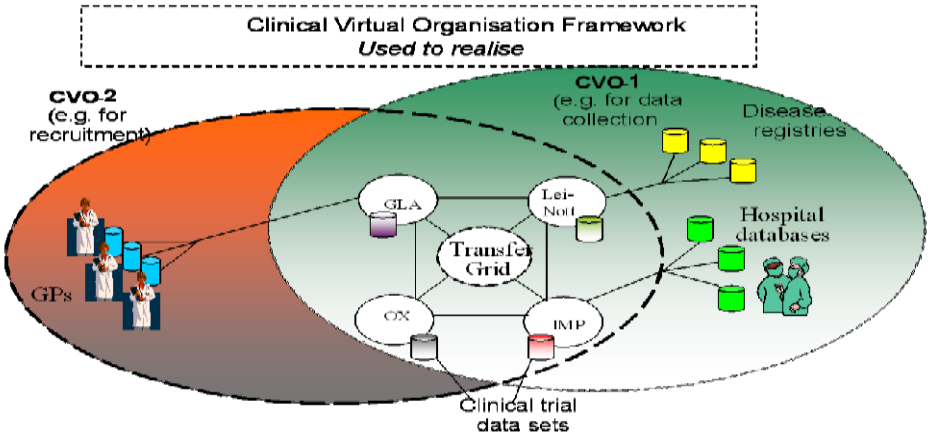
Once these roles are created and assigned to individuals they can be used to enforce finer grained authorisation at given data provider sites. To actually enforce these authorisation decisions, several software solutions can be adopted. The PERMIS RBAC (Chadwick, 2002) offers one such solution for role definition, assignment and its use in enforcing local access control. Anther popular solution is Virtual Organisation Membership Service (VOMS) (Alfieri, 2003) which provides a centralised repository of VO-roles and has been shown to interoperate with PERMIS in (Chadwick 2002). An essential part of authorisation is in understanding both when and how an authorisation decision needs to be made. The X812 standard (ITUT X812, 1995) defines concepts such as policy enforcement points (PEP) and policy decision points (PDP) which define a generic approach for deciding where (PEP) and how (PDP) authorisation decisions should be made. Trust plays a crucial role in any security system however, and the allocation of privileges and their use for authorisation needs to be augmented with user guidelines and agreements on best practices on accessing and using clinical data according to ethics and information governance.

To understand how these security concepts can be applied in the brain trauma domain, the National e-Science Centre (NeSC) at the University of Glasgow has undertaken a range of case studies exploring authorisation and Shibboleth technologies applied in the clinical domain. This work was undertaken as part of the Joint Information Systems Committee (JISC) funded GLASS (www.nesc.ac.uk/hub/projects/glass) and the Medical Research Council funded VOTES (www.nesc.ac.uk/hub/projects/votes) projects.

The VOTES project focused on realising a software solution meeting the challenges inherent in federating distributed clinical data, specifically in supporting the various stages of clinical trials and epidemiological studies such as patient recruitment; data collection and study management. The GLASS project was primarily focused upon the roll out of Shibboleth technology across the University of Glasgow and exploring its suitability in a range of e-Science and non-e-Science applications, e.g. to provide access to student records for authorised individuals in Glasgow.

Usability is at the heart of VOTES efforts and portal based solutions have been prototyped with this in mind. As identified previously, this usability should be for end users, administrators, investigators and the other stakeholders involved in clinical trials and studies. To support this, a

*Figure 3. VOTES node infrastructure and clinical VO overlays*



node infrastructure supporting the addition or removal of institutions and the resources that they wish to contribute is needed. This should support fine grained user-oriented VO-specific security. Figure 3 shows an overall schematic representation of such a VO.

