Human Mutation
OFFICIAL JOURNAL
HGVS
HUMAN GENOME VARIATION SOCIETY
www.hgvs.org

BioGrid Australia Facilitates Collaborative Medical and Bioinformatics Research Across Hospitals and Medical Research Institutes by Linking Data from Diverse Disease and Data Types

Robert B. Merriel,[1,2]* Peter Gibbs,[1–3] Terence J. O'Brien,[1,4] and Marienne Hibbert[4,5]

[1]*Melbourne Health, Melbourne, Australia;* [2]*BioGrid Australia Ltd, Melbourne, Australia;* [3]*Ludwig Institute for Cancer Research, Melbourne, Australia;* [4]*Department of Medicine, The Royal Melbourne Hospital, The University of Melbourne, Melbourne, Australia;* [5]*Victorian Partnership for Advanced Computing, Melbourne, Australia*

For the HVP Bioinformatics Special Issue
Received 17 September 2010; accepted revised manuscript 14 December 2010.
Published online in Wiley Online Library (www.wiley.com/humanmutation). DOI 10.1002/humu.21437

ABSTRACT: BioGrid Australia is a federated data linkage and integration infrastructure that uses the Internet to enable patient specific information to be utilized for research in a privacy protected manner, from multiple databases of various data types (e.g. clinical, treatment, genomic, image, histopathology and outcome), from a range of diseases (oncological, neurological, endocrine and respiratory) and across more than 20 health services, universities and medical research institutes. BioGrid has demonstrated an ability to facilitate powerful research into the causation of human disease and the prediction of disease and treatment outcomes. BioGrid has successfully implemented technology and processes that allow researchers to efficiently extract data from multiple sources, without compromising data security and privacy. This article reviews BioGrid's first seven years and how it has overcome 9 of its top 10 challenges.
Hum Mutat 32:1–9, 2011. © 2011 Wiley-Liss, Inc.

KEY WORDS: BioGrid; bioinformatics; data linkage; Bio21; Molecular Medicine Informatics Model; MMIM; ACCORD

Introduction

Bio21 Australia Limited (Bio21; http://www.bio21.org.au) is a biomedical, biotechnology research cluster of universities, tertiary health services and medical research institutes supporting collaborative projects and shared technology platforms. In 2001, Bio21 identified bioinformatics as a priority area of research for the coming decade. Melbourne Health (http://www.mh.org.au), a foundation member of Bio21, took the initiative and led a strategic project that defined the "current state" and "preferred future state" for clinical informatics and bioinformatics research. In the "current state" clinical and medical research data was an under utilized resource, held in data silos, with ethical and privacy issues limiting access, collaboration and data sharing within and between institutions. Although there was a positive collective will—there were limited resources, a lack of standards and an inconsistent approach to data collection, storage and utilization within and across Australian hospitals.

Market analysis plus consultations and workshops with researchers, clinicians and key government stakeholders defined the "preferred future state" and a pilot system, the Molecular Medicine Informatics Model (MMIM) was proposed. The vision was a virtual platform, where information is accessible to authorized users, yet the data remains physically in the custody of its curator, the physical connections creating a "virtual" data warehouse.

There was considerable skepticism as to whether the MMIM could accomplish its desired future goals. How would the project be able to overcome the funding, technology, standards, privacy, ethics, data linkage, intellectual property, scalability, governance and sustainability challenges? This article details the successes of MMIM and its evolution into BioGrid Australia (BioGrid).

Challenge ♯1: Funding

The project has been funded over seven years by three major government grants, which funded the pilot and technology scale up. In February 2004 the Victorian Government awarded Australian dollars (AUD) $1.66 million in funding for a pilot project, the MMIM (AUD 1.00 equals US$0.98). The former State Premier (then Minister for Innovation) John Brumby stated "the project, using leading edge computer technology, will see the hospitals and institutes link genetic and clinical research information which will ultimately lead to better treatments for diseases such as cancer, diabetes and epilepsy," further he said "short-term benefits will include the capacity to select patients for clinical trials based on their genetic profile, increasing Melbourne's attraction for pharmaceutical drug trials." In addition he set the goal of "development of leading edge computer technology and data management processes to overcome the many challenges of integrating research data stored by our hospitals and research institutes." (23 February 2004 Media release From the Minister for Innovation and the Minister for Health Victoria Australia)

The MMIM pilot was based in Parkville, Melbourne, Australia with Melbourne Health leading a collaboration of the Peter MacCallum Cancer Centre, Austin Health, Western Health, the Alfred Hospital, the Ludwig Institute for Cancer Research, the Walter and Eliza Hall Institute of Medical Research and the

*Correspondence to: Robert B. Merriel, Melbourne Health, BioGrid, Royal Melb Hospital, 300 Grattan St, Parkville, Victoria, 3050, Australia.
E-mail: robert.merriel@mh.org.au

University of Melbourne. In 2004 a public tender was commissioned to design and implement a technology solution for the MMIM. Responses were received from multiple vendors and IBM was selected. The decisive factor was IBM's federated data storage mode, where data is stored in remote but connected repositories versus proposals for a larger centralized data-warehouse. This federated model was decisive in overcoming many of the top 10 challenges BioGrid has faced in the last seven years.

In February 2005 the MMIM was publicly demonstrated at a Bio21 Bioinformatics Symposium. In a university lecture theatrette the audience was presented with a live demonstration over the Internet of how a researcher could use the MMIM infrastructure to run a query across multiple databases, multiple disease groups, and multiple institutions. This was a critical turning point, the pilot was successful, and by mid 2005 it had delivered on all of its objectives.

