

The Quality of Routinely Collected Data: Using the “Principal Diagnosis” in Emergency Department Databases as an Example

Siaw-Teng Liaw^{1,2,3}, Huei-Yang Chen¹, Della Maneze¹, Jane Taggart¹, Sarah Dennis¹, Sanjyot Vagholkar^{2,3}, and Jeremy Bunker^{2,3}

¹UNSW Centre for Primary Health Care and Equity

²Academic General Practice Unit, Fairfield Hospital

³UNSW School of Public Health & Community Medicine

Abstract

Objectives This paper aims to estimate the reliability of using “principal diagnosis” to identify people with diabetes mellitus (DM), cardiovascular diseases (CVD), and asthma or chronic obstructive pulmonary disease (COPD) in Firstnet, the emergency department (ED) module of the NSW Health Electronic Medical Record (eMR).

Methods A list of patients who attended a community hospital ED in 2009 with a specific “principal diagnosis” of DM, CVD, or asthma/COPD, or inferred based on possible keywords, was generated from Firstnet. This Firstnet list was compared with a list extracted from the underlying eMR database tables, using similar specific and possible coded terms. The concordance for an episode of care and for the overall was calculated. Patients on the Firstnet list who were admitted had their discharge summaries audited to confirm the principal diagnosis. The proportion of admitted patients correctly identified as having one of the chronic diseases was calculated.

Results The Firstnet list contained 2,559 patients with a principal diagnosis of DM, CVD, or asthma/COPD. The concordance (episode) of the Firstnet list with the eMR list were: 87% of CVD cases, 69% of DM and 38% of asthma/COPD cases. The audit of the discharge summaries of the Firstnet patients who were admitted confirmed the diagnosis of DM, asthma/COPD, and CVD for 79%, 66%, and 56% of the patients respectively.

Discussion An empirical method to examine the accuracy of the principal diagnosis in Firstnet is described. The incomplete concordance of diagnoses of the selected chronic diseases generated via different modules of the same information system raises doubts about the reliability of data and information quality collected, stored and used by the eMR. Further research is required to understand the determinants of data quality and develop tools to automate data quality assessment and management. This is particularly important with the increasing use of eMR in routine clinical practice and use of routinely collected clinical data for clinical and research purposes.

Keywords: Information; Data; Quality; EMR; EHR; Firstnet; Data Extraction; Discharge Summary

1 Introduction

Routinely collected electronic health care data, aggregated into large databases, are being used for audit, quality improvement, health service planning, epidemiological study and research. While these and other population health datasets are a potentially valuable resource for studying health outcomes, the quality of the information is not well understood. For example, a study of women who gave birth in a New South Wales (NSW) hospital in 2002 compared coded hospital discharge records with population maternal and obstetric health data sets. They found that there was a tendency to under-report maternal medical conditions; however specificities were high, indicating that false positives were uncommon. In a similar US study, coded data captured life-threatening peri-operative complications associated with vaginal hysterectomy; other complications were underestimated or missed entirely. The authors recommended that the use of coded data for outcomes assessments, especially of less life-threatening conditions, should be combined with other methods of recording to maximize validity. The quality of hospital discharge data appeared not to be good enough to support surveillance of hospital-acquired infection, as use of International Classification of Diseases, ninth revision (ICD-9) codes alone may misclassify bacterial infections. A study of infections among hospitalised rheumatoid arthritis patients found that misclassification varied with the specificity of the codes used and strength of evidence required to confirm bacterial infections. Combining codes with specified criteria identified “cases” with greatest accuracy. By using physician charts to validate linked administrative data, a Canadian study found that the sensitivity and specificity of the administrative data set was sufficiently high to support its use as a valid source of data to define cases of hypertension for surveillance and research purposes.

Studies of routinely collected primary care data suggest that the utility of these electronic data could contribute more to the process of health improvement if people working with these data fully understood the complexity of the context within which data entry takes place. These studies also demonstrated a need to know more about how to find relevant data, select appropriate search methods and ensure that the correct inferences are drawn. Presentation of data about mortality in their communities can enable general practices and primary care teams to reflect on their clinical policies, leading to improved quality of care and patient safety. At the health enterprise level, similar routinely collected information has been used by Kaiser Permanente to achieve better performance, as measured by quality metrics such

as prompt and appropriate diagnosis and treatment.