In this infrastructure, each node provides access to its resources according to local policies. A key part of this infrastructure is that it supports both heterogeneous data resources but equally importantly heterogeneous security infrastructures. Thus sites may recognise certain roles associated with a given clinical trial or not as the case may be. Furthermore, these roles and the privileges that are associated with them will have different interpretations at each site. Thus a clinical nurse role at one node might be allowed to query a range of clinical data including identifying data, whilst the same nurse role at a different site may be restricted to only accessing statistical information, e.g. the number of patients with a particular Glasgow Coma Score without knowing any further information on those patients.

The VOTES infrastructure supports a variety of Grid components to support secure access to federated clinical data. These include Grid services developed with the Globus toolkit version 4 (www.globus.org) and data federation technology such as the Open Grid Service Architecture Data Access and Integration (OGSA-DAI) technology (www.ogsadai.org.uk). The interface to the systems themselves is through portal based technologies based upon GridSphere (www.gridsphere.org).

To understand the interplay of role based access to federated resources we consider the scenario of a particular BrainIT trial undertaken with Glasgow Southern General Hospital. In this model we identified different roles that could access different demographic, physiological and monitoring data. For simplicity we simply assigned these roles directly to known and trusted individuals at Glasgow. This assignment was based upon adding the appropriate signed attribute certificates to the appropriate individuals in the Glasgow Identity Provider (IdP) given as an LDAP server although other authentication systems are possible. Extensions and refinements to this scenario allowing for example the delegation of these privileges is also supported through the JISC funded DyVOSE project (www.nesc.ac.uk/hub/projects/dyvose). In this model, the privileges (roles) can be pushed to remote trusted administrators who can subsequently assign them to local staff or students involved in that particular VO for example. Similarly, adding the attributes to separate attribute authorities (AAs) is supported rather than to a *single* institutional IdP. In this case, a user would log in to their IdP and the appropriate attributes linked and pushed (or pulled) from separate AAs set up for specific VOs. This linkage is possible through having unique identities in place across institutions and is described in more detail in (Sinnott, 2007; Watt, 2007).

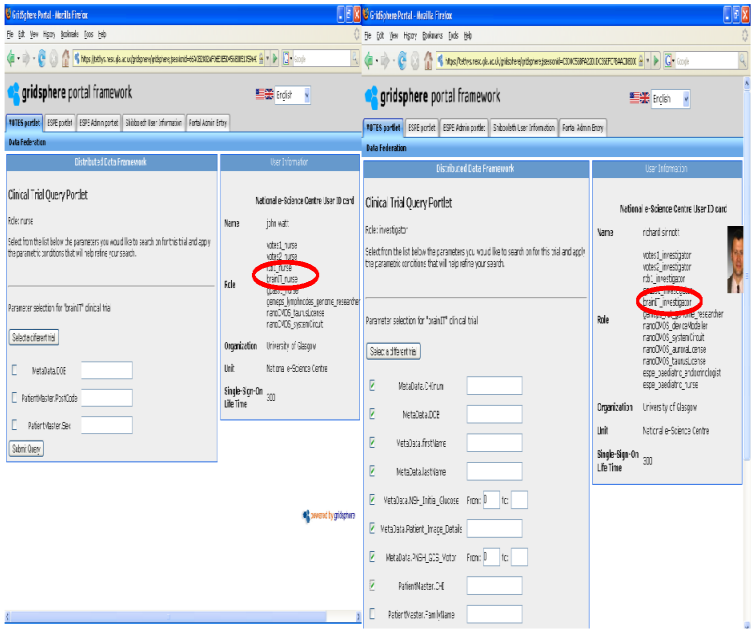
When a user attempts to access the BrainIT portal, they are redirected to their home institutions when they are asked to authenticate. After authenticating, the attributes needed for access to the BrainIT portal for authorised access to the federated clinical data are returned in a signed SAML assertion. We note that the tools for scoping of attributes, e.g. which sites to trust, and for defining and enforcing attribute release policies, e.g. to determine which attributes should be released to which sites have been implemented in the SPAM-GP project (www.nesc.ac.uk/hub/projects/spam-gp). The portal will then use these attributes to configure the contents of the portal, i.e. the users are restricted to see and do what their roles dictate. This is shown in Figure 4 where two different users have logged in to the system - one with the *brainit-investigator* role (right side) and one with the *brainit-nurse* role (left side). The former role has superior privileges in this particular trial and hence is allowed access to a wider range of information within the context of that trial, e.g. the GCS, the ICP, Glucose levels as well as identifying patient information. The nurse

role on the other hand is restricted to a subset of non-patient identifying information.