The project was then boosted by an AUD $4.37 million grant from the Australian Government (http://www.dest.gov.au/NR/rdonlyres/B7D0178C-9908-4722-8A1E-A2A8C6C08534/7726/Outcomesofthe2005CallForProposals2.doc). The Phase 2 grant enabled MMIM to move beyond the pilot stage to expand the site coverage from one to five states and territories, expand from 5 to 15 health services; integrate additional disease groups including cystic fibrosis (respiratory medicine) and more of the neurosciences—multiple sclerosis, neuropsychiatry and stroke; include more tumor types; and, link new data types, including digital images.

In 2006 the collaborative was awarded AUD $11 million funding from the Victorian government under the Healthy Futures Program—for Phase 3 "The Australian Cancer Grid" (ACG; http://www.business.vic.gov.au/busvicwr/_assets/main/lib60149/lifesci_web.pdf, page 34). This project aimed to expand BioGrid across Victorian metropolitan and regional cancer centres; and to encompass further cancer tumor streams. The grant included AUD $1 million for collaborative cancer research.

In late 2007 MMIM was renamed as BioGrid Australia, with the motto of "*health through information*," reflecting the clinical and genetic research—the biological domain—"Bio"; the national perspective on research infrastructure and data linkages—"Australia"/"grid"; and, the translational nature of the research—to improve "health."

Challenge #2: Technology

Data is connected to BioGrid by loading data from the source databases onto a server known as the Local Research Repository (LRR). The LRR is housed within each site's Computer facility. BioGrid does not dictate which source applications a site employs to capture its data. Each site can use the technology, applications design and platforms of their choice, ensuring sites can maintain their existing source systems. BioGrid implements Extract Transform and Load (ETL) software, (IBM Infosphere Datastage, IBM DB2 Warehouse Centre, MS Sequel Server SSIS, Talend ETL) between the source systems and the LRR. Each night the ETL processes enable the data that is stored in the disparate local systems to be integrated and loaded onto a single research repository at each site. The data on each LRR includes a table of personal identifiers, with completely separate tables containing the clinical data.

Each database at each site has a designated data owner who is responsible for the collection and quality of data, and directly controls access to the data on their LRR. Data is physically located at the site with the data owner and, to decrease complications, data is stored as much as possible in its original format. If changes need to be made to allow data linkages, these are kept to a minimum. Different sites use different databases for their LRR. Examples include IBM DB2, Microsoft® SQL Server, through Oracle 11 g to MySQL.

The next step was to unify the data from these diverse LRRs to create the virtual repository. This was achieved by creating a view of all the databases using the Federated Data Integrator (FDI). The FDI, (IBM Infosphere Federation Server) is installed on a separate server and contains the mappings or views of each LRR: basic information, such as the database that the source data was stored in, and the names and descriptions of the data fields.

BioGrid connects data across a heterogeneous technology environment. The different ETL tools in use and the federated integration accommodates connectivity to a broad range of databases, so the BioGrid sites are able to maintain their existing source systems and implement the technology of their choice for the LRR server.

BioGrid needed to create a way for researchers to query and search databases across the Internet. Initially a program called Query Management Facility (QMF) was trialed to act as a Web-based query tool for researchers to interrogate the virtual repositories. By September 2005 BioGrid had replaced QMF with SAS™ technology via a terminal server, which has proved a more user-friendly method of running web-enabled data queries.

In January 2007 BioGrid upgraded its technology offering to include SAS™ Business Enterprise Version. This software incorporated additional features and functionality to enable the system to be more user-friendly and deliver a variety of statistical analysis tools for multiple users. BioGrid has also implemented SAS Web Report Studio—enabling non-technical users to find, interact with, create, and share reports using BioGrid data. Features include a range of statistical and graphical representation and output types include PDF, images and HTML.

Researchers query BioGrid via the Internet and use SAS Enterprise Guide SPSS & STATA which are located on a terminal server. Researchers have virtualized access to the data they have been authorized to access on the LRRs as if they were a single source. Approximately 4,000 research queries are processed each month querying data across organizations and across state borders.

In 2009 a demonstration project illustrated how the data infrastructure met the needs of international collaborative clinical research. BioGrid linked cancer test data internationally between sites at the University of Melbourne and Vanderbilt University—the University Medical Centre and the Vanderbilt-Ingram Cancer Centre.

Challenge #3: Data Standards

BioGrid has developed collaborative and technical processes for managing differences in data standards and data definitions. Initially working groups of researchers, clinicians and data managers, with specialist knowledge of their disease area, met to map out the data elements that would form the consensus data set, to be collected at each participating site. BioGrid does not seek to create new data sets nor dictate data definitions but encourages researchers to agree on accepted standards and uniform design and format.

During the pilot, researchers could not search the system to identify what other data-tables were connected to the platform, the case numbers and what data they contained. Further, users could not see database table and column names without prior approval and often these names were not intuitive. To address these issues the project implemented IBM Infosphere Business Glossary providing functionality which includes a text description annotation of most of the elements of the stored data, a consistent naming convention for the data elements and hierarchy which classified the data fields into categories to assist content browsing. Also supported is text searching with synonym support so that equivalent data elements could be found using clinical language and linking of linking of the above descriptions, names, hierarchy and synonyms to the actual table and column names in the database.

The business glossary covers all of the BioGrid databases. The management of glossary and metadata is a challenging administrative task. The initial annotation involved the transfer of knowledge from data owner to data administrators. Although the data owners can manage their glossary terms on-line, there has been little uptake by data owners and so metadata management has been challenging to implement and maintain. Web portal technology and new version functionality will facilitate continued and expanded up-take and use.

Internet access to view the glossary is unrestricted. This open access affords any researcher, Australian or international, the opportunity to browse the categories of data available by description and their structures. The glossary provides a means to discover what kind of data is available via BioGrid, without accessing the patient level data.

