Protecting the privacy of individual general practice patient electronic records for geospatial epidemiology research

Soumya Mazumdar,[1] Paul Konings,[1] Michael Hewett,[1] Nasser Bagheri,[1] Ian McRae,[1] Peter Del Fante[2]

Abstract

Background: General practitioner (GP) practices in Australia are increasingly storing patient information in electronic databases. These practice databases can be accessed by clinical audit software to generate reports that inform clinical or population health decision making and public health surveillance. Many audit software applications also have the capacity to generate de-identified patient unit record data. However, the de-identified nature of the extracted data means that these records often lack geographic information. Without spatial references, it is impossible to build maps reflecting the spatial distribution of patients with particular conditions and needs. Links to socioeconomic, demographic, environmental or other geographically based information are also not possible. In some cases, relatively coarse geographies such as postcode are available, but these are of limited use and researchers cannot undertake precision spatial analyses such as calculating travel times.

Methods: We describe a method that allows researchers to implement meaningful mapping and spatial epidemiological analyses of practice level patient data while preserving privacy.

Results: This solution has been piloted in a diabetes risk research project in the patient population of a practice in Adelaide.

Conclusions and implications: The method offers researchers a powerful means of analysing geographic clinic data in a privacy-protected manner.

Key words: privacy, Geographical Information Systems (GIS), geospatial, general practice (GP) data collection, Australia, de-identified data, confidentiality

Public health researchers have traditionally relied on administrative databases, mandated outcome registries and population health surveys to inform their research. While each data source has its strengths, the individual patient record data held by primary healthcare providers is flexible to changing public health needs, burdens and issues. In Australia, most general practitioners (GPs) maintain these databases in their practices. GP databases from multiple practices can be aggregated to form what the Australian Institute of Health and Welfare (AIHW) has called a 'GP data collection'.[1] Such data collections are becoming increasingly common[1] with the proliferation of practice-level clinical and administrative e-databases, data extraction software, and data aggregation protocols. Not surprisingly, these data collections are being used to support clinical and business decision making[2] and are being used to inform population health studies.[3-5] Location information can enhance both clinical and business decision making, and inform public health research.[2,6] Ambulatory primary care data collection systems in the United States include the New York Department of Health's (NYDOH) 'Hub', which collects and analyses geographically attributed clinical data and is integral to NYDOH's decision-making process.[7] Similarly, a recently tested data protocol in Canada, implements spatial analyses such as spatial cluster analysis on data from 3,000 practices.[8] In the United Kingdom, with its near universal registration and extensive advanced computerisation of practices, data from GP practices have been used to map chronic obstructive pulmonary disease at the national scale[9] and type-2 diabetes at a local scale.[5]

Geographic GP practice data: the problem

In Australia, data extracted from GP data collections for research and analysis purposes rarely incorporate spatial information, with the exception of postcodes. The main reason for this is to ensure patient privacy by providing information in such a way that the information is not identifiable. Identifiable information should not be made available to researchers without explicit patient consent and, in general, obtaining consent of all patients in a practice is not practical, although in most cases opportunities are provided for patients to opt out of such data extractions. Since addresses can individually identify patients, they must be removed or otherwise obscured (as must other identifiers) from any data where each record represents a patient or a patient encounter if they are to

1. Australian Primary Healthcare Research Institute, Australian National University, Australian Capital Territory
2. Healthfirst Network, South Australia
Correspondence to: Dr Soumya Mazumdar, Australian Primary Healthcare Research Institute, Australian National University, Building 63, Cnr Mills and Eggleston Rds, Canberra, ACT 0200; e-mail: soumyamazumdar@yahoo.com
Submitted: December 2013; Revision requested: February 2014; Accepted: May 2014
The authors have stated they have no conflict of interest.
Aust NZ J Public Health. 2014; 38:548-52; doi: 10.1111/1753-6405.12262