For clinical data providers, it is unlikely that they would simply allow access to their resources based upon a role which has been used to configure a portal where the portal itself may well exist outside of their own domain of control. The VOTES e-Infrastructure has been designed with security throughout however. This includes security of the local Grid and data services as well as security at potentially remote data providers so that they are able to make their own autonomous authorisation decisions. Thus the roles are used to configure the portal, but then importantly the distributed data providers accessible from this portal have their own local authorisation policies which must be satisfied by the information provided to the portal. It is quite possible that further information than the supplied role needs to be obtained before a given access decision is made. This might be a notification from an ethical oversight committee for example. The infrastructure allows for a range of authorisation possibilities to be supported, including scenarios where additional attributes from one or more attribute authorities are required to authorise specific requests. Examples of these different kinds of model are described in (Stell, 2008).

We note that data security offers challenges that existing Grid security infrastructures are not well aligned to and hence cannot be easily supported. In large scale federated data infrastructures, it is often the case that access to individual subsets of data within different federated repositories is required. Many mainstream Grid based solutions today work primarily at securing services or methods that services support. Whilst it is possible to develop targeted services to specific data resources for a nurse or an investigator, a more scalable solution is to have one or more generic services that differ in the security models that they offer. Thus, to support a range of studies it should not be necessary to support individual services at each service provider for each study. Rather, the ideal scenario is to have a single data access service, which can distinguish between the different trial/study specific security identifiers. To support such scenarios, the VOTES project has

*Figure 4. Role-based access to portal and restrictions to data access*



implemented RBAC models based upon access control matrices. With these models, users and roles are defined with specific relationships over the data models themselves at each data provider. Thus, a given data provider may decide that a subset of their tables should be made available with specific columns and rows accessible for different roles within a given study. Through tools and services implemented in VOTES it is possible to associate those users and their roles with specific subsets of the databases. Details on how this is supported, is described in (Sinnott, 2008) along with the benefits and drawbacks compared to other RBAC systems. This model is based on the premise that the access matrix will be available at every node in the VO and will be regularly updated. Consequently, every user that has access to the VO in some form will go through this matrix model to access any other resource, be it local or remote, within the VO with each data provider defining its own local access policies for each role.

Having defined the roles and local policies on how these are interpreted when enforcing policy decisions it is necessary to consider the ramifications of the data sets that are made available through the combination of the policies. Thus a given provider might be completely satisfied that the data sets that they release are in accordance with local policy and information governance frameworks. However, when linked or joined with other sets, the consequences need to be understood within the context of the VO. For example, a provider may be happy to release anonymised data but if this is linked with other data sets containing identifying data for given patients for example, then this needs to be clearly understood by all stakeholders across the VO. We note that it is also possible to link data and remove (anonymised) the fields that have been used for the data linkage. The algorithms for supporting these kinds of scenarios are described in detail in (Ajayi, 2008).