For BioGrid to enable high quality research outputs, it is vital that the data in the system is of a high standard and poor data quality can be caused by a variety of factors. If data entry systems lack quality checks, are poorly documented; data entry standards are not created or adhered to; or users don't understand the impact of missing or inaccurate data, then data quality will be compromised. Lack of clarification of data ownership and failure to incorporate business/validation rules into data entry systems can also cause data quality problems.

BioGrid has assisted data quality in multiple ways. BioGrid funded data officers have worked collaboratively with data owners to optimize the data fields and formatting of the data to be collected. A data quality and cleansing process has been implemented, including the possibility of cross referencing disparate databases that contain similar information, providing an opportunity for checking for discrepancies and improving the data accuracy in both data sets. Where possible BioGrid has used technology assets to check data at the point of entry, flagging where unlikely or impossible data has been entered.

Challenge #4: Privacy

Patients in Australia increasingly receive health care at multiple sites and from multiple organizations. For example it is not uncommon for a cancer patient to receive surgery at one hospital, a PET scan at another and radiotherapy at a third completely separate organization. The Australian health system creates multiple, site based, patient identification numbers and includes them in a mostly paper based medical record. BioGrid has implemented an electronic privacy protection process that allocates one identification number per patient and stores that information as a Unique Subject Index (USI), a linkage key enabling the same individual's record to be linked where they appear across multiple databases or sites without revealing to the researcher any indentifying information.

The USI application is a secure "black-box" operation that contains no health data. Using probabilistic matching the USI application assigns a number to each new patient in BioGrid. The USI allows different records to be linked for the same patient across multiple databases and organizations while retaining patient privacy. The USI is created using the Java Caps (Oracle) application, which is used by many Australian hospitals for integration and record matching. Its basic implementation is to determine patient matches using 6 fields: surname; given name; middle initial; date of birth; gender; and 5 digits from the patient's Medicare number (http://www.medicareaustralia.gov.au/). The USI server is a separate computer which houses the JCAPS program and tables. This contains the 6 fields for each patient who has been added to a hospital's LRR and the assigned USI. Each night, all the LRRs are scanned for any new patients or for those whose identifiers have changed. Each record is compared against the "master" so that a patient who already has a USI, is allocated the same number, and a new patient gets a new number.

The matching algorithm calculates the probability of a true match or a mismatch. Defaults are used in some fields, however a valid name and date of birth is needed—if these are missing the data owner receives an error report. The matching process handles challenges common to any probabilistic matching process.

BioGrid has extremely high standards for the protection of individual patient privacy. The project has received extensive legal advice on the requirements for compliance with national and state privacy laws governing information where that individual's identity is "apparent or is reasonably ascertainable." Essentially, privacy is guarded by separating personal information (such as names) from the clinical data by the USI allocation function. Clinical information is held on each LRR, whereas the six identifiers are stored on the central USI server. To safeguard the patient's privacy no clinical data is stored centrally, the USI server does not hold health information and the USI is stored in the LRR in encrypted form. In addition administration of the server is managed by a Health service, not by BioGrid, thereby separating the identifier and USI management from BioGrid.

BioGrid therefore links very detailed information about a person's health without disclosing that person's identity; researchers only have access to de-identified data.

In the majority of BioGrid data linkages the probabilistic matching, as outlined above, uses a direct comparison of two sets of patient identifiers and assigns a weight to each identifier and totals the score of the matched fields. This total is then checked against a threshold value to determine if the two sets of identifiers are the same person. Probabilistic matching requires that the two sets of identifiers be brought together in the one place for comparison. This is not possible if the legal and ethical requirements for one or both data custodians preclude the release of this information to the other party. In this case, BioGrid uses an alternative—the "hashing" algorithm.

BioGrid has worked collaboratively to implement a sophisticated hashing algorithm called Grhanite (www.grhanite.com). This hashing algorithm is run at the selected data custodian's site on the identifying data. For each set of identifying data, it creates a set of strings of 256 characters. The same identifying data will always produce the same hash value, but the hash is mathematically proven to be non-reversible. In other words, it is not possible to decipher the hash to retrieve the underlying identifying data. In addition, the hash value that is produced is "pseudo-random"; that is, values that are similar but not exactly the same will produce hashes that are not similar, making it impossible to determine the identifiers by a technique known as "dictionary attack."

The same hashing algorithm is used to create hashes at the target site where data is stored and within BioGrid for all patient identifiers. The hashes are then brought to BioGrid and assigned a USI. No identifying information ever leaves a site. Consequently, the strict legal and ethical requirements are satisfied and the data is available as part of the virtual repository.

Additional privacy protection is obtained by all researchers wishing to use BioGrid signing an agreement confirming they will adhere to the privacy and ethics requirements and to collaborative research principles of BioGrid prior to receiving access to any data.

In June 2010 the Australian Government passed Individual Healthcare Identifier legislation (http://www.health.gov.au/Internet/ministers/publishing.nsf/Content/mr-yr10-nr-nr135.htm) authorizing the issuing of individual 16-digit identifier numbers to all Australian patients. BioGrid is capable of incorporating the use of the individual healthcare identifiers to further improve the USI process.

Challenge ♯5: Ethics & Security

Every site that seeks connection to BioGrid must first obtain ethics approval from their local governing human research ethics committee (HREC). While their specific requirements and methodologies vary, the ethics committees of all 25 members have approved the BioGrid processes for managing data, subject de-identification, the linkage process, access management, and data collection. Subsequently, ethics approvals are not required for almost all further research projects conducted, provided they comply with the ethically approved standard processes, a major practical advantage for researchers. An exception is projects involving Government data owners, where specific ethics approval is required before data can be used.