1.1 Context of study

The NSW Health Electronic Medical Record (eMR), established in the early 1990s, was envisaged as a single integrated patient-centred enterprise-wide clinical information system to enable clinicians to securely access clinical information regarding their patients via one point of access from anywhere and at any time. The eMR would become the primary source of information for health care. By providing real time access to and communicating legible information in conjunction with automated decision support tools, the eMR can reduce adverse events and improve patient safety and quality of care. All information recorded and modified will meet clinical, legal and administrative requirements, enabling audit for clinical and medico-legal purposes. The eMR program modules include Person Management, Scheduling, Powerchart for results reporting, orders, discharge referrals, Surginet for operating theatres and Firstnet for emergency departments (ED). All eMR modules have their own specific files and share files with the eMR. The eMR is now live in 43 sites, with 75,000 users across the state.

Firstnet has been implemented in the EDs of 22 public hospitals in NSW. The information recorded in Firstnet includes the urgency of treatment, demographics of patients, time of arrival, mode of arrival, time seen by doctor and nurse, the description of symptoms, the diagnosis, length of stay and discharge disposition. Firstnet enables ED clinicians to easily access patients' previous ED attendance and inpatient admission information e.g. pathology results and discharge summaries, from any Area Health Service (AHS) unit. Firstnet provides better audit trail capabilities than the old system (ED Information System, EDIS), as it records all changes made to patient records; however, the changes cannot be monitored easily by users. Another disadvantage is that only one “principal diagnosis”, such as diabetes mellitus, is recorded in the database as a Systematic Nomenclature of Medicine-Clinical Terms (SNOMED-CT) code; this limits the ability to identify cases of interest should it not be categorised as the “principal diagnosis”. Finally, there does not appear to be any explicit guidelines on data entry, including type of data such as symptom or diagnosis, and who enters the data such as doctors or nurses.

To date, there are no formal or reported studies into the data quality of the eMR and Firstnet to assess the accuracy and reliability of the information generated from the eMR directly or through its various modules. The aim of this paper is to estimate the accuracy of us-

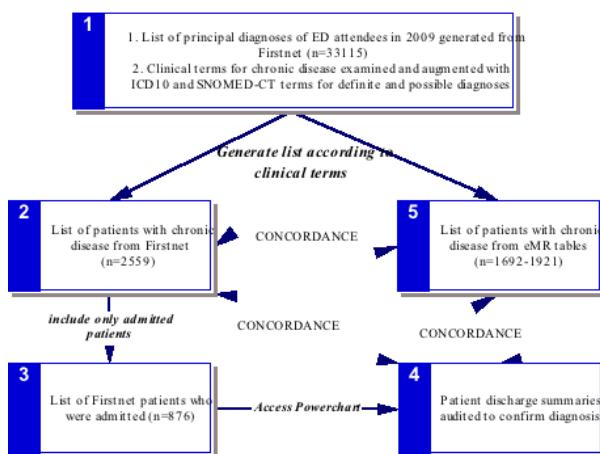


Figure 1: Firstnet data quality study protocol

ing “principal diagnosis” in Firstnet to identify patients with chronic diseases, using diabetes mellitus (DM), cardiovascular disease (CVD) and asthma or chronic obstructive pulmonary disease (COPD) as examples.

2 Methods

The accuracy of the “principal diagnosis” derived from Firstnet was examined by (1) comparing the Firstnet list with a similar list generated from the eMR, and (2) comparing the principal diagnosis of patients from the Firstnet list who were admitted with the diagnosis in discharge summary. Figure 1 summarises the methodology adopted.

The search terms (keywords) for the selected chronic diseases were developed with reference to clinical terms and terms used in the International Statistical Classification of Diseases and Related Health Problems, Tenth Revision, Australian Modification (ICD-10-AM) and SNOMED-CT. The list of “principal diagnosis” from Firstnet, which are coded with relevant terms from SNOMED-CT, was crosschecked by two clinician members of the team (DM and STL) to ensure that all synonyms and acronyms for DM, CVD, and asthma/COPD in the list were included as search terms. These Firstnet terms were augmented with relevant terms from ICD-10-AM and used to identify all patients with a definite diagnosis. However, some patients with the selected chronic diseases may have a less obvious but related principal diagnosis recorded in Firstnet. We therefore identified a second set of terms to identify possible diagnoses of the selected chronic diseases, for example, chest pain for CVD or shortness of breath for asthma/COPD. Table 1 below describes the terms that describe the definite and possible cases.