be transferred to external parties for purposes such as research and surveillance. There are many approaches to the protection of privacy of unit record data including the addition or subtraction of a small random number to the value being protected or suppression of data from areas with small counts.[10] For example, a random number of years may be added to the age of people in unit record data to protect their privacy. However, a more common approach to protecting patient privacy while incorporating location attributes is areal aggregation. The principle of areal aggregation is simply that if there are sufficient numbers of records in a defined geographic area such that it is impossible to identify individuals, then individuals are protected. Census bureaus in most jurisdictions, including the Australian Bureau of Statistics (ABS), aggregate their geographic population data before release. For example, the smallest geography in Australian Statistical Geography Standard (ASGS)[11] is a Mesh Block, an aggregation of 30 to 60 dwellings. Areal aggregation also satisfies an important principle of privacy protection known as the *k*-anonymity principle,[12,13] where an individual's record can be 'confused' with at least *k* other records.

While 'on-the-fly' aggregation of clinical data and patient addresses to small geographies may achieve the optimal solution of providing data privacy while allowing meaningful analyses, such solutions are difficult to implement and not usually available. One solution is to extract patient postcodes from GP data collections, which effectively provides privacy protection through de facto aggregation, since postcodes in urban Australia have a more than enough people (median census population in postcodes is 3,500), contingent on the number of attributes associated with each person.

In Australia, several utilities exist that can extract clinical data from practice management databases but most do not extract geography, and those that do usually extract postcodes.[1] These tools include GRHANITE (GeneRic HeAlth Network Information Technology for the Enterprise),[14] Canning Division Tool and Pen CAT.[15] The lack of geography in data extracted from GP practices may in part be because the clinicians, researchers or decision-makers using GP practice data have not needed the spatial dimension of the patient data. Thus for example, while the GRHANITE data extraction and aggregation tool has been used to create data collections with patient postcode,[1] the Collaborative Network and Data Using IT (CONDUIT)[1] program, a nationwide surveillance program utilising the tool,[3] does not incorporate or collect any patient geography. However, other tools do extract postcodes that allow the production of spatially explicit reports[2] at this relatively broad level.

Postcodes as a geography have a number of drawbacks.[16] Their geographical representations within the ABS geographical framework – Postal Areas – are only approximations of the underlying postcodes;[17] they change over time, cover immense areas in rural Australia and may be too large in some densely populated urban areas for observing small area geographical variations. Further, as they are defined for postal purposes, many do not accord with social or political boundaries, and some have areas that are not contiguous. However, with the current suite of available practice data extraction tools, a researcher often has no option but to use the postcode geography.

Data custodians may consider requests from researchers for spatial health data at a finer precision than postcodes, since such parties may be considered as 'trusted' or 'semi-trusted' as opposed to release of data to the general public.[8,18] Indeed, there exists a privacy-access continuum with individually identifiable restricted data at one end and aggregated publicly accessible data at another. Researchers may need individual-level data but, unlike GPs, may not require identifiable individual data. Many geospatial research problems can be successfully addressed with data at a reasonable degree of aggregation.[19] However, currently in the Australian GP practice data context, there is no opportunity to release data at a geographic precision between patient address and postcode. That is to say that it is an all-or-nothing approach – release data (albeit to trusted users) that includes individually identifiable addresses, or limit the location accuracy to postcode. Having the flexibility to choose a geography that is appropriate for a given project is important. An appropriate level often lies between address and postcode, since this can both refine the analyses and retain confidentiality.

Geographic GP data: a possible solution

One compromise approach requires that either the researcher or a computer program have access to the address data and the clinical data, but not both at the same time. Therein lies the solution. In fact, for some time researchers have realised that the key to de-identifying individual level or "unit record" data is the separation of identifying information such as names and birthdays from other clinical and demographic information.[5,20] In the Australian context, there is a large body of literature on the process of de-identification and linkage of administrative data.[21-27] In addition to separating the identifying information from the clinical data, these processes seek to link various datasets together while separating the linkage process from the datasets. The identifying information are removed, linked and encrypted. Unit records in the datasets are then assigned the encrypted keys that can be used to link them together on the fly, or they can be analysed independently if so desired without any breach of confidentiality. In addition to identifiable information being separated from clinical data, there is role separation between individuals who work with the clinical data and those who work with the identifiable information such as addresses. Such protocols have been used for some large surveys.[23] Of the various administrative data linkage units in Australia, such as Data Linkage Western Australia, few incorporate geography beyond postcodes[24] and these usually do not link to GP practice data (other than through *ad hoc* requests for linkage), although given the richness of such datasets their inclusion could greatly enhance the research potential of the data.