BioGrid has worked with the data owners to record the "Participant Consent" or "opt-out" consent process for each data collection and to ensure the system respects the chosen option. The method that applies differs from data collection to data collection, with ethics approvals for some projects requiring an "opt in" whereas the majority are managed by "opt out" processes. In the context of data being collected from multiple sources for future de-identified research, the important aspects for patient consent are as follows. The patient is informed that their de-identified ie., codified data could be used for research purposes, that the research may extend across a number of institutions and that any researcher must have appropriate ethics approval to conduct the research. The effects of providing or withholding consent; and that there is the opportunity to withhold consent without compromising care, are also made clear to the patient.

Once consent requirements are satisfied, there is no obstacle to obtaining consent in the same way from a patient where their information is held at two institutions. An information sheet provided to patients should make it clear that the patient would be consenting to collection of all information relevant to their treatment; information relevant to their treatment is likely to be held at multiple institutions; and, researchers are only interested in information held at institutions that is relevant to the course of treatment.

There are four models of consent for prospective data collection: (1) Research intervention projects—explicit consent for initial and ongoing use of codified data; (2) Genetic data—explicit consent to ongoing use of codified data; (3) Clinical "audit" data—opt-out consent to ongoing use of codified data; and (4) Administrative data sets—no consent as approved by HREC committee—but no re-identification is allowed. An Ethics committee may waive the requirement for consent in certain circumstances and so all models must be considered.

In the event that a researcher makes a discovery that could help patients already within the BioGrid databases, the BioGrid process enables, with ethics committee approval, patient re-identification. This is a requirement of Australian code of conduct for human research.

BioGrid has passed two security audits of its processes and technology, engaging an independent expert to conduct each audit and verify the adequacy of security controls and processes, taking into consideration any data or control changes that have occurred since the previous review.

Challenge ♯6: Data Linkages

The BioGrid model supports linkage of a variety of data types beyond clinical data. These include genomic, imaging, and histopathology. Although linking to the publicly accessible databases GenBank (a repository of gene sequences and their functions) and SwissProt (a database of protein information, including structure and function) was provided in the pilot, there was little demand for this access and it is no longer provided.

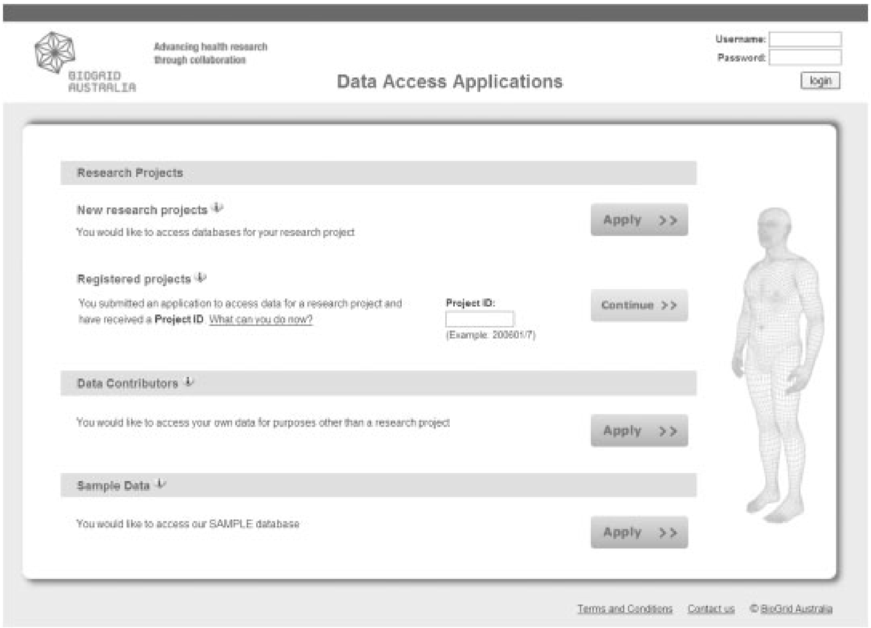
The Victorian Cancer Biobank (VCB; http://www.viccancerbiobank.org.au) provides a coordinated program of bio-specimen collection and sample distribution to facilitate cancer research. BioGrid collaborates with the VCB, adding significant value by linking data about the collection and storage of the human tissue samples; with the clinical information about the patient, their treatment(s) and outcomes. The provision of high-quality, clinically annotated bio-specimens enables researchers to differentiate between cancer origins and types and facilitates the discovery of predictive and prognostic biomarkers, as well as the development of targeted therapies, all of which will lead to improved patient outcomes.

Digital images, such as MRI, CT and PET, are increasingly being acquired as part of clinical care and research. Image data can be linked through BioGrid, however the average file sizes of these images are large, especially as an individual may generate a series of images (300 MB or greater) creating a challenge when linking data across the Internet. In 2006 BioGrid chose Oracle 10 g RDBMS as the technical solution for DICOM image storage, and the metadata extraction and retrieval, because of its interMedia feature which specializes in working with images as objects in the database. BioGrid joined the Oracle 11 g Beta testing program as specific enhancements to the DICOM image handling were part of that release. The successful test of the Oracle 11 g Windows Beta release demonstrated that a view of the image metadata can be built in Oracle 11 g and the IBM Infosphere Information Integrator can successfully map to that view.

The BioGrid Access Management System (AMS) is an online application that manages user access, access requests and information on BioGrid databases, disease groups, current users, institutions and research projects (Fig. 1 is an example of the user interface). The AMS has three interfaces:

1. *the Access Request Interface* (for anyone); Researchers requesting access to BioGrid data fill in the online form and submit their application. Each person included in this project must agree online to abide by the BioGrid Terms and Conditions. Data owners are automatically notified via email of the new request to access their database(s) and they then provide a decision by email, allowing or denying the use of their data;
2. *the Administration Interface* is for authorized BioGrid staff members and a login is required. This functionality covers the ability to view, add and edit detailed information about BioGrid members, institutions, research projects, disease groups, databases and submitted access requests and their status;
3. *the Data Owner Interface* enables an overview of who has requested access to each data owner's database(s), which projects and people have had access to this data and a list of any new access requests.