This enhanced list of terms was used to define the parameters to generate, from Firstnet, a list of patients

who attended the ED of a NSW hospital in 2009 and had a definite or possible “principal diagnosis” of DM, CVD, or asthma/COPD. This terms list was also used by eMR technical support staff to generate a similar patients list directly from the eMR. The two lists were then compared to determine the concordance.

The accuracy of the chronic disease label in the Firstnet list of patients with chronic diseases was verified with an audit of the discharge summaries of the ED attendees who were admitted to hospital. The discharge summaries of the corresponding admission were reviewed by clinician members of the research team (DM, SV, JB, and STL) independently, to confirm whether patients had presented with DM, CVD, and asthma/COPD. This process determined the concordance for this episode of care – the Concordance (episode). If the discharge summary for the corresponding admission was not available, the discharge summary of the subsequent admission was audited. The discharge summaries of the previous admission were checked if the discharge summaries of the corresponding and subsequent admission were not available. This process determined the concordance for the period - the Concordance (period). Finally, other reports (eg. Cardiology reports or Aged Care and Rehabilitation documents) were also reviewed in some cases.

3 Results

In 2009, a total of 2,559 patients with a principal diagnosis suggestive of DM (n=118), CVD (n=1,392), or asthma/COPD (n=1,049) attended ED at the hospital (Table 2). Of these patients, 876 were admitted; the proportions of admission were similar across the diagnoses, with the lowest proportions for asthma/COPD (31%) and highest for DM (39.8%).

Table 3 shows the comparisons between the Firstnet and eMR lists. Compared to Firstnet data (definite n= 1,059; total n= 2,559), fewer cases were extracted from the eMR tables (definite n= 764; total n= 1,921). The concordance was comparable among the chronic diseases: CVD (58% for specific episode and 81% for period), DM (60% and 74%), and asthma/COPD (51% and 69%). When the possible diagnoses (n=1,500) are included, the concordance was: CVD (87% for episode and 92% for period), DM (69% for episode and 85% for period), and asthma/COPD (38% for episode and 51% for period).

Table 4 presents the accuracy of the “principal diagnosis” as confirmed by an audit of the discharge summary. For the CVD patients who were admitted (n=504), 47% were confirmed as having CVD by episode and a further

| Disease | Terms that suggest Definite Diagnoses | Terms that suggest Possible Diagnoses |
|--------------|---|--|
| CVD | Acute coronary syndrome, Angina, Ischaemic heart disease, Myocardial infarct, ST segment elevation myoc, and NSTEMI; | Chest wall pain, Chest pain (discomfort/tightness), and Atypical chest pain. |
| DM | Diabetes, Hyperglycaemia, Hyperosmolar non-ketotic, Metformin, Insulin Injection, and NIDDM; | Hypoglycaemia, Metabolic acidosis, and Persistent hyper-insulinaemic |
| COPD/ Asthma | COPD (Chronic obstructive pulmonary disease), Asthma, Chronic obstructive airway, Chronic airflow limitation, Chronic airway disease/obstruction, Bronchiectasis, Bronchitis, Bronchial spasm, and Emphysema; | Shortness of breath, Respiratory distress, and Dyspnoea |

Table 1: Terms that suggest definite and possible diagnoses of specific chronic diseases

| Diagnosis | ED | Admitted to Ward | (%) |
|-------------|-------|------------------|--------|
| DM | 118 | 47 | (39.8) |
| CVD | 1,392 | 504 | (36.2) |
| Asthma/COPD | 1,049 | 325 | (31.0) |
| Total | 2,559 | 876 | (34.2) |

Table 2: Numbers of ED attendance and admission to ward by types of chronic diseases

9% were confirmed by period. A quarter (24%) of CVD patients had no information to confirm the CVD, while another 20% did not have any information on which to make a decision. For the DM patients who were admitted ($n=47$), 57% were verifiable by checking the corresponding episode, and checking other reports yielded an extra 21%. Only 2% of DM patients could not be confirmed as DM. The proportion of patients with missing discharge summaries was 19%. Among admitted asthma/COPD patients ($n=325$), 43% were confirmed as having asthma or COPD by the corresponding discharge summaries/other reports; an additional 23% were confirmed by documents from subsequent or previous admissions. Approximately 23% of asthma/COPD patients were not confirmed as having asthma or COPD. One-tenth (11%) of asthma/COPD patients did not have discharge summaries/reports available to confirm the diagnosis.