The typical interactions of the components involved in supporting the access to and usage of federated resources is as follows. Firstly the user must log into the portal either through Shibboleth or directly. In the Shibboleth case the various user roles and privileges are provided to personalise/authorise their access to a variety of services as described above. To achieve this, the portal server checks the local resource files to discover the available grid servers, data servers and associated databases accessible to this particular user with the associated privileges. The appropriate Grid service consults the local access matrix and returns the parameters for the resources that the user can query for that particular trial. These are presented to the user as a list of check-boxes, with the option to specify conditions if desired (as shown in Figure 4). Following this, the user makes their selection of parameters and submits them. These are constructed into a query which is distributed, i.e. federated, across the various resources associated with that particular trial. To support this query distribution, the query is sent from the Grid server to the data server, where it is wrapped as an OGSA-DAI service request before being passed to a local driving database. This driving database subsequently co-ordinates the execution of the distributed queries over the resources associated with that trial. Where possible, for example where web services are offered and no-direct connections to a remote database are offered or possible, the federating of the sub-queries can result in combinations of direct database queries through JDBC for example, as well as SOAP messaging to remote services. Once all sub-queries are returned the results are joined together and returned for display in the portal after conversion to HTML. We note that it is quite possible for the results to be stored into databases for future access by researchers. The actual joining of results itself is made possible through the unique identifier assigned to patients in Scotland – the Community Health Index number. This greatly simplifies the development of federated queries

since a unique joining key is available. In other situations, this is not always available hence other mechanisms and identifiers have to be identified through which data linkage can be made. This might be based upon study specific identifiers generated for particular patients. With this approach, joining of data within a given neurological centre can be achieved however joining across centres is not possible. Since it is unlikely that the same individual will be found in different neurological data centres, this is not overly constraining. Rather in this case queries are federated to the different centres, the data joined on local identifiers, and then returned to the portal where the union for example of the data sets can be made. The result of a particular query is shown in Figure 5.

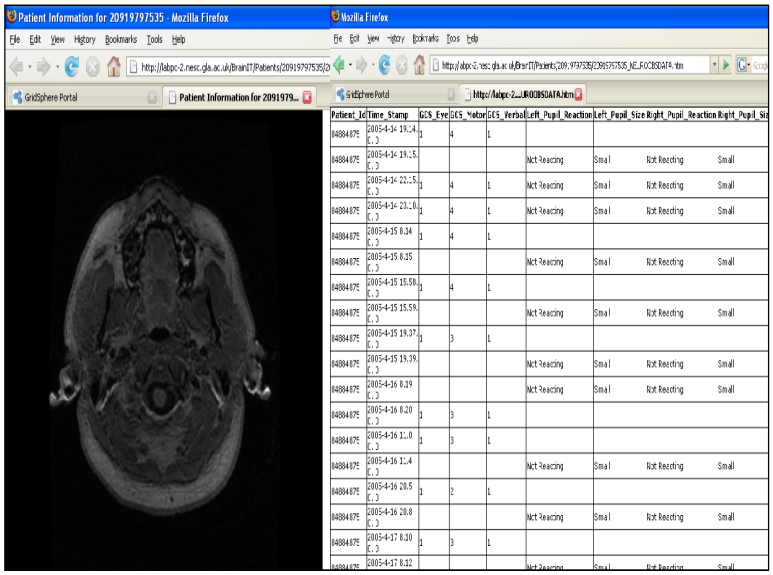
In the results shown in Figure 5 we see the DICOM images and various physiological data associated with a given patient that includes any changes in their condition or their overall Glasgow Coma Score in response to the specific treatments they may be receiving. A variety of other information can be returned in the existing systems including the current treatment, ICP measurements amongst other factors.

The above systems were developed in a controlled test-bed at the NeSC in Glasgow. This comprised a variety of SQL Server, MySQL and Oracle databases hosting the various neurological data sets as well as a consent database.