Each request receives two scientific reviews and each project must meet the required standards (importance of research questions, quality of the science, availability of required data and adequate sample size) before researchers can be permitted to access data, and finally the application is reviewed and approved



Data Access Applications

Research Projects

New research projects

You would like to access databases for your research project

Registered projects

You submitted an application to access data for a research project and have received a **Project ID**. What can you do now?

Project ID:

(Example: 200601/7)

Data Contributors

You would like to access your own data for purposes other than a research project

Sample Data

You would like to access our SAMPLE database

Figure 1. BioGrid User Interface.

by the BioGrid Management Committee. At any stage projects can be sent for ethical review/advice before granting final approval.

BioGrid tracks all queries to the data by using DB2 Query Patroller, a query management tool employed to monitor the flow of queries against the Infosphere Federation. This helps staff monitor the use of BioGrid by approved users and track queries for regulatory and governance requirements. The software collates information about completed queries and is used to audit queries, heavy users, and frequently used tables and indexes. It also enables database administrators to monitor the types of requests that researchers submit and allows delaying or even placing a query on hold, if it will unreasonably slow the system.

Challenge #7: Intellectual Property

Concerns about losing control of intellectual property (IP) can discourage researchers from sharing data. Recognizing this BioGrid worked together with legal advisers to develop a standard process for the management of IP so that data owners present and future IP rights are protected. Detailed guidelines were developed, and before gaining access to BioGrid, researchers and their institutions must agree to abide by this protocol.

The IP of BioGrid projects is defined as either Background IP, which is IP relating to the databases or other IP that each institution or research group will contribute; or Project IP, which is the IP arising from research conducted using BioGrid. Background IP is retained by the party that supplied it. Each party grants a non-exclusive license for other project participants to use its nominated Background IP appropriately within their specific research project. BioGrid users must comply with any additional IP conditions set by the institution supplying the data.

Before beginning a project, each research group must appoint a Principal Investigator, identify the Background IP that they are bringing to the project, the data they wish to access and identify which party will be the Commercialization Lead should patentable IP result from the collaboration. The Commercialization Lead is responsible for any efforts made to commercialize the Project IP. Each project participant agrees to provide all possible assistance in protecting any Project IP. There is a default position that allows for rapid project start-up.

Challenge #8: Scalability

An enhanced technical architecture that enables robustness on a national scale was successfully implemented in early 2009. This technology upgrade provided scalability as it extended the IT architecture and upgraded the infrastructure to take BioGrid into the next phase: a sustainable technological operation. BioGrid's technology is now scalable, robust, responsive and reliable. The main areas of accomplishment in this upgrade included the addition of a distinct development environment for nearly all components of the architecture. This is where BioGrid develops and tests for changes to existing IT systems or for any new IT systems. This facilitates innovation and enhancement while protecting and maintaining a reliable service in the production environment. Also included were a commitment to the use of server virtualization, where servers are partitioned into smaller "guest" autonomous servers. This facilitates the greater use of the physical servers, reduces space requirements, reduces power consumption, simplifies backup plans and contributes to budgetary savings. There are 31 servers in the central BioGrid infrastructure, a mix of physical (14 of which are virtual server hosts) overlaid with 17 virtual servers. Further, BioGrid added new technologies to the infrastructure including new data analysis and metadata discovery tools, IBM InfoSphere Information Analyser and InfoSphere Business Glossary Anywhere. Web security, IBM Tivoli Access Manager and Directory Server were also added. The use of web services has been demonstrated through a web service delivery of our record linkage service. In addition the facility to host external institutions' data sets on a BioGrid independent research repository was added as an alternative to them installing a LRR at their site, which is a feature attractive to registry data owners. Finally, BioGrid installed IBM Portal Express which resulted in greater ownership and total renovation of our website and delivered the potential for continuous growth.

Challenge #9: Governance

During the first two stages the project governance was managed as an unincorporated joint venture of collaborative institutions, each of whom signed a Collaboration agreement providing equal rights of representation on the BioGrid Management Committee, equal benefits of membership and equal responsibilities.

The original collaboration agreement expired in October 2008 and the project subsequently transitioned to a new Governance structure. BioGrid Australia Limited, a company limited by member guarantee, was incorporated to streamline management and membership processes, to address potential legal liabilities, and to provide greater flexibility in services such as data provision and access. The company was licensed by the members to operate the data linkage and integration platform on their behalf.

BioGrid is a not-for-profit company with any surplus income to be reinvested in future operations. It can have up to 7 directors. The members have rights to collectively appoint two Directors to the company Board and individually to be represented on the BioGrid Management Committee. The management committee's role transitioned with the change to a company structure and the committee is no longer directly responsible for the daily operations of the project.

BioGrid has two Science Advisory Committees (SAC's) comprising representatives from member organizations and affiliates, one focused on cancer and the other on non cancer disease areas. The SAC membership has been selected based on representatives' expertise in particular tumor stream areas/diseases. The role of each SAC is to lead and oversee the science and research activities of BioGrid, with representation and input from each tumor stream/disease.

A staffing structure has been developed to manage BioGrid operations on behalf of the Board and requires experienced and highly skilled personnel. Current staff have been involved in the development of the BioGrid platform, medical research and leadership. The team encompasses expertise in the areas of research methodology, data management, database design and development, research analysis, information technology, systems and architecture, and business management.

BioGrid manages its multi-party collaborations by formal agreements that either provide membership to the collaborator or establishing a research collaboration for a specific project.