A similar approach has also been adopted in the GP practice spatial-data domain in a spatial analysis project in Tower Hamlets, London. In this project a pseudonymised identifier was created for each record in the databases[5] of participating GP practices. The clinical data attached to the identifier and the postcodes attached to the identifier were extracted separately and reattached later for analysis. Note that six-digit postcodes in the UK are only slightly larger than ASGS Mesh Blocks. A similar exercise in Australia would have to be done at a geography with fine enough a resolution to achieve the required precision of analysis – a criteria that, as discussed above, Australian postcodes will often not satisfy.

Since the location information is to be separated from the clinical information and assigned unidentifiable keys, finer precision location information held separately from

the clinical data does not additionally compromise the data over coarser precision data, as long as the geography is suitably coarse when reintegrated with the clinical data. For example, the use of addresses over postcodes does not in any way additionally compromise the data, as long as it is kept separately. In the section that follows, we describe a methodology we have developed and implemented that relies on the principle of separation and de-identification described above to analyse GP practice geospatial data in Australia and that allows external researchers to implement spatial analyses.

Method

The methods described here are those implemented by a team at the National Centre for Geographic and Resource Analysis in Primary Health Care (GRAPHC) in the Australian Primary Health Care Research Institute at the Australian National University. They are mostly described in a generalised form as the methods can be implemented in many ways but, where necessary, the terminology of the GRAPHC implementation is used for convenience.

In a typical scenario, a researcher wishes to analyse practice data from one or more general practices. The researcher registers with a secure dedicated server maintained by the GRAPHC team, after obtaining appropriate ethics clearances. The general practice data extraction process typically results in two datasets: de-identified clinical data; and the patient-identifying data that includes an address. Addresses from the patient data file are submitted to the secure dedicated server that returns a unique identifier for each address – GRAPHC uses a Globally Unique Identifier[28] (GUID), which has a very low likelihood of non-uniqueness. The identifier has no qualities that can be used to determine or derive an address. The new identifiers are known as GTAGs and, for convenience, we refer to them in this way throughout this discussion. As each identifier is returned, it is assigned to the appropriate de-identified clinical data record. In this way, the de-identified data can be handed to researchers and analysts without any directly embedded location data, but with a link that permits spatial referencing or spatial analysis at a later time.

Researchers can request particular spatial attributes or linked data for each relevant GTAG, or spatial analyses for batches of GTAGs. Addresses need to be geocoded to latitude and longitude within the secure server to permit the necessary analyses. Addresses and co-ordinates are never made available to external users, as this would breach privacy. The linked information requested could be:

- the region in which the address is located, generally in our context the ABS geographies such as the ASGS Statistical Area-1 (SA1) or the larger SA-2s or SA-3s;
- analyses such as the distances of a batch of GTAGs from a specific location such as a hospital;
- demographic, socioeconomic or other area-based population indicators available from GRAPHC (e.g. the GP-to-population-ratio in an area).

