The Molecular Medicine Informatics Model (MMIM)

Marienne HIBBERT [a,1], Peter GIBBS [b], Terence O'BRIEN [c], Peter COLMAN [d], Robert MERRIEL [e,] Naomi RAFAEL [f.] Michael GEORGEFF[g.]

[a] *Project Director, Melbourne Health, VPAC and University of Melbourne,*[b] *Chief Scientist , Melbourne Health and Ludwig Institute ,*[c] *Neurosciences leader, Melbourne Health and University of Melbourne ,*[d] *Diabetes leader, Melbourne Health and Walter and Eliza Institute,* [e]*Steering Committee Chairman, Melbourne Health ,*[f] *Senior Database Administrator, Melbourne Health,* [g] *Solution Manager, Monash University. MMIM, Melbourne, Australia*

Abstract. In 2005, a major collaboration in Melbourne Australia successfully completed implementing a major medical informatics infrastructure – this is now being used for discovery research and has won significant expansion funding for 2006 - 2009. The convergence of life sciences, healthcare, and information technology is now driving research into the fundamentals of disease causation. Key to enabling this is collating data in sufficient numbers of patients to ensure studies are adequately powered. The Molecular Medicine Informatics Model (MMIM) is a 'virtual' research repository of clinical, laboratory and genetic data sets. Integrated data, physically located within independent hospital and research organisations can be searched and queried seamlessly via a federated data integrator. Researchers must gain authorisation to access data, and inform/obtain permission from the data owners, before the data can be accessed. The legal and ethical issues surrounding the use of this health data have been addressed so data complies with privacy requirements. The MMIM platform has also solved the issue of record linking individual cases and integrating data sources across multiple institutions and multiple clinical specialties. Significant research outcomes already enabled by the MMIM research platform include epilepsy seizure analyses for responders / non responders to therapy; sensitivity of faecal occult blood testing for asymptomatic colorectal cancer and advanced adenomas over a 25-year experience in colorectal cancer screening; subsite-specific colorectal cancer in diabetic and non diabetic patients; and the influence of language spoken on colorectal cancer diagnosis, management and outcomes. Ultimately the infrastructure of MMIM enables discovery research to be accessible via the Web with security, intellectual property and privacy addressed.

Keywords. Molecular Medicine Informatics Model, MMIM, Research, Grid, Record Linkage

Introduction

The convergence of life sciences, healthcare, and information technology is revolutionizing the discovery of new treatments and the optimal use of available

[1] Corresponding Author: Dr Marienne Hibbert, PhD, MMIM Project Director, 6 North, Main Block, Melbourne Health, Grattan St, Parkville, Victoria, 3050, Australia; E-mail: Marienne.Hibbert@mh.org.au

therapies. Researchers now have the capability to analyse human biology at the finest level through genomics and link to clinical outcomes data giving them the potential to understand the fundamental causation of human disease and predict outcomes. This will drive the development of new drugs, new diagnostics, and lead us to the era of personalised medicine. Key to making the required associations between genotype and phenotype is access to detailed clinical data in sufficient numbers of patients to ensure studies are adequately powered. Few institutions alone have sufficient numbers to perform meaningful analyses, particularly where stratification is performed to look at specific disease attributes. Further, clinicians need to look beyond their own specialty into datasets of other disease groupings, analyzing the impact of co-morbidity. To achieve the research objectives promised by this new era in science, sharing of clinical data between research groups and institutions becomes critical.

The impetus for the development of MMIM came from a recognition of the need to maximise collaborative research across Australia and internationally. A cohesive approach between disciplines was needed so that research data collection became a one time only exercise with the data subsequently available to assist in answering multiple research questions across various clinical disciplines and jurisdictions. In addition new and emerging data sets such as genomic data could be linked to more traditional clinical and outcome data. Thus researchers could examine genetic, genomic and proteomic profiles, all factors that may influence treatment outcome, with respect to toxicity and potential benefit.

The MMIM Project enables research from multiple perspectives, including:

- Genetic predisposition;
- Environmental exposures;
- Health Screening activity;
- Genomics, proteomics & epigenetics;
- Co morbidities;
- Treatment strategies;
- Outcomes.

The objective of the project is to maximise collaborative research efforts, both in Australia and internationally through the development of a federated data integration infrastructure that is enabling:

- Linking and testing of multiple hypotheses without collecting / recollecting their own data;
- Identifying patient numbers for clinical trials based on clinical information or genetic profile;
- Researching suitable pre-symptomatic testing and early intervention based on genotype data;
- Analysing summary/statistical information across institutions and from diverse databases.