BioGrid membership has expanded greatly over its years of operation. There are currently 25 member organizations covering 34 sites with a further 9 organizations currently in various stages of engagement. A six stage classification system is employed to monitor progress with installations and membership. The six stages are:

1. proposed new member seeking ethics approval;
2. with formal ethics approval the proposed new member is seeking executive/corporate approval;
3. with formal ethics and executive approval the member implements the information technology hardware and software resources to enable data linkage through the LRR;
4. with the information technology in place the member is working on annotating and making available the database(s) for the ETL process from source systems into the LRR;
5. with the information technology and ETL processes in place the member reviews and refines the quality of their data; and
6. the member has all the ethics, corporate, IT approvals and data management systems in place and data linkage has begun.

The membership classification, as at 1st September 2010, is shown in Table 1.

Table 1. BioGrid Membership Classification

Stage	Description	Sites at this stage of membership
1	Seeking Ethics approval	5
2	Seeking Corporate Governance	4
3	Implementing IT Connectivity	1
4	Documenting Metadata	0
5	Reviewing Data Quality	1
6	Active data linkage	21

In addition to the information in Table 1, BioGrid has research agreements and memorandums of understanding with Monash University and three additional medical research organizations.

The Commonwealth Scientific and Research Organization (CSIRO), Colo-Rectal Surgical Society of Australia and New Zealand (CSSANZ), VCB, Cancer Trials Australia (CTA), Australian Institute of Health and Welfare (AIHW) as well as caBIG in the USA and the National Cancer Research Initiative UK Cancer Informatics Initiative (NCRI) are examples of local and international collaborators who are not BioGrid members.

Challenge #10: Sustainability

BioGrid is yet to achieve financial sustainability, despite its many research and technical achievements to date, BioGrid has found its biggest challenge to be establishing an appropriate mix of revenue streams that will support its long term financial security. While the technology transformation is complete and the research results internationally recognized—the transition from the 2004 "current state" 100% government grant funded research project, to the 2011 "preferred future state" a self sustaining business model with a mix of government and private sector revenues, with fee for service provision to clients and a pricing model that covers all infrastructure costs, is BioGrid's priority challenge in the next 12 months.

We believe BioGrid's financial sustainability will flow from expanding the services which utilize the infrastructure, which will attract increased greater research interest and funding opportunities with expansion of data integration across sites, organizations, borders and disease types. BioGrid's achievements, both research and technical, demonstrate its value and the opportunities for financial sustainability.

Biogrid's Achievements

BioGrid has attracted clinical trials to member hospitals—through the ability to quickly analyze patient populations, responding to clinical trial patient profile requests from pharmaceutical companies, electronically capturing patient data and linking clinical information to stored tissue samples.

BioGrid has changed clinical practice—researchers have published a study that showed that interval Faecal Immune Testing between scheduled colonoscopies for individuals who are on colonoscopic surveillance programs, can detect missed rapidly developing neoplastic lesions [Lane et al., 2010]. This was achieved by linking databases, via BioGrid, on the results of Faecal Immune Testing with data on patients having surgery for colorectal cancer.

BioGrid has enabled doctors to better understand why drugs work for some patients and not others—Personalized Medicine. Petrovski and colleagues [Petrovski et al., 2009] reported the first "proof-of-concept" study for applying machine learning methods to identify patterns of multiple genetic markers that predicted with a high degree of accuracy the chance that a patient would achieve successful seizure control after starting treatment with a medication for epilepsy. This year the same group published a further paper [Petrovski et al., 2010] which extended this work to incorporate non-genomic data showing that adding information from MRI and a neuropsychiatric questionnaire added additional predictive value to the genomic classifier. This work was enabled by BioGrid data linkages and analysis tools and it provides the "blue print" for an approach to developing such classifiers which can be applied to many other common diseases, and potentially improve the quality and safety of drug prescribing.

BioGrid researchers have identified challenges to the management of genetic risk factors for disease [Wong et al., 2008] by linking data from the genetic clinic and the colorectal cancer clinical database to demonstrate that many patients that were likely to have an inherited risk of bowel cancer were not being referred for genetic testing and advice, and that many that were referred failed to attend the appointments.

The first published international multicentre study of the pharmacogenomics of epilepsy treatment [Szoeke et al., 2009] reported that a genetic polymorphism that had previously been implicated, does not have a strong role in determining the efficacy of initial drug therapy in controlling seizures in newly diagnosed epilepsy. The study also highlighted the issues that arise in combining pharmacogenetic datasets from different ethnic regions and health systems, an approach which is essential to advance this field.

BioGrid has enabled research projects with sample sizes greater than 10,000, and has linked data across Organizations, State and International borders. Encouraged and informed by proof of concept work for identifying multigenic predictive models for the outcome of newly treated epilepsy, Petrovski and colleagues have initiated a multicentre international collaborative project combining cohorts of patients from Melbourne, Perth, England and Scotland prospectively followed for the outcome of newly treated epilepsy (approximately 2,000 patients in total). These patients have been genotyped for 550,000 SNPs genome-wide on the Illumina 660^{TM} platform. The project team will collate these large genotype-phenotype databases, to identify and validate more accurate multi-SNP pharmacogenomic models that provide clinically useful predictions of the likelihood of an individual initiating treatment for epilepsy achieving seizure control with a range of commonly prescribed anti-epileptic drugs.

BioGrid has linked digital images with clinical data. In one of the largest published series [Jones et al., 2010] of patients with video-EEG confirmed psychogenic non-epileptic seizures with detailed neuropsychiatric assessments researchers demonstrated that many patients experienced long delays in diagnosis and high rates ($>80\%$) of prolonged, inappropriate, treatment with anti-epileptic drugs. Patients assessed at follow-up exhibited poor long-term outcomes with ongoing Psychogenic Non-Epileptic Seizures (PNES), high rates of psychopathology, low rates of specialist follow-up, poor quality of life and poor overall levels of functioning. These results demonstrate the need for earlier diagnosis of PNES and co-morbidities, and highlight the need for diagnostic and therapeutic approaches which combine neurological and psychiatric perspectives.