4 Discussion

As far as we can ascertain, this is the first reported independent study to examine the data quality of the eMR in NSW. A significant discordance between the list of patients with selected chronic diseases generated from Firstnet and the list generated directly from the eMR was found. We could have used more resource

intensive and comprehensive statistical methods, but opted for this more economical method which focused on the sensitivity, rather than the specificity, part of the concordance equation. There is no easy way of checking the accuracy of the patients' diagnosis if they were not admitted; this would have to be done by examining subsequent ED attendances or accessing the information systems of the patients' GP. This access may be possible in the future with a range of current initiatives into data linkage and information sharing in the health system.

A reason for the discordance between the reports generated via Firstnet and directly from the eMR tables, using SNOMED-CT codes equivalent to the clinical terms used in Firstnet is that the list of enhanced clinical terms developed could not be fully implemented and less specific SNOMED-CT codes had to be used to retrieve. This will therefore miss some patients with the diagnosis, resulting in a lower yield. This is an important finding which raises important considerations about data entry by a range of clinicians, data coding by coding clerks, and data extraction by information managers. The nexus between clinical practice and information management is a complex one requiring much consensus, standards and mutual understanding.

The discordance was compounded by the finding that the principal diagnosis for about half of the patients (43-57%) could not be verified by diagnostic information documented in the discharge summaries or other reports in Powerchart during the same admission into the hospital. Further examination of discharge summaries from previous and subsequent admissions improved the yield by about 10-20%. Missing discharge summaries/reports meant that the diagnoses of approximately 10-20% of patients could not be verified. Discharge summaries are most likely to be missing because they were not done in a timely manner by busy clinicians. However, possible technical or eMR reasons for missing information should also be examined by having routine informa-

| Comparison of ED attendances | Firstnet Diagnoses | Tables of the eMR | | | |
|--------------------------------|-----------------------|-------------------|---|------|------|
| | | N | n | % | n |
| CVD definite diagnoses | 228 | 132 | | 57.9 | 184 |
| CVD possible diagnoses | 1164 | 1077 | | 92.5 | 1100 |
| Subtotal | 1392 | 1209 | | 86.9 | 1284 |
| DM definite diagnoses | 70 | 42 | | 60.0 | 52 |
| DM possible diagnoses | 48 | 39 | | 81.3 | 48 |
| Subtotal | 118 | 81 | | 68.6 | 100 |
| COPD/Asthma definite diagnoses | 761 | 393 | | 51.6 | 528 |
| COPD/Asthma possible diagnoses | 288 | 9 | | 3.1 | 9 |
| Subtotal | 1049 | 402 | | 38.3 | 537 |
| Total definite diagnoses | 1059 | 567 | | 53.5 | 764 |
| Total possible diagnoses | 1500 | 1125 | | 75.0 | 1157 |
| Grand total | 2559 | 1692 | | 66.1 | 1921 |

Table 3: Comparing the Firstnet and eMR lists. * = The numbers of Concordance (episode) were included in Concordance (period).

| | Firstnet Diagnoses | Audit of discharge summary | | | |
|--------------------------------|-----------------------|----------------------------|---|------|------|
| | | N | % | % | % |
| CVD definite diagnoses | 198 | 55.1 | | 66.2 | 19.7 |
| CVD possible diagnoses | 306 | 41.5 | | 49.3 | 26.8 |
| Subtotal | 504 | 46.8 | | 55.9 | 24.0 |
| DM definite diagnoses | 33 | 63.6 | | 81.8 | 3.0 |
| DM possible diagnoses | 14 | 42.9 | | 71.5 | 0.0 |
| Subtotal | 47 | 57.4 | | 78.7 | 2.1 |
| COPD/Asthma definite diagnoses | 205 | 54.1 | | 79.0 | 8.8 |
| COPD/Asthma possible diagnoses | 120 | 24.2 | | 43.4 | 47.5 |
| Subtotal | 325 | 43.1 | | 65.9 | 23.1 |
| Total definite diagnoses | 436 | 55.3 | | 73.4 | 13.3 |
| Total possible diagnoses | 440 | 36.8 | | 48.4 | 31.6 |
| Grand total | 876 | 46.0 | | 60.8 | 22.5 |

Table 4: Audit of discharge summaries of admitted patients, comparing the Firstnet and eMR lists. * = The numbers of Concordance (episode) were included in Concordance (period).

tion quality monitoring procedures in place at all points in the information life-cycle. This will also prevent information-related errors and adverse events in health care.