1. Materials and Methods

1.1. MMIM Background

Phase 1 of the MMIM project was successfully completed in 2005. It was a pilot project funded by the Science, Technology and Innovation Infrastructure Grant program (STI) of the Victorian Government Department of Innovation, Industry and Regional Development (DIIRD) through Bio21 Australia Limited. This phase delivered the successful integration of data across five hospital sites (The Alfred, Austin Health, The Royal Melbourne Hospital, Peter MacCallum Cancer Centre, and Western Hospital) and two medical research institutes (Ludwig Institute for Cancer Research, and Walter and Eliza Hall Institute). This stage of the project involved three disease types, namely, colorectal cancer, epilepsy and diabetes.

Phase 2 of the MMIM Project is currently funded until 2007 by the Australian Government Department of Education, Science and Training (DEST) through the University of Melbourne. Phase 2 is integrating data across additional Victorian and interstate hospitals including: Box Hill Hospital; Cabrini Health; Flinders Medical Centre; Monash Medical Centre; The Queen Elizabeth Hospital; Royal Adelaide Hospital; Royal Children's Hospital; Royal Hobart Hospital/Menzies Research Institute; Royal Women's Hospital and St Vincent's Health. Additional disease types to be integrated include multiple sclerosis, stroke, Parkinson's disease, cystic fibrosis, asthma, and brain cancer.

Phase 3 of the project is funded by a grant from DIIRD over a three year period until June 2009 to provide support for the creation of an Australian Cancer Grid and:

- The Infrastructure expansion – the data grid;
- The research activity and outcomes;

1.2. Overview of the MMIM Federated Mode - Technologies

The MMIM project is a federated model where each participating site retains ownership and control over their own data sources and data collection systems. The architecture can be seen diagrammatically in Figure 1.

The data sources used for integration were established as clinical research databases written and maintained by specialist clinicians in their own healthcare facilities. These typically have highly detailed clinical information including surveillance and treatment subsets.

Data was uploaded from these database systems nightly or manually loaded (for static datasets) into a 'cache' database, a local DB2 UDB database termed a Local Research Repository (LRR) located at each site. The distributed LRR databases were federated using IBM Websphere Information Integrator running on a single federating server termed the Federated Data Integrator (FDI). Information Integrator makes remote databases appear as local DB2 table views, allowing single SQL queries to be executed against all federated data.

Public domain databases were also federated into the system including a local XML flat file and resources from the National Library of Medicine (Genbank, Medline and Uniprot) via an Internet web service. Both data sources appeared relational to the end user even though they were not.

A unique number was assigned to each patient termed a Unique Subject Index (USI) by transferring certain identifying information to Sun (SeeBeyond) e-Ways and replicated back to the LRRs via the FDI.

The security system included a number of notable features. Each LRR was connected to the FDI via Virtual Private Network (VPN) connections, which ensure data privacy and encryption. Views block all identifying information, allowing end users to see only the clinical data in conjunction with the USI. User access to these views on the FDI is controlled by assigning DB2 database roles which define privileges to the table/view level. DB2 Query Patroller is used on the FDI to track all queries for audit purposes. Access is controlled by assigning permission at the table level.

SAS Enterprise Guide was used as the interface for researchers to perform queries, statistical analysis and construct reports.

1.3. The MMIM Architecture

Diagrammatically the phase 2 architecture is shown in the diagram below.