BioGrid has converted 12 years worth of MRI Brain images from over 1000 DAT format tapes onto a 6 terabyte disk array. The MRI images were converted from a proprietary format to the DICOM format. These images were loaded onto a new online disk system, making them available almost immediately for use by clinicians. BioGrid then progressed to make CT and PET brain images similarly available.

BioGrid has helped its researchers win international recognition for their output. Awards given to BioGrid and researchers utilizing BioGrid include:

- 2007, Dr Cassandra Szoeke was awarded the Bio21 Industry Fellowship;
- 2008, BioGrid was awarded a Laureate Medal by Computer World;
- 2009, two BioGrid Australia Cancer Grid Research Fellows received Merit Awards at the Gastrointestinal Cancers Symposium USA;
- 2009 Slavé Petrovski was awarded the American Australian "Sir Keith Murdoch Fellowship" for medical research for 2010;
- 2009 Dr. Jeanne Tie, was awarded the 2009 Bradley Stuart Beller Merit Award.

BioGrid has attracted more than AUD $1.2 million in commercial projects and has generated significant interest in collaborative research among pharmaceutical companies. Twenty-five projects have been completed, 9 projects are underway and many more are in discussion. The largest collaboration is a 4 year project with Roche Products, a prospective study into how clinicians' make treatment decisions in metastatic colorectal cancer. This study will be the first of its kind and aims to increase our understanding of the impact of current treatments on patient outcomes in advanced colorectal cancer (29 October 2009 Media release from BioGrid Australia and Roche Products).

BioGrid has fostered the development of commercially valuable IP. The multigenic predictive model for the outcomes of the treatment of epilepsy [Petrovski et al., 2009] was patented in 2009. Using clinical and genetic data linked through BioGrid, researchers have developed a test for adverse drug outcome in epilepsy using multiple genomic markers. It is anticipated that such a model would improve patient care by reducing the duration of trialing different antiepileptic drugs and reducing the possibility of having an adverse drug reaction or not maintaining seizure control.

BioGrid has developed multiple software applications that support clinical care and research data collection. The Australian Comprehensive Cancer and Research Database (ACCORD) is a reliable and easily accessible web based database. It has been developed by BioGrid for hospitals that require a simple data collection and entry system and allows clinicians to analyze the clinical and demographic characteristics of cancer patients at their hospital. Modules have been developed specifically for cancer of the Bowel; Brain; Head and Neck; Prostate; Sarcoma; and Kidney; and for Chronic Lymphocytic Leukaemia; with a module for Breast cancer being finalised. New modules, where funding is available and planning is underway include Lung; Liver; and, Thyroid cancer.

An "electronic" chemotherapy prescribing ACCORD module has been developed for colorectal cancer, enabling clinicians to record and prescribe neo-adjuvant (prior to surgery), adjuvant and palliative chemotherapy regimens. It documents all aspects of the treatment regimen; the doses administered, side-effects, dose adjustments, and early termination of treatment. The program also records details of treatment response, time to cancer relapse and subsequent therapy. The electronic format minimizes the amount of missing data, improves the quality of recorded information and is user-friendly. Importantly it has increased patient safety as chemotherapy dosages are calculated by the computer and printed prescriptions minimize the chance of human error in prescribing and consequently administering doses of chemotherapy. The program has a number of failsafe mechanisms to ensure only appropriate doses of chemotherapy are ordered, and all patient drug allergies are recorded.

Web based Diabetes Survey/Clinical System—this enables clinicians to rapidly document care of type 1, type 2 and gestational diabetes-related complications and treatments. It uses Microsoft SQL Server as its database; enables unlimited concurrent users; links to the hospital Patient Master Index and pathology data, and collects information on all types of diabetic complications, additional risk factors, co-morbidities, and medications. It functions as an efficient tool for generating concise letters for general practitioners and comprehensive management summaries for the discharge of patients from clinic.

BioGrid has introduced two Neuropsychiatry systems. The first system enables results from the Neuropsychiatry Unit Cognitive assessment tool [Walterfang et al., 2006] to be recorded and displayed via a web interface. The system records the test scores and graphs the changing results for a patient over time, allowing monitoring of progression of cognitive impairment and/or response to treatment. Studies are underway in epilepsy, multiple sclerosis, adrenomyeloneuropathy and Fabry's disease. The second system records and graphically displays the results of a range of neuropsychological tools and test items used in clinical practice.

BioGrid has developed a pilot web based Clinical Viewer as a parallel system to the research linkage, enabling data linkage of current clinical data to inform the treatment decisions for patients who are being treated by several health services. Identified clinical data, rather than de-identified research data is linked via a web based portal sharing the data with all clinicians treating the patient even though they are spread across a number of sites and medical disciplines. This was written using mediaflux technology and a customized referral/consent authorization module.

BioGrid has been a member of teams which have won in excess of AUD $6 million in national competitive research grants. The Victorian Cancer Agency awarded BioGrid a Platform Technology Capacity Building grant to develop training packages for people working in Cancer in Victoria. This project is producing professional-quality web based training materials to teach current and potential cancer researchers how to use record linked data in their work and bring them closer to the data required for translational research.

BioGrid is an important component of the NHMRC Centre for Research Excellence (CRE) in Translational Neuroscience (http://www.nhmrc.gov.au/grants/) which was awarded in 2010. This CRE, which will be based within the Melbourne Brain Centre at Royal Melbourne Hospital and the associated University of Melbourne Academic Centre, is a comprehensive program specifically designed to take up and generate new knowledge in clinical neuroscience and to translate this into improved patient care. The Centre will have a major focus on training, with a particular focus on implementing health practice change. The model is modular and readily adaptable to non-neuroscience disease areas in Australia and beyond. A key enabler of the research to be undertaken in the CRE is the linking of clinical and para-clinical information using BioGrid.

