A Systematic Mechanism for the Collection and Interpretation of Display Format Pathology Test Results from Australian Primary Care Records

Douglas Boyle[1], Fabian Kong[2]

[1]*Rural Health Academic Centre, Melbourne Medical School, University of Melbourne, Australia*
[2]*Centre for Population Health, Macfarlane Burnet Institute for Medical Research and Public Health on behalf of the ACCES Collaboration, Melbourne, Australia*

Abstract:

Objectives: To demonstrate a mechanism for access to primary care patient data and the use and systematic interpretation of display format textual pathology test results held in primary care – in this case for Chlamydia surveillance.

Background: In Australia, it is common for pathology test results to be stored electronically in primary care medical record systems. At this time, there are few options for wide-scale access to such data that complies with privacy legislation, that supports record-linkage to other clinical databases and that can properly ensure textual results are correctly interpreted.

Methods: The GRHANITE™ software system (offering generic computer system interfacing, consent management, and privacy-preserving record linkage capabilities) was utilized to ethically extract patient consultation, pathology test requests and associated pathology test results for Chlamydia from Primary Care. In a sample of 10 practices (two different GP computer systems), results from 16 separate laboratories were obtained – all of which report test results differently. A rule-set for the parsing of this data was created and a C# program created to auto-generate rule-testing and rule interpretation SQL code.

Results: Utilising the rule-set interpreter, we were able to systematically verify that: All laboratories supplying data to the practices were included in the analysis, all tests for Chlamydia were identified (often several tests from each laboratory), all specimen types were accounted for and all results were correctly interpreted. In total 7,072 test results were analysed producing 236 distinct rules for the interpretation of the data from 16 laboratories. The rule-set generator created 3,102 lines of SQL code for rule verification and 9,135 lines for data consolidation. Later analysis against laboratory records confirmed the parsing strategy to be accurate in all cases.

Implications: GRHANITE™ has shown itself capable of meeting ethical requirements for data access and extracting data at a patient level (that is record-linkable) from a number of practices across multiple distinct general practice systems. The diversity of recording formats utilised by laboratories makes the electronic interpretation of textual pathology data difficult. The rule-set interpreter mechanism utilized here provides a reliable, extensible solution. We now have a sustainable and fast mechanism for future interpretation of any textual pathology data as held in Australian GP computer systems.

Keywords: Health Care Evaluation Mechanisms; Electronic Health Records; Epidemiologic Methods; Pathology; