Electronic health records and disease registries to support integrated care in a health neighbourhood: an ontology-based methodology

Siaw-Teng Liaw[a,b,c], Jane Taggart[a], Hairong Yu[a], Alireza Rahimi[a]

[a]*University of New South Wales, Australia;* [b]*SW Sydney Local Health District, Australia,* [c]*Ingham Institute of Applied Medical Research*

Abstract

Disease registries derived from Electronic Health Records (EHRs) are widely used for chronic disease management (CDM). However, unlike national registries which are specialised data collections, they are usually specific to an EHR or organization such as a medical home. We approached registries from the perspective of integrated care in a health neighbourhood, considering data quality issues such as semantic interoperability (consistency), accuracy, completeness and duplication. Our proposition is that a realist ontological approach is required to systematically and accurately identify patients in an EHR or data repository of EHRs, assess intrinsic data quality and fitness for use by members of the multidisciplinary integrated care team. We report on this approach as applied to routinely collected data in an electronic practice based research network in Australia.

Keywords:

EHR, patient registries, data quality, routinely collected data, data repository, health neighbourhood, integrated care.

Introduction

Disease registries derived from Electronic Health Records (EHR) are widely used for chronic disease management (CDM). However, not enough is known about the quality of EHR-based registers in the UK (1, 2) and Australia (3). There are publications about large administrative or population health databases, but little about disease registries created from multiple EHRs. Even less information is available about whether improved quality of EHR-based disease registries improve CDM, patient safety or quality outcomes. In addition to research, the increasing use of EHR-based registries, created through "blackbox" extraction tools, for clinical care can increase the likelihood and scope of data errors and adverse events (4).

The design and development of EHR-based disease registries does not appear systematic or comprehensive (5). Aspects of quality of disease registries have been examined in the UK (2, 5) and through our own work on the consistency and quality of diabetes registries within an electronic Practice Based Research Network (ePBRN) in Australia (6).

Our proposition (7) is that a realist (8) and ontological (9) approach is required to systematically and accurately identify patients in an EHR (10), or data repository of information from multiple EHRs, and assess intrinsic data quality and fitness for use by stakeholders such as members of the multidisciplinary integrated care team or researchers (6). The realist approach (8) adopted for this evolving yet complex domain includes:

- **Context:** CDM, integrated care, evidence based practice;
- **Mechanisms:** systematic methods to assess and manage the quality of data integration, knowledge integration, clinical integration and interdisciplinary integration;
- **Impacts/outcomes:** improved data quality and fitness for use of disease registries, and, over the longer term, safety and quality of integrated care.

The ontological approach to EHR-based registers includes the collection of formal, machine-processable and human-interpretable representations of the entities, and the relations among those entities, within a defined domain (11). Ontologies also provide regimentations of terminology that can support the reusability and integration of data, thereby supporting the development of automated systems for data annotation, information retrieval, and natural-language processing (11). By incorporating defined rules, ontologies can generate logical inferences and control the inclusion/exclusion of relevant objects (12), such as the patient with a diagnosis of diabetes mellitus (DM), abnormal pathology (Path) test, DM medication (Rx), or a DM cycle of care Medicare service payment item (10). In addition, a formal ontological model of the domain data and metadata can specify a unified context which allows intelligent software agents to act in spite of differences in concepts and terminology from different primary care EHRs. This will enable the systematic development of automated, valid and reliable methods to extract, link and manage data as well as assess the data quality and semantic interoperability issues.

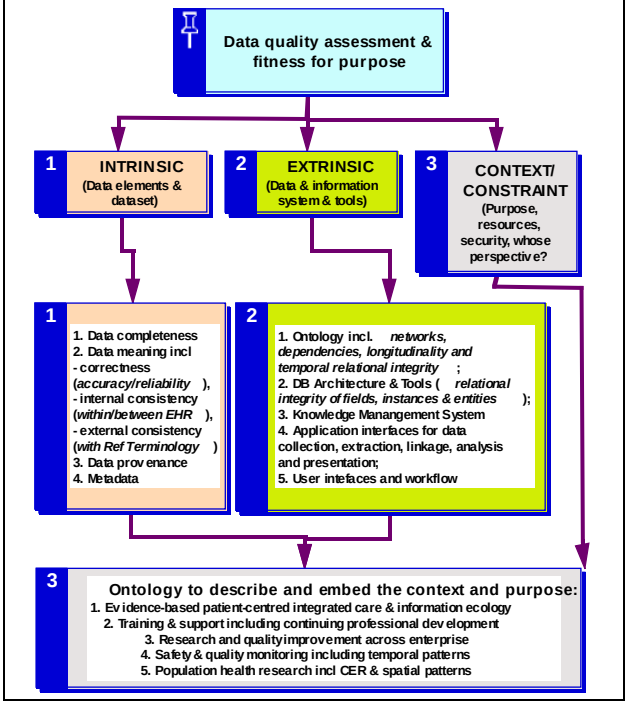
We have reported on our realist ontological approach ("Context-mechanisms-impact") to the quality of routinely collected data and integrated care, the relevant concepts and their relationships (13). The context is focused on the need for complete, correct, consistent and timely information about the cycle of care, risk factors, disease indicators, quality of life and patient satisfaction. The mechanism is the development and validation of ontologies to conceptualise and formalize the information and methods required to implement evidence-based integrated care in a range of contexts. This will allow the development of software agents to find cases to create disease registries, assess the intrinsic data quality and determine fitness for integrated care.

The quality of registries is influenced by the quality of EHR data, the case-finding system and associated quality processes, including currency and integrity, and the context such as clinical, insurance or other functions or objectives. Data quality

(DQ) is defined by the International Standards Organisation as: "the totality of features and characteristics of an entity that bears on its ability to satisfy stated and implied needs" (ISO 8402-1986, Quality Vocabulary). This "fitness for purpose/use"(14) definition is necessarily multidimensional requiring all intrinsic components and extrinsic associations of the entity to meet benchmarks and work together to achieve the purpose or meet the requirements.

An examination of the data quality literature (6, 15, 16) have led us to develop a more specific conceptual framework for data quality (DQ) and fitness for purpose (Figure 1).

Figure 1. Data quality & fitness for purpose framework



The framework comprises intrinsic, extrinsic and contextual dimensions, each with their concepts and relationships.

1. The intrinsic concepts cover the data elements and dataset, including the metadata, semantics (data meaning), provenance (who authored, where, when?) and constraints to the data meanings.
2. The extrinsic concepts cover the information system, including concept representation, ontology, temporal relationships system architecture and user interface.
3. The contextual determinants include the objectives of stakeholders such as the integrated care practitioner, resource constraints, security requirements, legislation, etc.

Data elements are assessed intrinsically in terms of consistency, correctness; data sets in terms of completeness and duplicate records (6). We are developing ontology-based tools to assess the information required to support integrated care in terms of timeliness and relational, historical and temporal integrity between concepts. Temporal and conceptual relationships may be dependent or independent factors. Relationships may be at a number of levels e.g. at the concept or table levels. The contextual determinants have been assessed qualitatively, aiming to guide clinical and organizational strategies to improve data quality to ensure fitness for purpose. The unified context will allow intelligent software agents to act in an environment of different concepts and terminology from different EHRs.

This paper will report and discuss this realist and ontological approach to developing automated, valid and reliable methods to define "cases" for a registry, manage data quality and determine fitness for purpose. We used the integrated care of diabetes mellitus in a health neighbourhood, as represented by the ePBRN, as a case study of the methodology of this work.

Materials and Methods

Setting: The ePBRN pilot group of 4 general practices has tested and validated the ePBRN data, processes and management in context, depending on the purpose. The internal validation of the ePBRN involved regular checking of the data and metadata using both automated and manual methods to examine the data repository. The data are also checked with probabilistic matching to assess the extent of duplicate patients and patients shared within the geographic region, the local health neighbourhood. The methodology was implemented with Microsoft SQL Server and an extension, Transact-SQL™ to link the server objects in the SQL Server with the heterogeneous datasets from multiple EHRs (17). The external validation of the ePBRN extraction tool involved a comparison against two other commercial data extraction tools (4).

Case-finding: The ePBRN ontological approach (10) used defined rules to generate logical inferences and control the inclusion/exclusion of the patient with a diagnosis of diabetes mellitus (DM), diabetes reason for visit (RFV), abnormal pathology (e.g. HbA1C, glucose tolerance test), diabetes medication (Rx) or glucose testing scripts, or a DM cycle of care item in the Medicare Benefit Schedule (MBS) (10). Following the query, the results were also analysed to exclude duplicate records/patients from the final result. This ontological approach was implemented and tested using SPSS and SQL, Each method acted as a control/validator for the other's accuracy. The benchmark was established with a manual examination of the results of SPSS and SQL queries on the smallest participating practice (Practice 1) contributing to the ePBRN data repository.