BioGrid has provided web portal technology for consumer involved research and communication. BioGrid developed a web-based Rare Tumor Database branded as CART-WHEEL.org (Centre for Analysis of Rare Tumors) (http://www.cart-wheel.org) which provides an ethically approved portal for consumer-driven data entry, enabling research into rare tumors and molecular sub-types of common tumors, using the infrastructure of BioGrid Australia. This tool assists with consumer participation in research, particularly as patients with rare tumors make up 20% of all cancers and 30% of all cancer deaths, yet receive less than 5% of research funding (AIHW, 2006; Cancer Research Reports Vic, 2006).

BioGrid has facilitated more than 87 journal publications. In the majority of cases, BioGrid has integrated important data from a number of sources and enabled them to be accessed from the virtual repository, saving hours of work trying to access data from different sources. A broad range of issues in treatment and diagnosis have been explored in oncology, diabetes, the neurosciences and respiratory diseases. Publication topics range from impact of cultural background on timelines of diagnosis, to the uptake of chemotherapy in treating colorectal cancer.

BioGrid research and development has been presented in 22 Posters and 64 oral presentations at national and international conferences. Presentations have been made in many locations including Melbourne, Sydney, Brisbane, Perth, Adelaide, Bangalore, London, Washington, Las Vegas, Dallas, Brussels and Geneva.

BioGrid has developed a communication network with the medical research and bio-informatics community. BioGrid has produced 8 e-newsletters and 16 hard copy newsletters. Its website is a source of information and opportunity.

The Future

BioGrid has identified a number of key data connections and is in the process of implementing linkage with state government health databases (e.g. hospital admissions); national government registries (e.g., including the national death index which records date of death and cause of death); is in discussion regarding connection to national government health databases (eg, including prescribing and procedure data); disease specific registries; and, various ‘omics data to support systems biology research.

BioGrid is a collaborator on the Australian Node of the Human Variome Project. This project is in its initial stage and will collect information about every genetic variant reported by an Australian diagnostic laboratory and store it in a secure online repository. This information will be able to be linked to clinical data from Australian clinics using BioGrid. The data collection, including information on genes, variants and phenotype will be shared with Locus specific databases internationally. BioGrid infrastructure is enabling the privacy protected connection between pathology laboratories and clinical data so the Australian Node can share pathology interpretations and hence improve consistency in diagnosis, initially in Australia, but ultimately also enabling such information to be accessible to authorised users internationally [Cotton et al., 2008].

The future vision is for BioGrid as an integrated research portal which will support treatment and research centres in all disease types, integrating data across bio-specimen, treatment, clinical outcome and clinical trial databases. BioGrid will integrate data for analysis, mining and visualization; helping to identify biomarkers to increase knowledge on the causes and outcomes of all diseases. Linkage to systems biology initiatives will increase research opportunities. Supporting greater participation in clinical trials should accelerate the development of targeted therapies and commercialization of research and attract pharmaceutical research investment to participating centres.

References

Cotton RGH, Axton M, Bankier A, Brais B, Cavedon L, du Sart D, George P, Goldgar D, Harrison T, Hibbert M, Hopper J, Macrae F, O'Keefe CM, Ravine D, Savarirayan R, Sheffield L, Smith T, Stokes N, Sundararajan V, Thorburn D, Winship I. The Human Variome Project: Suggested Actions from the Melbourne Information Seminar Nature Proceedings: doi:10.1038/npre.2008.1784.2: Posted 11 April 2008.

Jones SG, O'Brien TJ, Adams SJ, Mocellin R, Kilpatrick CJ, Yerra R, Lloyd JH, Velakoulis D. 2010. Clinical characteristics and outcome in patients with psychogenic non-epileptic seizures. Psychosomatic Medicine 72:487–497.

Lane JM, Chow E, Young GP, Good N, Smith A, Bull J, Sandford J, Morcom JM, Bampton PA, Cole SR. 2010. Interval fecal immunochemical testing in a colonoscopic surveillance program for increased risk for colorectal cancer. Gastroenterology (in Press).

Petrovski S, Szoeke CEI, Jones NC, Salzberg MR, Sheffield LJ, Huggins RM, O'Brien TJ. 2010. Pre-Treatment Patient-Perceived Neurocognitive Symptomatology Predicts Seizure Recurrence in Newly Treated Patients. Neurology 75:1015–1021.

Petrovski S, Szoeke CEI, Sheffield LJ, D'Souza W, Huggins R, O'Brien TJ. 2009. A multi-SNP pharmacogenomic classifier is superior to single SNP models for predicting drug outcome in complex diseases. Pharmacogenetics and Genomics 19:147–152.

Szoeke C, Sills GJ, Kwan P, Petrovski S, Newton M, Hitiris N, Baum L, Berkovic SF, Brodie MJ, Sheffield L, O'Brien TJ. 2009. Multidrug resistant (ABCB1) Genotype & Seizure Recurrence in Newly Treated Epilepsy: Data from International Pharmacogenetic Cohorts. Epilepsia 50:1689–1696.

Walterfang M, Siu R, Velakoulis D. 2006. Establish the validity and reliability of a cognitive—or brain function tool. MMINM has computerised and using NUCOG and incorporating the data. Aust N Z J Psychiatry 40:995–1002.

Wong C, Gibbs P, Johns J, Jones I, Faragher I, Lynch E, Macrae F, Lipton L. 2008. Value of database linkage: are patients at risk of familial colorectal cancer being referred for genetic counseling and testing? IMJ 38:328–333.