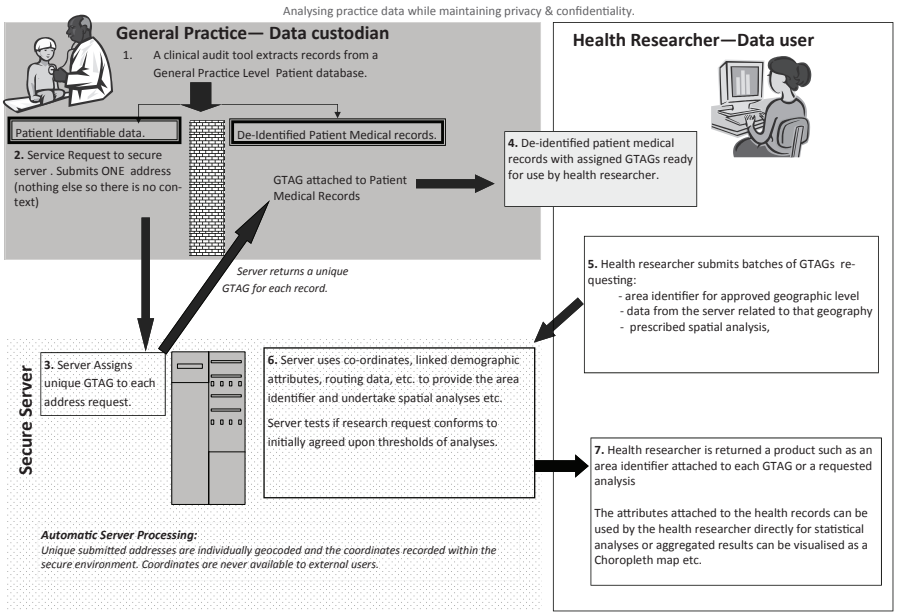
Results are sent back to the researcher, either attached to GTAGs so they can be linked directly to the clinical records or as a report (such as maps and graphs), or both. This method allows the researcher to implement analyses at a meaningful and relevant spatial precision without compromising the researchers' ethical clearances or patient confidentiality. Figure 1 summarises the workflow described above.

In most situations, the researcher operates independently of the GRAPHC team, and thus does not have access to the clinical data and individual address information at one time. In rare circumstances, when the researcher is based within the GRAPHC team (as in the example described below), the researcher is not permitted to access the GRAPHC server that implements the above protocol.

Results

The method described above has been implemented in the context of one GP practice in Adelaide.[29] The clinical information consisted of de-identified health records on 31,940 unique patients for the period January 2009 to June 2012. Patients who were not 'active' were removed, using the Royal Australian College of General Practitioners' definition that considers patients who have been seen once in the past 15 months to be active.[30] This resulted in a dataset of 14,969 (46.8%) active patients. GTAGs were attached to de-identified clinical data using the above protocol. The addresses were then geocoded on the secure server and geocodes attributed to SA1s.[11] This geography offers a much better spatial resolution than postcodes and is designed to be longitudinally stable. There are 55,000 SA1s in Australia with a median resident population of around 400. Most patients attending the practice lived relatively nearby, meaning that large patient populations were present in these SA1s, however some of the more distant SA1s had relatively small patient populations and to ensure confidentiality all SA1s with five or fewer patients were deleted. Of the 14,969 active patients, 97% were successfully geocoded to 282 unique SA1s. A researcher developed models of diabetes risk and calculated rates of diabetes within the SA1

Figure 1: Privacy protection workflow.



areas with this data.[29] A sample map from this research is displayed in Figure 2. It shows a smoothed map of diagnosed diabetes rates per 1,000 people within the clinical data set.

Discussion

The problems faced by researchers wishing to apply spatial analyses to data collected from GP practices while maintaining patient privacy are significant. The approach outlined in this paper provides a means of allowing researchers to attach a spatial marker (like SA1 or SA2 identifiers), or implement additional analyses to the clinical demographic data extracted from a practice database, without the need to access any personal identifying information. The methodology allows additional analyses with finer levels of spatial data such as travel distances than could be implemented by researchers having access to only the areal spatial identifiers. To our knowledge, this is the first time in Australia that GP practice data has been used for geospatial analysis at a fine resolution in a privacy-protected manner using remote servers. Beyond this pilot project, the system and paradigm described here is being tested on a number of additional projects.