## 4. FUTURE TRENDS

The above systems have shown the proof of concept upon which clinical data can be accessed and used within a secure framework. There are numerous enhancements to this system that are being explored within the AVERT-IT project. We outline upon some of these here since we believe they capture future trends in Grid-enabled brain trauma research. Firstly, it is clear that the models and systems described here need to deal with the actual concerns regarding access to and usage of live systems supporting multi-centre trials. Proof of concept VOs and federated authorization systems as described here need to reflect the particular needs and requirements of particular neurological data centres. In particular it is essential to understand what data needs to be made accessible and

*Figure 5. Results of executing a particular brain trauma query*



linked across the various centres and importantly for what reason. Within the AVERT-IT project we propose to capture near real time information from the various neurological data centres to explore the potential prediction of hypotensive events. To support this, we are exploring how determination of hypotensive event might be predicted via empirical analysis, univariate analysis and multivariable linear regression techniques of the weighted association between multiple patient parameters (drawn from demographic, periodic and episodic datasets) and subsequent arterial hypotension. This association will then form the basis for the initial definition of a novel Bayesian Neural Network, to be trained against the BrainIT dataset prior to undertaking a novel clinical trial demonstrating the effectiveness of the AVERT-IT project concept.

From a technological perspective, the AVERT-IT system will focus on the development of a novel IT-based decision support system ("HypoPredict"), appropriate for deployment within intensive and high dependency care units which is capable of:

- Automatically and continually monitoring at least four in-vivo patient parameters (ECG, arterial blood pressure, Oxygen saturation and core temperature), together with open interfaces providing input of key demographic data (age, gender etc.) and periodic data (clinical pathology results etc.) related to the patient.
- Outputting a continuous Hypotension Prediction Index (HPi) in the range 0 – 100 (0 = no risk; 100 = patient is currently hypotensive);
- Providing primary (P1) and secondary (P2) weighted (0-100; 0 = not considered relevant; 100 = critical importance) causal data (current values of input parameters) in parallel with HPi to facilitate appropriate intervention selection by clinician (for example, elevated core temperature could be indicative of sepsis, a common precursor to hypotension);
- Providing updated HPi, P1 and P2 values immediately changes are detected in the patient parameter input set.

Such predictive approaches offer one of the greatest opportunities from the application of Grid middleware to link different clinical data centres. Can we determine the best way to treat individuals suffering major brain trauma thereby extending the state of the art knowledge in brain trauma treatment from initial assessment, diagnosis and treatment to continued software support systems which augment the whole process of improved healthcare and treatment. We emphasis here that these systems are not and are never intended as a replacement for informed clinical decisions in brain trauma patient management, but as a guide and tool that helps intensivists and brain trauma specialists better understand the whole course of brain trauma treatment with improved monitoring and diagnostics.

However from past and on-going experience in projects such as VOTES, we recognise the direct transfer of Grid technology within a clinical setting often requires a degree of pragmatism and especially consideration of the clinical IT staff and their existing systems in place. As noted it is unrealistic to expect a hospital IT administrator to be knowledgeable in Grid and/or advanced authorisation technologies. Furthermore, one of the major issues with most Grid middleware today is their lack of stability. This in turn directly impacts upon the validation of the software. Before widespread take up of Grid middleware in the clinical domain can be achieved, software validation is needed. This has to ensure both that the middleware satisfies the basic tests that it is fit for purpose, but also that it is resilient enough to withstand attempts at breaking the software from third parties, e.g. hackers. This has not been the focus until now, but is clearly needed for the wider community to have faith. To address such

aspects within the AVERT-IT systems we are initially proposing models based upon pushed data transfer only, i.e. the clinical data provider firewalls will reject all incoming data requests, but will push data out of their firewalls into demilitarised zones set up as a buffer between the Grid research domain and the domain of live healthcare provision. It is our intention, however to eventually develop systems that allow to integrate the various neurological data sets within a given clinical data provider domain and push them out of the firewall. In supporting this, we are acutely aware of the pragmatic approach that is needed and are not yet proposing any overly complicated middleware solutions, but instead focusing upon lighter-weight data integration clients that can query various clinical data resources directly, e.g. over JDBC or ODBC connections before anonymising, encrypting and ultimately sending them to the centralised AVERT-IT repository accessible to the partner sites. We are currently in the progress of evaluating the different data models and software used within the individual partner sites and how these map onto the BrainIT core data set identified previously so that the queries to extract the needed data sets can be defined and implemented in the lightweight clients. This is non-trivial given the international dimension of the AVERT-IT project which involves collaborators from the UK, Sweden, Italy, Germany and Lithuania with each site having their own software and data infrastructures.