Data quality management: The conceptualization of the DQ ontology (Figure 1) included operationalising the reported core dimensions such as accuracy, currency and completeness (15) or completeness, correctness, consistency and timeliness (6, 16) and including duplicates (to account for aggregating multiple EHRs), temporal pattern (to account for the constantly changing clinical "big data") and timeliness which is important in integrated care. Validation of the conceptualization included discussions with practitioners and consumers of health care. The specification of the data quality ontology started with the definitions of completeness, consistency and correctness of data that we have reported previously (6).

Formalisation: To formalize the disease registry and DQ ontologies, we drew on the prevalent technical mechanisms and methodologies for ontology development, including knowledge acquisition, conceptualisation, semantic modelling, knowledge representation and validation (18, 19). Most used a layered approach (20) to incorporate clinical guidelines and rule-based approaches. The development tools used include: Protégé, a popular open source ontology editor and knowledgebase framework (http://protege.stanford.edu/); reference terminology (SNOMED-CT-Au); representation languages (Web Ontology Language (OWL), XML and RDF (Resource Description Framework)); query languages

(SPARQL Protocol and RDF Query Language); rules languages (Semantic Web Rule Language (SWRL)); logic ontology reasoners to provide automated support for reasoning tasks in ontology and instance checking through -ontopPro- (http://ontop.inf.unibz.it/), an ontology based data access (OBDA) application (21). The patient data, associated with instances of ontology classes or properties, is populated through -ontopPro-. The knowledge component of the infrastructure, related to conceptual terminologies defined by the specified ontology, was built using SNOMED CT-AU and Web Ontology Language (OWL: http://www.w3.org/TR/owl-features/) through Protégé. Details have been reported elsewhere (17) on how the RDF schema is mapped to logics to support formal semantics and reasoning. Formal semantics describes precisely the meaning of knowledge i.e. the semantics does not refer to subjective intuitions, nor is it open to different interpretations by different actors or machines (22). We used the layered ontology methodology to address semantic interoperability issues amongst different EHR in the ePBRN (23-27). This approach enables intelligent software agents to act in various semantic contexts in collaborative environments. We implemented and tested the DQ ontology, using SPSS and SQL tools, with the pilot ePBRN (N=95,056) data repository.

Results

Ontological approach to find cases for a diabetes registry

An overall prevalence rate of 2.8%, lower than expected for diabetes, was found for this pilot dataset. Table 1 shows data completeness of relevant indicators (RFV, Rx, Path) used for this paper and highlights that the ontological approach was more sensitive, finding more cases than a single database table query. The range of 0.2-4.8% for single factor and 1.1-5.7% for the ontological approach across practices, suggest that data quality is a significant factor. The pathology and medication tables contributed most. Case finding was improved, but the main limitation had been data quality dimensions like data completeness and consistency (5). The denominator was also important in assessing prevalence as some practices do not accurately represent active and inactive patients in the EHRs.

Table 1. Diabetes patients identified by diagnosis (RFV), HbA1C, medication, and ePBRN ontological approach

N = EHR flagged active patients	Practice 1 (N=3863)	Practice 2 (N=7028)	Practice 3 (N=23,162)	Practice 4 (N=30,717)	ePBRN (N=64,770)
Completeness of data:					
• All RFV (All DM RFV)	95% (4.3%)	87% (5.7%)	92% (4.9%)	99% (6.5%)	95% (5.8%)
• All Rx (All DM Rx)	80% (2.4%)	94% (8.4%)	96% (5.4%)	96% (6.6%)	95% (6.4%)
• All Path (HbA1C)	16% (0.8%)	61% (8.0%)	63% (1.3%)	66% (1.5%)	62% 2.4%)
• All 3 (RFV+Rx+Path)	82%	90%	90%	92%	90%
Diabetes indentified by:	N (%)	N (%)	N (%)	N (%)	N (%)
• Reason for visit (RFV)	37 (0.9)	231 (3.3)	387 (1.4)	787 (2.6)	1,442 (2.2)
• Diabetes medication	19 (0.5)	332 (4.7)	446 (1.9)	803 (2.6)	1,600 (2.5)
• HbA1c	8 (0.2)	334 (4.8)	468 (2.0)	809 (2.6)	1,619 (2.5)
• ePBRN ontological approach	43 (1.1)	403 (5.7)	602 (2.5)	1,042 (3.4)	2,090 (3.2)

Duplication and other dimensions of data quality

Table 2 shows up to 13% patient records matched across the participating EHR neighbourhood, suggesting that data quality assessment and management should include the extent of duplication of data with information sharing across the neighbourhood as well as within practices where there can be up to 3% duplication (Table 3). This has significance for clinical use of EHR data in integrated and shared care as well as secondary uses for research, population health and policy guidance.

Table 2. Record matching across general practices in a neighbourhood – shared patients

N=EHR active patients	**Pract 1 (N=3863)**	**Pract 2 (N=7028)**	**Pract 3 (N=23,162)**	**Pract 4 (N=30,717)**	**ePBRN (N=64,770)**
Practice (postcode)	**Records (%)**	**Records (%)**	**Records (%)**	**Records (%)**	**Records (%)**
Practice 1 (2176)		175 (2.5)	142 (0.6)	405 (13)	722 (1.1)
Practice 2 (2164)	173 (4.4)		327 (1.4)	691 (2.2)	1,191 (1.8)
Practice 3 (2171)	139 (3.4)	333 (4.7)		3,011 (9.8)	3,483 (5.4)
Practice 4 (2176)	400 (10)	692 (9.8)	3,005 (13)		4,097 (6.3)
Total	712 (18)	1200 (17)	3,474 (15)	4,107 (13)	9,493 (15)

Table 3. Record matching within general practices – duplicated records

Suburb (postcode)	EHR Active patients	Matched patients (%)	Matched records (%)
Practice 1 (2176)	3,863	10 (0.2%)	20 (0.5%)
Practice 2 (2164)	7,028	97 (1.3%)	198 (2.8%)
Practice 3 (2171)	23,162	220 (0.9%)	447 (1.9%)
Practice 4 (2176)	30,717	413 (1.3%)	830 (2.7%)
Total	64,770	740 (1.1%)	1,495 (2.3%)

Specifying and formalising the ontological approach

In addition to SQL tools, we have used the various ontology development tools mentioned to formalize the ontology work. The formal specification of the ontologies developed is available as Protégé files. Testing has being conducted with one of the participating practice (Practice 1) in the ePBRN, using – ontopPro- to map to the relational ePBRN data repository and implement the built-in reasoners. SPARQL and SWRL were used as the underlying query languages. However, this is the subject of another paper in preparation, which will also compare the utility and validity of SQL-based inductive versus ontology-based approaches and tools to create accurate patient/disease registries and assess/manage the quality of routinely collected data in the ePBRN data repository and its source EHRs.

Discussion

Research into the quality of routinely collected data in EHRs and EHR-based disease registries, especially in primary care, is an evolving field. While standards and benchmarks are being developed in this research domain, a realist and ontological approach is the most appropriate to understand what is being done in what context and with what impact, given that the processes and knowledge base are continually evolving, requiring ongoing monitoring, evaluation and reflection. The ePBRN research confirms this need to ground the research and development work in context and in the real world of health practice, where data is noisy and continually changing.

The ontological approach to case-finding identified a greater number of cases for inclusion in a disease/patient registry, highlighting the importance of this approach in the real world where data collection is suboptimal. Data quality management of aggregated information from multiple EHRs in a health neighbourhood to support integrated care must include the detection and management of duplicated records. Duplicates also lead to inaccurate public health and epidemiological research.

Ontologies deal with reality (**being**) and the transformation (**becoming**) of concepts as they interact with one another over time. An ontologically rich approach to the creation of patient registries from EHRs is essential to optimise accuracy (10). The effect of data quality is predictable as the disease registry is only as good as the EHR from which it is created – and there is much room for improvement in EHR data quality (6, 16). The improvement requires realist ecological approaches to the governance and provenance of data quality across the data cycle from collection to management to display and secondary use in other applications such as electronic decision support (16, 28). This approach recognises that the quality of electronic data collected as part of routine clinical practice is determined by more than just the GIGO – garbage in garbage out - principle. For instance, data models are influenced by the database management system, security and access management software, organisational processes for data collection and management, and the people in the organisation who enter and use data (4). The ePBRN foundational work reported here, along with others, has confirmed this to a significant extent.