What and where are the likely sources of inaccuracies in an information system? The data model, 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 are evolving constantly, depending on the prevailing needs. The data remains relatively constant. Evolving objects are prone to errors. The data model is important to promote information quality. For instance, a study of hospital discharge abstracts in New York and California found that about 80% of the secondary diagnosis codes indicating urinary tract infections were flagged as present on admission (POA), suggesting that the addition of POA indicators in Medicare claims would increase the positive predictive value (PPV) up to 86% and sensitivity up to 79% in identifying hospital-acquired catheter associated urinary tract infections (CAUTIs). The strategic use of related fields relevant to research questions can improve the accuracy and “fitness for purpose” of the information and information system.

What can we do to improve data quality? An automated process, building on this relatively simple and inexpensive methodology, will provide a cost-effective tool to support the routine monitoring of information quality among data repositories of different services across the range of settings of care. A framework to approach data quality assessment and management can include (1) user errors such as incomplete data entry, errors in spelling or coding and mal-compliance with organisational policy and protocols to support information quality; (2) technical errors such as corruption of the database architecture or data extraction (or report generation) algorithms and tools; and/or (3) environmental factors such as distractors and poor training and support to facilitate data entry or quality assurance to ensure that data collected is relevant, accurate, valid and reliable,

Promoting data quality and preventing data inaccuracies begins with careful information systems procurement, especially user requirements specifications, software requirements document and the hardware specifications. The data model which must be relevant and reflect the objectives of the health service. Tools and strategies are needed to support health services staff to enter data accurately and completely. Software design including database architecture, user-interface and drop down menus, is important to promote structured data entry. We need to know how to find relevant data, select

appropriate research methods and ensure that the correct inferences are drawn. Automated monitoring and audit of data quality and feedback is essential alongside workplace training and support in the consistent use of the eMR and associated tools. People working with these data need to understand fully the complexity of the context within which data entry takes place. Metadata are important: explicit statements are needed to explain the source, context of recording, validity check and processing method of any routinely collected data used in research. Education providers must incorporate informatics competencies in their curricula, aiming to graduate specialist health informaticans as well as eHealth-savvy clinicians and managers. Other challenges that need to be addressed in dealing with large routinely collected data sets include integrating systems where there is often no reliable unique identifier and between health (person-based records) and social care (care-based records e.g. child protection), informed consent and achieving appropriate levels of information security and privacy and legal and social issues related to health care policy, financing and professional practice.

Good data is essential to support a systematic Research and Development (R&D) program to provide an understanding of the strategies required to maintain a comprehensive and accurate picture of the health of individuals and communities, facilitating improved policy development, planning and implementation, and quality monitoring as well as addressing the costs of external data failure and the complementary costs of data-quality assurance. Relevant questions include: What are the guidelines for data entry such as whether to enter only diagnoses or to include symptoms? Are chronic diseases being under- or over-recorded as reasons for patient attendances at EDs? Are the instructions for list generation (or data extraction) pulling out too many false positives? The clinical course may have changed and the chronic disease-related reasons or diagnoses not confirmed during the admission. Are the reasons for the discordance related to the user or are they related to technical and design issues? Is the problem with how the database is designed and managed? Is the data model not consistent with the purpose of Firstnet? These questions warrant a systematic approach, covering sociotechnical domains, to examine data quality in the eMR and component modules.

Finally, this study highlighted the need for linkages between primary and secondary care systems. This study focused on the proportion of agreement as a measure of accuracy and completeness because the absence of this linkage made the collection of the required information to calculate kappa and specificity difficult

and expensive. This linkage should be promoted and supported because it will also improve the quality of information exchange to support optimum clinical handover between the levels of care.