One challenge that remains to be solved in the AVERT-IT system and within the wider e-Research community more generally is with regard to real time or near real-time data. Monitoring data that needs to be streamed to support the Bayesian adverse event prediction algorithms offers new challenges that have hitherto not been addressed within the VOTES project for example. In this case, it is likely that data streaming models with authorisation capabilities are supported, or enhancements to systems that allow for frequent periodic queries to be undertaken with periods of the order of 5 minutes or less being likely. This is made further challenging since these near real-time data sets need to be joined with other non-real time data sets each time. Obviously to predict the onset of an adverse hypotensive event, the closer to real time the information is the better and more accurate the identification will be and hence less false positives identified.

These challenges are not specific to the brain trauma domain but apply more generally to wider healthcare support where real time information to patient information is essential. However real time access to patient data even by clinical healthcare providers is often a fraught process as exemplified by the recent furore in the UK to establish a national consent database (Guardian, 2007). It is interesting to note here that many patients were largely in favour of such national data resources for improved healthcare, whereas general practitioners and trusts where more reluctant to make their patient data available to national-level resources.

## 5. CONCLUSION

The Grid paradigm provides a compelling model for secure access to clinical data and the proof of concept systems have shown that this vision of seamless access to federated data can be made a reality. We recognise that a proof of concept system and a live system that has been validated requires a step change in the way in which Grid based systems are currently developed. The Grid middleware evolution or revolution in many cases needs to solidify to hardened software stacks that clinical and healthcare providers understand and are comfortable in supporting. Indeed the Grid community as a whole need to agree upon best practice of establishing the numerous different flavours of Grids that exist. Why should software developed for distributing petabytes of elementary particle data from the Large Hadron Collider in CERN be applicable to accessing a hospital data-

base in Scotland? Surprisingly, such applications of Grid middleware are still being explored. From direct experience of working with healthcare providers and their IT support staff, the deployment of complex open source Grid middleware stacks from a variety of sources will never be accepted. It is essential that in future health Grid efforts this is recognised and a common core set of functionality agreed upon, documented and the detailed pro's and con's of such software defined for others to use. It is the case that the weakest link in any system is the one that will be exploited. Ensuring that all nodes within a particular VO understand the consequences of being in that VO for all parties is crucial. Similarly, whilst compromises of Grid facilities can be accepted to a degree in some domains, the healthcare domain is not one of them. The reputation of the Grid is perhaps as important to convince healthcare providers to provide access to their resources via the Grid as are the Grid technologies themselves. The authentication-only based models of most large scale Grids such as the National Grid Service (www.ngs.ac.uk) in the UK will not convince these providers – thus it will never be the case that a researcher is allowed access to a clinical provider database to "*do stuff*". Rather, clearly defined rules and regulations on what can be accessed and used and in what context are needed, and Grid technologies need to *demonstrably* support this process.

## 6. REFERENCES


Ajayi, O., Sinnott, R. O., Stell, A. J. & Young, A. (2008). *Towards a Virtual Anonymisation Grid for Unified Access to Remote Clinical Data*. Proceedings of 6th International HealthGrid conference, Chicago.

Ajayi, O., Sinnott, R. O., Watt, J., Jiang, J. & Stell, A. J. (2006). *Single-Sign on and Authorization for Dynamic Virtual Organizations.* Proceedings of International Conference on Virtual Enterprises, (PRO-VE'06), Helsinki, Finland.

Alfieri, R., et al (2003). *VOMS: an authorization system for virtual organizations.* Proceedings of 1st European across Grids conference, Santiago de Compostela, Spain.