As we validate the formal ontology tools developed in the ePBRN program and apply them to the development of fully automated methods to address the data quality of EHR and data repositories of ever increasing sizes, it is anticipated that this will build greater evidence for ontological approaches in the clinical and informational domains. The final tested ontologies and software tools can enable the systematic development of automated, valid and reliable methods to extract, link and manage data as well as assess/manage the data quality and semantic interoperability challenges.

Limitations:

This is a work in progress, evolving from a pilot phase to an established representative practice-based research network (and, given resources, a health information exchange to support evidence-based clinical practice). Having said that, the ePBRN foundational work has been systematic and robust in the methodology adopted:

1. to establish the ePBRN to reflect a local health neighbourhood with hospital, community health, general practice and other primary care services;
2. to refine and test the tools to extract, link and manage the data repository of routinely collected data in multiple EHRs; and
3. to make the transition from traditional management of "big data" from SQL and schematic relational databases to an ontological approach using semantic web principles and tools.

The data reported is neither representative nor timely; it is part of a pilot ePBRN to conduct our experiments to validate our methodologies with real world data from primary and secondary care settings. Our data across all projects shows that the quality of routinely collected data in EHRs is not only variable and suboptimal (6), but also continually evolving and changing with time. This emphasizes the need for cost-effective and validated automated methods to assess and manage data and information systems in a timely manner. The ePBRN program demonstrates that the challenge is great but surmountable.

Conclusion

The specification of a unified context to enable intelligent software agents to act, in spite of differences in concepts and terminology from different EHRs, will enable the systematic development of automated, relevant, valid and reliable me-

thods to extract, link and manage data as well as manage the data quality and semantic interoperability issues. This ontological approach to collecting, annotating, analysing and presenting clinical and scientific data is probably the only practical and sustainable solution to the information and data explosion. This is important to optimize the availability of good quality and relevant information to facilitate the safety and quality of integrated care as well as accurate and valid research.

Acknowledgments

The ePBRN research is supported in part by the University of New South Wales (UNSW) Major Research Equipment Infrastructure Initiative (2010, 2012), UNSW Medicine, Ingham Institute for Applied Medical Research and the Health Contribution Fund (HCF) Research Foundation (2013-14). We thank the participating general practices for their support and guidance, the conceptual input of Simon de Lusignan and Craig Kuziemsky, and contributions of co-researchers from the UNSW School of Public Health & Community Medicine and the following UNSW Research Centres: Primary Health Care & Equity, Health Informatics and Asia-Pacific Ubiquitous Health Care.