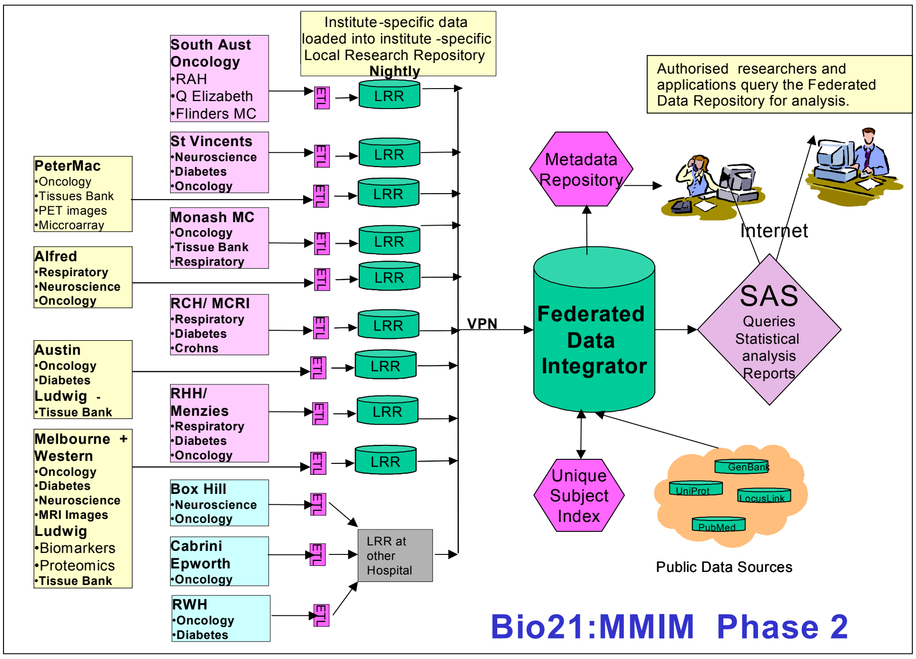

Figure 1. Schematic Architecture of the Bio21:MMIM data grid showing the secure data flow and technologies. Authorized researchers can access the integrated and de-identified data via the Internet.

1.4. MMIM Data Flows-Phase 1 and 2

The key features of the system include:

Connectivity:

Each participating research institution has a local data storage facility (the LRR). All data stores are connected with a secure technology involving double encryption Virtual Private Network and DB2-DB2 encryption. This is the technology commonly used in industries to link various sites of operations for major corporations, such as the banks).

Data loading:

On a nightly or ad-hoc basis, the clinical research data is loaded into the LRR at each individual site. This loading process uses an extraction, transform and load software feature that is installed on each LRR.

Unique Subject Index (USI) data flow:

The USI is a unique number given by the system to each patient; so data accessible to researchers for an individual is linked across databases, but de-identified. In order to de-identify patients, selected identifying data is sent on a nightly basis, from each LRR to a Unique Subject Index (USI) program, where the unique number is generated and stored in an Oracle database. This number is then pushed back to the LRR and stored in encrypted form.

End-user query data flow:

Researchers can only access de-identified data, except where they are performing research on their own databases. The identifying data remains stored on the USI data store with extremely restricted and controlled access. No Health Data is stored externally. Only authorized researchers can log in to MMIM via the Statistical Analysis System (SAS) and perform queries on the de-identified clinical research data. This data flows from the LRRs through the Federated Data Integrator (FDI), the engine that brings together the data from the various institutions (LRRs) to enable analysis. This de-identified data is then put into the researcher's secure folder for statistical analysis (the SAS server). All data queries performed are tracked and logged for security audit purposes.

Security levels:

There are two main levels of security for access to MMIM. Firstly, authorization and usernames and password must be obtained so only authorised researchers who have access to the FDI can query the de-identified clinical data. Secondly, the MMIM System Administrators who are responsible for building and maintaining the system have access to all parts of the system (currently there are only two MMIM project authorised personnel with this level of access).

1.5. MMIM Data Flows Phase 3

As MMIM expands and technology changes, new software called DataStage (IBM) has been added.

This enables the following:

- Improvements in data quality by correcting systematic errors in the data;
- Transformation of data into a simple structure.

As with all the MMIM servers, the DataStage server has only restricted access by system administrators.

The key extra process in this Phase 3 change is data 'passing' through DataStage. Data will be sent from the LRR at each site to the "DataStage server" where it is transformed and cleansed (its quality improved and validated) in transit, and then sent into the LRR at the site. As with all the data flows, identifying data is always separated from clinical information. All these operations on the data are performed in computer memory and not stored.

1.6. Improving the Quality of Source Data

In MMIM Phase 1 source data was accepted as provided by participant sites with the project being in a pilot phase. With MMIM having passed acceptance testing for the pilot and entering Phase 2 and 3, it was clear that the databases and systems used for clinical data collection vary at local sites and often are not robust enough to ensure high quality and consistent data. As a simple example, the field ‘sex’ may be stored as ‘M/F’, ‘male/female’, ‘0/1’ or the field may contain a mixture of all 3 or other characters such as ‘?’.

The project disease/tumour team leaders (in particular the colorectal cancer group) have been working with colleagues in other hospitals and disease areas to standardise data fields and collection processes as far as possible.

Further MMIM has been working cooperatively with groups such as the Cancer Council Victoria (CCV) which has through Victorian state funded projects such as the Victorian Cancer Outcomes Network (VCON) been trialing standardised cancer data capture models.

All of these initiatives will in time contribute to standardised and better quality data for cancer and other diseases of relevance to researchers wishing to utilise MMIM.

1.7. Record Linkage - The Unique Subject Index

The Unique Subject Index (USI) is the key element in linking patient records across disparate data sources within and across institutions. It ensures compliance with privacy.