Bates, D. G. A. (2003). Improving safety with information technology. *New England Journal of Medicine, 348*, 2526-2534.

Bulger, E. & Nathens, A. (2002). Management of severe head injury: institutional variations in care and effect on outcome. *Critical Care Med, 30*, 1870-1876.

Chadwick, D. W., Otenko, O. (2002). The PERMIS X.509 Role Based Privilege Management Infrastructure. *Future Generation Computer Systems, 936*, 1–13.

Guardian (2007). *Family doctors shun national database of patients' records, 20 Nov 2007.* Retrieved from www.guardian.co.uk/society/2007/nov/20/nhs.health

Housley, R. & Polk, T. (2001). *Planning for PKI: Best Practices Guide for Deploying Public Key Infrastructures*. Hoboken, NJ: Wiley Computer Publishing.

Internet2 Shibboleth, (2006). *Shibboleth Architecture Protocols and Profiles*. Retrieved from http://shibboleth.internet2.edu/docs/draft-mace-shibboleth-arch-protocols-latest.pdf

Internet2 Shibboleth, (2006). *Shibboleth Architecture Technical Overview.* Retrieved from http://shibboleth.internet2.edu/docs/draft-mace-shibboleth-tech-overview-latest.pdf

ITU-T Recommendation X.509 (2001) | ISO/IEC 9594-8: 2001, Information technology – Open Systems Interconnection – Public-Key and Attribute Certificate Frameworks.

ITU-T Recommendation X.812 (1995). | ISO/IEC 10181-3:1995, Security Frameworks for open systems: Access control framework.

Kuperman, G. J. (1999). Improving response to critical laboratory values with automation: results of a randomised controlled trial. *J Am Med Inform Assoc, 6*, 512-522.

Narayan, R. M. (2002). Clinical Trials in Head Injury. *Journal of Neurotrauma, 19*, 503-557.

OASIS (2003). *Assertions and Protocol for the OASIS Security Assertion Markup Language (SAML) v1.1.* Retrieved from http://www.oasis-open.org/committees/security/

Olesen, J. L. M. (2003). The burden of brain disease in Europe. *European Journal of Neurology, 10*, 471-477.

Piper, I., Citerio, G., & the BrainIT Group (2003). The BrainIT Group: Concept and Core Dataset Definition. *Acta Neurochirurgica, 145*, 615-629

Rosenfield, B. A. & Dorman, T. (2000). Intensive care unit telemedicine; alternative paradigm for providing continuous intensive care. *Crit. Care Med, 28*, 3925-3931.

Sandhu, R., Coyne, E. J., Feinstein, H. L. & Youman, C.E. (1996). Role-Based Access Control Models. *IEEE Computer 29* (2), 38-47.

Shaw, M. & Piper, I. (2008). The Brain Monitoring with Information Technology (BrainIT) collaborative network: Data Validation Results. *Acta Neurochirurgica, ICP Symposium Journal*.

Sinnott, R. O., Jiang, J., Stewart, G., Stell, A. J., Martin, D., Doherty, T., Su, L. & Watt, J. (2008). *Advanced Security for Virtual Organizations: Exploring the Pros and Cons of Centralized vs Decentralized Security Models.* Proceedings of 8th IEEE International Symposium on Cluster Computing and the Grid (CCGrid 2008), Lyon, France.

Sinnott, R. O., Stell, A. J., Chadwick, D. W., & Otenko, O. (2005). Experiences of Applying Advanced Grid Authorisation Infrastructures. In P.M.A. Sloot, et al, (eds.), *Proceedings of European Grid Conference (EGC)*, Amsterdam, the Netherlands (pp. 265-275).

Sinnott, R. O., Stell, A. J., Chadwick, D. W., Otenko, O., Watt, J., Koetsier, J. & Nguyen, T. A. (2006). *Supporting Decentralized, Security focused Dynamic Virtual Organizations across the Grid*. Proceedings of 2nd IEEE International Conference on e-Science and Grid Computing, Amsterdam, Holland.