As with any system that manages sensitive information, it is important to assess possible scenarios through which data re-identification may be possible.[8] Two such scenarios are relevant to this system:

Re-identification through researcher

This paradigm ensures that the address and the patient data never come together. Nevertheless, it is important to understand that the system cannot provide protection beyond what the researcher requests. It will, however, protect the researcher from inadvertent re-identification to a significant extent. If, for example, the researcher is permitted by the ethics committee to code addresses at the SA3 level (which are ASGS geographies with 30,000 to 130,000 people) and the researcher has registered this with the server administrators, then GRAPHC's GTAG System ensures that attempts to geo-attribute to smaller geographies will not be successful.

The system cannot protect from re-identification through the introduction of secondary data.[13] It is possible that the use of small areas could introduce potential opportunities for re-identification of a person with otherwise unusual attributes (such as age, country of origin, unusual health conditions or more likely a combination of these). In these cases, the ethics approval could specify that the de-identified clinical data be stripped of such secondary indicators. Alternatively, they could be aggregated (for example, to broader age groups) before geographic identifiers are added to comply with reasonable expectations of small area sample anonymity.

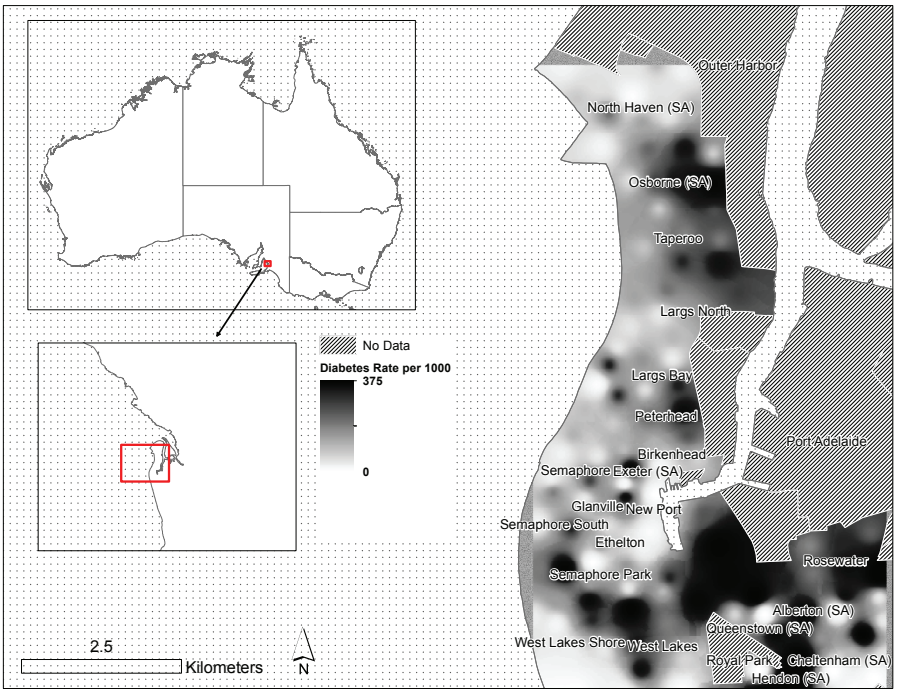
It is important to underscore the role of the researcher in protecting the privacy of health data. The onus is on the researcher to ensure the data are appropriately privacy protected at all stages. For example, the map in Figure 2 illustrates how a researcher may protect the privacy of data when publishing at small geographies. A choropleth map of diagnosed diabetes rates at SA1s may not necessarily be privacy protected. Since SA1s are small geographies, it may be possible to derive the number of cases in a geography from a map of rates and if, for example, only one case is found, to identify the person with additional information on the person/locality (in the case of this map, of course, no SA1 has less than 5 cases). However, if a map is 'smoothed' as in Figure 2, in addition to the displayed rates being statistically more stable than if the rates were displayed at small geographies,[31] identification of individuals is impossible.