References

1. de Lusignan S, Sadek N, Mulnier H, Tahir A, Russell-Jones D, Khunti K. Miscoding, misclassification and misdiagnosis of diabetes in primary care. Diabet Med. 2012 2012 Feb;29(2):181-9.
2. Martin D, Wright J. Disease prevalence in the English population: a comparison of primary care registers and prevalence models. Soc Sci & Med 2009;68(2):266-74.
3. Ford D, Knight A. The Australian Primary Care Collaboratives: an Australian general practice success story. Med J Aust. 2010;193(2):90-1.
4. Liaw S, Taggart J, Yu H, de Lusignan S. Data extraction from electronic health records – existing tools may be unreliable and potentially unsafe. Aust Fam Physician. 2013;42(11):820-3.
5. Mehta A. The how (and why) of disease registers. Early Human Development. 2010;86(11):723-8.
6. Liaw S, Taggart J, Dennis S, Yeo A. Data quality and fitness for purpose of routinely collected data – a case study from an electronic Practice-Based Research Network (ePBRN). AMIA Annual Symposium 2011; Washington DC: Springer Verlag; 2011.
7. Liyanage H, Liaw S, Kuziemsky C, de Lusignan S. Ontologies to improve chronic disease management research and quality improvement studies – a conceptual framework. In: Aronsky D, Leong S, editors. Medinfo 2013; Copenhagen: Elsevier Press; 2013.
8. Pawson R, Greenhalgh T, Harvey G, Walshe K. Realist review - a new method of systematic review designed for complex policy interventions. J Health Serv Res Policy. 2005;10(Suppl 1):21-34.
9. Gruber TR. Toward principles for the design of ontologies used for knowledge sharing. Int J Human-Comput Stud. 1995;43(5-6).
10. de Lusignan S, Liaw S, Michalakidis G, Jones S. Defining data sets and creating data dictionaries for quality improvement and research in chronic disease using routinely collected data: an ontology driven approach BCS Informatics in Primary Care. 2011;19(3):127-34(8).
11. Rubin D, Lewis S, Mungall C, et al. National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge. OMICS (Summer). 2006;10(2):185-98.
12. Perez-Rey D, Maojo V, Garcia-Remesal M, et al. ONTOFUSION: ontology-based integration of genomic and clinical databases. Comput Biol Med. 2006;36(7-8):712-30.
13. Liaw S, Rahimi A, Ray P, et al. Towards an ontology for data quality in integrated chronic disease: a realist review of the literature. . Int J Med Inform 2013;82(1):10–24.
14. Wang RY. A product perspective on total data quality management. Communications of the ACM. 1998;41(2 (Feb)):58-65.
15. Wang R, Strong D, Guarascio L. Beyond accuracy: what data quality means to data consumers. J Management Information Systems. 1996;12(4):5-33.
16. Liaw S, Chen H, Maneze D, et al. Health reform: is current electronic information fit for purpose? Emergency Medicine Australasia. 2011 Feb 2012;24(1):57-63.
17. Yu H, Liaw S, Taggart J, Rahimi A. Using Ontologies to Identify Patients with Diabetes in Electronic Health Records. Poster/demo. International Semantic Web Conference 2013; 2013; Sydney Australia: Springer-Verlag Berlin Heidelberg.
18. Kuziemsky C, Lau F. A four stage approach for ontology-based health information system design. Artificial Intelligence in Medicine 2010 2010;50:18.
19. Ying W, Wimalasiri J, Ray P, Chattopadhyay s, Wilson C. An Ontology Driven Multi-Agent Approach to Integrated e-Health Systems International Journal of E-Health and Medical Communications (IJEHMC). 2010 2010;1(1):12.
20. Colombo G, Merico D, Boncoraglio G, et al. An ontological modeling approach to cerebrovascular disease studies: The NEUROWEB case. Journal of Biomedical Informatics. 2010;43(4):469-84.
21. Rodríguez-Muro M, Calvanese D. Quest, a System for Ontology Based Data Access. KRDB Research Centre for Knowledge and Data, Free University of Bozen-Bolzano, 2012.
22. Dean M, Schreiber G, Bechhofer S, et al. OWL web ontology language reference. W3C Recommendation. 2004.
23. Chalortham N, Buranarach M, Supnithi T. Ontology Development for Type II Diabetes Mellitus Clinical Support System2009 3 March 2011. Available from: http://text.hlt.nectec.or.th/ontology/sites/default/files/CRdm2css_0.pdf.
24. Ganendran G, Tran Q, Ganguly P, Ray P, Low G. An Ontology-driven Multi-agent approach for Healthcare. HIC2002.
25. Ganguly P, Ray P, Parameswaran N. Semantic Interoperability in Telemedicine through Ontology-Driven Services. Telemedicine & e-Health. 2005;11(3):8.
26. Hadzic M, Chang E, editors. Ontology-based multi-agent systems support human disease study and control. International Conference on Self Organization and Adaptation of Multi-Agent and Grid Systems (SOAS); 2005 Dec 11; Glasgow, UK. Amsterdam, The Netherlands: IOS Press.
27. Hadzic M, Dillon DS, Dillon TS, editors. Use and Modeling of Multi-agent Systems in Medicine. Proceedings of the 20th International Workshop on Database and Expert Systems Application; 2009.
28. de Lusignan S, Liaw S, Krause P, et al. Key concepts to assess the readiness of data for International research: Data quality, lineage and provenance, extraction and processing errors, traceability, and curation. IMIA Yearbook of Medical Informatics. 2011:112-21.

Address for correspondence

Professor Siaw-Teng Liaw

The General Practice Unit, Fairfield Hospital

PO Box 5, Fairfield, New South Wales 1860, Australia

Email: siaw@unsw.edu.au

Work phone: +61 2 96168520

Work fax: +61 2 96168400