Sinnott, R. O., Watt, J., & Jiang, J. (2007). *The GLASS Project: Supporting Secure Shibboleth-based Single Sign-On to Campus Resources.* Proceedings of UK e-Science All Hands Meeting, Nottingham, UK.

Sinnott, R. O., Watt, J., Ajayi, O. & Jiang, J. (2006). *Shibboleth-based Access to and Usage of Grid Resources.* Proceedings of IEEE International Conference on Grid Computing, Barcelona, Spain.

Stell, A. J., Sinnott, R. O. & Ajayi, O. (2008). *Supporting Nationwide e-Clinical Trials and Studies.* Proceedings of 15th Mardi Gras Conference, Baton Rouge, LA.

Stell, A. J., Sinnott, R. O. & Ajayi, O. (2006). S*ecure, Reliable and Dynamic Access to Distributed Clinical Data*. Proceedings of Life Science Grid Conference, Yokohama, Japan.

Stell, A. J., Sinnott, R. O., & Watt, J. (2005). *Comparison of Advanced Authorisation Infrastructures for Grid Computing.* Proceedings of International Conference on High Performance Computing Systems and Applications, Guelph, Canada.

Watt, J., Sinnott, R. O., Ajayi, O., Jiang, J. & Stell, A. J. (2006). *User Oriented Access to Secure Biomedical Resources through the Grid.* Proceedings of Life Science Grid Conference, Yokohama, Japan.

Watt, J., Sinnott, R. O., Jiang, J., Stewart, G., Stell, A. J., Martin, D. & Doherty, T. (2007).

Federated Authentication and Authorisation for e-Science. In *Proceedings of APAC conference*, Perth, Australia.

## KEY TERMS AND DEFINITIONS

**Adverse Hypotensive Event:** Hypotension lasting 5 minutes or longer falling below the commonly accepted threshold of 90 mmHg systolic OR 70 mmHg mean pressure.

**Authentication:** The act of establishing or confirming something or someone as authentic. In the security domain this might for example involve electronically confirming the identity of a person wishing to access a given software or hardware resource. Authentication can be achieved in many ways, e.g. usernames/passwords, certificate based systems etc.

**Authorisation:** The process of restricting access to resources only to those permitted to use them. In the security domain this will typically entail the definition of security policies associated with resources, the assignment of privileges to individuals that should be able to access those resources and the subsequent enactment of those policies when requests for access are received by individuals. Standards have been identified to support both the definition of policies, where access decisions need to be enforced, and how such decisions are made. Authorisation typically augments authentication and allows finer grained access control to be supported.

**Certificate Authority (CA):** An entity which issues digital certificates for use by other parties (including individuals and computers). Through trusting a CA and the process by which it issues and revokes certificates, the certificates can be used for accessing multiple resources seamlessly to support one of the basic tenets of Grid: single sign-on.

**Glasgow Coma Score (GCS):** Widely used clinical score assessing brain stem function following a suspected brain injury.

**Hypotension:** Abnormally low blood pressure which can be especially dangerous for brain injury patients.

**Pressor:** Pharmacological agent which acts to increase the work of the heart and often includes actions on increasing systemic vascular resistance towards raising blood pressure when hypotensive.

**Registration Authority (RA):** Typically an individual at an institution that supports the processes required by a CA to verify the identity of individuals applying for digital certificates. Typically this is through the certificate requestor presenting in person a visual identity such as a passport or student matriculation card to the RA.

**Single Sign-On:** The ability to securely access and use a variety of distributed resources without the need for multiple usernames/passwords or authentication challenge/responses. In the Grid world this is typically achived through trust of the CA that issued the certificate and local policy on whether that individual with that certificate is allowed access. For many Grids, this can be through a mapping of the Distinguished Name associated with the certificate to a local system account.