Re-identification through unauthorised access

The separation of patient addresses from their context data is the key to the power of this paradigm. A data hack[32] would have to access the secure server and also obtain clinical information from the practice or the researcher to be able to breach the privacy of the data. This scenario is extremely unlikely, given that researcher computers are separate from the secure server. It would be much easier to hack the practice database with patient clinical/address information.[19]

Thus, this paradigm offers the research community the ability to analyse GP practice or other confidential data in a spatially explicit manner, without undue challenges to privacy.

Figure 2: Rates of diagnosed diabetes in the Lefevre Peninsula, Adelaide: Rates were smoothed using an Inverse Distance Interpolation algorithm. Numerators are number of diagnosed diabetes cases in the clinic data while denominators are all patients in the clinic.



Conclusions

The separation of identifiable and non-identifiable clinical information, the use of a secure system to provide geographic information and the ability to link geography back to the confidential data provides an efficient and secure means of enabling spatial analysis of clinical data. While researchers and ethics committees must always apply care to setting the limits of such analyses to protect privacy, this system opens a door to a range of research that was otherwise not possible. The methods applied here can equally be

applied to confidential data from sources other than clinical practices, and potentially have a very wide range of usage.

References

1. Australian Institute of Health and Welfare. *Review and Evaluation of Australian Information about Primary Health Care*. Canberra (AUST): AIHW; 2008.
2. Del Fante P, Allan D, Babidge E. Getting the most out of your practice. The Practice Health Atlas and business modelling opportunities. *Aust Fam Physician*. 2006;35(1/2):34-8.
3. Guy RJ, Kong F, Goller J, Franklin N, Bergeri I, Dimech W, et al. A new national chlamydia sentinel surveillance system in Australia: Evaluation of the first stage of implementation. *Commun Dis Intell*. 2010;34(1):319.
4. Liljeqvist GTH, Staff M, Puech M, Blom H, Torvaldsen S. Automated data extraction from general practice records in an Australian setting: Trends in influenza-like illness in sentinel general practices and emergency departments. *BMC Public Health*. 2011;11(1):435.
5. Noble D, Smith D, Mathur R, Robson J, Greenhalgh T. Feasibility study of geospatial mapping of chronic disease risk to inform public health commissioning. *BMJ Open*. 2012;2:1-11.
6. Dubowitz T, Williams M, Steiner ED, Weden MM, Miyashiro L, Jacobson D, et al. Using geographic information systems to match local health needs with public health services and programs. *Am J Public Health*. 2011;101(9):1664-5.
7. Buck MD, Anane S, Taverna J, Amirfar S, Stubbs-Dame R, Singer J. The Hub Population Health System: Distributed ad hoc queries and alerts. *J Am Med Inform Assoc*. 2012;19(1e):e46-50.
8. El Emam K, Hu J, Mercer J, Peyton L, Kantarcioglu M, Malin B, et al. A secure protocol for protecting the identity of providers when disclosing data for disease surveillance. *J Am Med Inform Assoc*. 2011;18(3):212-17.
9. Nacul E, Soljak M, Samarasundera E, Hopkinson NS, Lacerda E, Indulkar T, et al. COPD in England: A comparison of expected, model-based prevalence and observed prevalence from general practice data. *J Public Health*. 2010;33(1):108-16.
10. Hampton KH, Fitch MK, Allshouse WB, Doherty IA, Gesink DC, Leone PA, et al. Mapping health data: Improved privacy protection with donut method geomasking. *Am J Epidemiol*. 2010;172(9):1062-9.
11. Australian Bureau of Statistics. *Australian Statistical Geography Standard (ASGS)* [Internet]. Canberra (AUST): ABS; 2011 [cited 2013 Jun 30]. Available from: http://www.abs.gov.au/