Linking patient/subject records and assigning USI identifiers to data allows patients to be linked across multiple institutions and databases while also observing legal, ethical, privacy and data ownership constraints

The USI is developed based on matching of six key demographic data items:

- Surname
- Given name
- Middle name or Initial
- Date of Birth
- Gender/sex
- Digits 5 to 9 of the Medicare Number

The software checks new records for a match against existing subjects, using probabilistic matching and a score is assigned on the basis of match / non-match for each data item. “Fuzzy logic” is used for transpositions, soundex matches, common “dummy” names (e.g. Babe of, Twin 1). Manual checking of subjects in the “grey area” between thresholds can be undertaken by the data owners.

1.8. The metadata management – business glossary

The MMIM system provides the ability to search for the information in MMIM and discover whether the required data is available - the metadata as opposed to access the data itself – the technical data. MMIM users can search and discover the

- clinical areas covered (diseases & database purpose)
- sites contributing
- types of data collected (pathology, procedures, genetic)
- detail of data elements
- have enough information to request access

MMIM chose IBM's Websphere Business Glossary (WBG) to provide terminology management capabilities. Definitions of standard terms, attaching standard terms to items (databases, tables, or columns), defining the hierarchies of terms, specifying the preferred terms, synonyms and having categories of terms with hierarchies have been implemented. This functionality is important in search and discovery of metadata as users may use non standard or non familiar terminology (or may misspell words) when describing items. For example, "date of birth" may be described as "dob", "gender" may be described as "sex", etc.

The Business Glossary has open access from the Internet and searching for information can occur by browsing, by drilling down the 'trees' or searching using keywords. interpret the data fields, to run queries and to understand the data models so they have the knowledge to join data across databases.

1.9. Addressing the Issue of Privacy

*MMIM infrastructure + processes***:** The project obtained independent legal advice from lawyers and privacy experts at all stages to ensure that measures taken to protect privacy continue to be timely and relevant as the project grows.

Site and project specific Human Research Ethics Committee approval has been obtained for all participating sites, as a prerequisite to proceeding with implementation of MMIM at participating sites. All data outside of the hospital LRR is de-identified. All health data is de-identified. Log-ons and passwords are used, Virtual Private Networks are utilised for transmission of data with secure internet access. Researcher access is provided to specified tables in MMIM only on application with the research/purpose fully described and only after approval of the application by the MMIM steering committee and data is only available to researchers in de-identified form.

1.10. Intellectual Property

A Collaboration agreement that all participating sites must sign to join MMIM explicitly provides for recognition of both Background and Project Intellectual Property.

The project has a set of standard IP management and commercialization processes. Default IP positions are agreed. However, individual research projects are free to negotiate appropriate terms on a case by case basis.

2. Research Results

Examples of research outcomes to date the areas epilepsy and neuropsychiatry, (1-3) evaluating the sensitivity and specificity of FOBT compared to colonoscopy over a 25 year period, (4) and the evaluation of colorectal cancer patients in the areas of biomarkers and therapy. (5 – 8)

3. Discussion

3.1. Building the Federated Data Integration Infrastructure

Phase One of MMIM successfully built the system infrastructure and federated database integration capability outlined in the methodology section above. This technology allowed the issue of patient privacy, patient record-linkage as well as researcher intellectual property to be protected. Acceptance of the system have meant a further data sets were successfully integrated in five Victorian public sector sites and two medical research institutes (Ludwig Institute for Cancer Research, and Walter and Eliza Hall Institute). This stage of the project involved data sets (clinical, genomic, tissue bank & biomarkers) for three disease types, namely, colorectal cancer, epilepsy and diabetes.

Phase two of MMIM is building on this success to include a further ten public health sites in Victoria and three states and territories. This phase involve a further four medical research institutes (*p*health CSIRO, Murdoch Children's Research Institute, Monash Institute of Medical Research, Neurosciences Victoria) and link more than 35 disease databases.

Phase 3 will expand the technology across the Regional Integrated Cancer Services within Victoria and the Metropolitan Melbourne Hospitals which together with the Victorian and interstate metropolitan public hospital sites implemented in Phase 1 and 2 will create the South eastern Australia part of an Australian Cancer Grid.

3.2. Powering Future Research

The MMIM Project has transformed the way that research can be undertaken giving approved and authorised researchers unprecedented access via the internet to a virtual repository of privacy–protected data not previously available.

From their own work stations researchers can now:

- Link genomic data to clinical / outcome data in Colorectal Cancer and Epilepsy;
- Test multiple hypotheses without collecting / recollecting their own data (with data owner approval);
- Research suitable pre-symptomatic testing and early intervention based on genotype data;
- Research genetic, genomic and proteomic profiles, factors that may influence treatment outcome, with respect to toxicity and potential benefit;
- Analyse summary/statistical information across institutions and from diverse databases.