5 Conclusion

This study tested a relatively inexpensive method to examine the accuracy of a module of the NSW eMR, which is applicable to other modules as well as to other area health services in NSW. The findings suggest that the data in the eMR and its Firstnet module is probably not accurate or complete enough to rely on for the various uses via eHealth tools to support safety and quality of care. This study suggests some socio-technical approaches, including careful implementation and change management strategies to encourage and support accurate and more complete data entry by time poor clinicians. There is a need for a systematic R&D program in data and information quality in health information systems.

Acknowledgements

The SSWAHS-Divisions of GP Executive Liaison Committee and key staff from SSWAHS Information Management and Technology Division and Fairfield Hospital. Co-author Jeremy Bunker passed away on 04 May 2011.

Conflict of Interests

None

References

1. Hadfield RM, Lain SJ, Cameron CA, Bell JC, Morris JM, Roberts CL. The prevalence of maternal medical conditions during pregnancy and a validation of their reporting in hospital discharge data. *Aust N Z J Obstet Gynaecol.* 2008 Feb;48(1):78-82.
2. Heisler CA, Melton LJ, 3rd, Weaver AL, Gebhart JB. Determining perioperative complications associated with vaginal hysterectomy: code classification versus chart review. *J Am Coll Surg.* 2009 Jul;209(1):119-22.
3. Moro ML, Morsillo F. Can hospital discharge diagnoses be used for surveillance of surgical-site infections? *J Hosp Infect.* 2004 Mar;56(3):239-41.
4. Patkar NM, Curtis JR, Teng GG, Allison JJ, Saag M, Martin C, Saag KG. Administrative codes combined with medical records based criteria accurately identified bacterial infections among rheumatoid arthritis patients. *J Clin Epidemiol.* 2009 Mar;62(3):321-7.
5. Quan H, Khan N, Hemmelgarn BR, Tu K, Chen G, Campbell N, Hill MD, Ghali WA, McAlister FA, Hypertension O, Surveillance Team of the Canadian Hypertension Education P. Validation of a case definition to define hypertension using administrative data. *Hypertension.* 2009 Dec;54(6):1423-8.
6. de Lusignan S, Hague N, van Vlymen J, Kumarpeli P. Routinely-collected general practice data are complex, but with systematic processing can be used for quality improvement and research. *Inform Prim Care.* 2006;14(1):59-66.
7. de Lusignan S, van Weel C. The use of routinely collected computer data for research in primary care: opportunities and challenges. *Family Practice.* 2006 April 1, 2006;23(2):253-63.
8. Sullivan E, Baker R, Jones D, Blackledge H, Rashid A, Farooqi A, Allen J. Primary healthcare teams' views on using mortality data to review clinical policies. *Qual Saf Health Care.* 2007 Oct;16(5):359-62.
9. Feachem R, Sekhri N, White K, Dixon J, Berwick D, Enthoven A. Getting more for their dollar: a comparison of the NHS with California's Kaiser Permanente Commentary: Funding is not the only factor Commentary: Same price, better care Commentary: Competition made them do it. *BMJ.* 2002 January 19, 2002;324(7330):135-43.
10. Robertson L. SSWAHS eMR Information. Sydney South West Area Health Service (SSWAHS), NSW Health 2009.
11. Zhan C, Elixhauser A, Richards CL, Jr., Wang Y, Baine WB, Pineau M, Verzier N, Kliman R, Hunt D. Identification of hospital-acquired catheter-associated urinary tract infections from Medicare claims: sensitivity and positive predictive value. *Med Care.* 2009 Mar;47(3):364-9.
12. de Lusignan S, Metsemakers J, Houwink P, Gunnarsdottir V, van der Lei J. Routinely collected general practice data: goldmines for research?

A report of the European Federation for Medical Informatics Primary Care Informatics Working Group (EFMI PCIWG) from MIE2006, Maas-tricht, The Netherlands. *Inform Prim Care.* 2006;14(3):203-9.

13. Wang RY, Pierce EM, Madnick SE, Fisher CW, editors. *Information Quality.* Armonk, NY: ME Sharpe Inc; 2005.
14. Wang RY, Storey V, Firth C. A framework for analysis of data quality research. *IEEE Transactions on Knowledge and Data Engineering.* 1995;7(4 (Aug)):623-40.

Correspondence

Professor Siaw-Teng Liaw,
General Practice Unit,
Fairfield Hospital,
PO Box 5, Fairfield, NSW 1860,
Australia.

siaw@unsw.edu.au