12. Spruill NL. The confidentiality and analytic usefulness of masked business micro-data. *Proceedings of the American Statistical Association Section on Survey Research Methods*; 1983; vol XVIII .p. 602-7; American Statistical Association Alexandria, VA.
13. Sweeney L. k-Annonymity: A model for Protecting Privacy. *Int J Unc Fuzz Knowl Based Syst*. 2002;10(05): 557-70.
14. Liaw S-T, Boyle D. Secure Data Linkage and Information Sharing With GRHANITE. In: Grain H, editor. *HIC 2008 Australia's Health Informatics Conference*. Melbourne (AUST): Health Informatics Society of Australia; 2008.
15. Schattner P, Saunders M, Stanger L, Speak M, Russo K. Clinical data extraction and feedback in general practice: a case study from Australian primary care. *Inform Prim Care*. 2010;18(3):205-12.
16. Mullan N, Boyd J, Konings P, Butler D, Veenendaal B, West G, et al. Spatial Data Infrastructures and Geocoding of Health Data in Australia. In: ESRI, editor. *International Geospatial Geocoding Conference*; 2011. Redlands, CA; 2011.
17. Jones SD, Eagleson S, Escobar FJ, Hunter GJ. Lost in the Mail: The inherent errors of mapping Australia post postcodes to ABS derived postal areas. *Aust Geogr Stud*. 2003;41(2):171–9.
18. El Emam K, Jonker E, Arbuckle L, Malin B. A systematic review of re-identification attacks on health data. *PLoS One*. 2011;6(12):e28071.
19. Exeter DJ, Rodgers SE, Sabel CE. "Whose data is it anyway?" The implications of putting small area-level health and social data online. *Health Policy*. 2014;114(1):88-96.
20. Rodgers SE, Demmler JC, Dsilva R, Lyons RA. Protecting health data privacy while using residence-based environment and demographic data. *Health Place*. 2012;18(2):209-17.
21. Breen KJ. Consent for the linkage of data for public health research: Is it (or should it be) an absolute pre-requisite? *Aust N Z J Public Health*. 2001;25(5):423-5.
22. Churches T. A proposed architecture and method of operation for improving the protection of privacy and confidentiality in disease registers. *BMC Med Res Methodol*. 2003;3:1.
23. Holman CD. The impracticable nature of consent for research use of linked administrative health records. *Aust N Z J Public Health*. 2001;25(5):421-2.
24. Holman C, Bass JA, Rosman DL, Smith MB, Semmens JB, Glasson EJ, et al. A decade of data linkage in Western Australia: Strategic design, applications and benefits of the WA data linkage system. *Aust Health Rev*. 2008;32(4):766-77.
25. Lovett R, Fisher J, Al-Yaman F, Dance P, Vally H. A review of Australian health privacy regulation regarding the use and disclosure of identified data to conduct data linkage. *Aust N Z J Public Health*. 2008;32(3):282-5.
26. Kelman CW, Bass AJ, Holman CDJ. Research use of linked health data - a best practice protocol. *Aust N Z J Public Health*. 2002;26(3):251-5.
27. Cadilhac DA, Sundararajan V, Andrew N, Kilkenny MF, Flack F, Anderson P, et al. Using linked data to more comprehensively measure the quality of care for stroke – understanding the issues. *Australas Epidemiol*. 2013;20(1):15-9.
28. Microsoft Developer Network. *GUID Structure* [Internet]. Seattle (WA): Microsoft; 2013 [cited 2013 Jun]. Available from: http://msdn.microsoft.com/en-us/library/aa373931%28VS.85%29.aspx
29. Bagheri N, McRae I, Konings P, Butler D, Douglas K, Fante PD, et al. Undiagnosed diabetes from cross-sectional GP practice data: an approach to identify communities with high likelihood of undiagnosed diabetes. *BMJ Open*. 2014;4(7):e005305.
30. Royal Australian College of General Practitioners. *Standards for General Practices*. 4th ed. Melbourne (AUST): RACGP; 2012.
31. Beyer K, Tiwari C, Rushton G. Five Essential Properties of Disease Maps. *Ann Assoc Am Geogr*. 2012;102(5): 1067-75.
32. Armstrong MP, Rushton G, Zimmerman DL. Geographically masking health data to preserve confidentiality. *Stat Med*. 1999;18(5):497–525.