The following table summarises the traditional approach to research data collection and assembling of databases with that offered by the MIIM Project.

Table 1. Comparative Advantages of Using MMIM

Using Traditional Standalone Research Databases	**Using the MMIM Data Grid**
Static Data at one point in time Often one-off 'dump' Linked once	*Dynamic* Data refreshed & updated Live link to clinical research data Links made on-demand
Often anonymised data	Codified–ethically re-identifiable in exceptional circumstances and privacy protected
Data leaves 'owners' control	Data owners control access
Minimum data sets	Minimum + legacy data
Must specify exact data/query up-front - can only answer one off specific research question	*Discovery* Research analysis / explorative 'Quality' type clinical reports Clinical Data collected at healthcare site Discovery tools and potential for iterative and exploratory research on a theme (to approved data)
Usually population based studies	Clinical and Biomedical Data collected at healthcare site and population data

3.3. Research, Publications and Teaching

The MMIM virtual repository has enabled collaboration between multiple institutions, both within and across disease specialties, and between clinical researchers, bio-informaticians and Information Technology specialists. This in turn has expanded research capacity and productivity as follows:

- Within and across disease groups;
- Between data owners;
- Between data owners and researchers across academic institutions in Australia;
- Between data owners and researchers overseas;
- New research data types –e.g. imaging (MRI)

At the same time MMIM can enable expanded teaching and training resources (data sets and tools) in the health research field (medical informatics, genomics, proteomics) which is being developed including the publication of thesaurus/glossary - metadata schemas and ontologies for medical research.

4. Conclusion

MMIM provides a privacy protected data grid for connecting heterogeneous and dispersed data for medical researchers.

The platform is operational and is growing and developing to become scalable and sustainable. It continues to incorporate new data and provide tools for researchers.

References

[1] Yerra R, Kilpatrick C, Matkovic Z, King B, Rafael N, Hibbert M, Brand C, and O'Brien, First Seizure Clinic Experience: Heterogeneity of patient population and prognosis, *Epilepsia* **46** suppl 8 (2005), 360.

[2] Yerra R, Kilpatrick C, Matkovic Z, Belbin S, Rafael N, Hibbert M, Brand C, and O'Brien T, Syndromic diagnosis of epilepsy in the First Seizure Clinic population, *Journal of Neurosciences*, **238** (2005), S150

[3] Mark Walterfang, Ronald Siu, Dennis Velakoulis. The NUCOG: validity and reliability of a brief cognitive screening tool in neuropsychiatry patients. *Australian and New Zealand Journal of Psychiatry* **40** (2006), 995-1002

[4] Macrae FA, Slattery MA, Brown GJ, O'Dwyer MA, Murphy C, Hibbert ME, St John DJB. Sensitivity of faecal occult blood testing (FOBT) for asymptomatic colorectal cancer and advanced adenomas over a 25 year experience in colorectal cancer screening. Gastroenterological Society of Australia Australian Gastroenterology Week, Brisbane 2005*, Journal of Gastroenterology and Hepatology* **20** (2005), A34.

[5] Lim E, Jones IT, Gibbs P, McLaughlan S, Faragher I, Skinner I, Chao MW, and Johns J. Subsite-specific colorectal cancer in diabetic and non-diabetic patients . *Cancer Epidemiology, Biomarkers & Prevention* **14** (2005), 1579–1582.

[6] Rodrigues J, Lim E, McLaughlin S, Faragher I, Skinner I, Chao M, Croxford M, Chapman M, Johns J, Gibbs P The Influence of Language Spoken on Colorectal Cancer Diagnosis and Management'. *ANZ Journal of Surgery* **76** (2006), 671.

[7] Gibbs P, McLaughlin S, Skinner I, Jones I, Hayes I, Chapman M, Johns J, Lim L, Faragher Re: Completion of therapy by Medicare patients with stage III colon cancer. . *Journal of the National Cancer Institute* **98**(21) (2006), 1582.

[8] Reiger NA, Barnett FS, Moore JWE, Neo E, Badahdah F, Ryan AJ, Ananda SS, Croxford M, Johns J, Gibbs P The quality of pathology reporting impacts on lymph node yield in colon cancer*, Journal Clinical Oncology* **25** (2007), 463.

Acknowledgments. This project was funded by the Victorian Government Department of Innovation and Regional Development via Bio21 Australia Limited. Phase 2 was funded by the Commonwealth Government of Australian, Department of Education, Science and Training, via their Systemic Infrastructure Initiative. Phase 3 is funded by the Victorian Government Department of Innovation and Regional Development via the University of Melbourne.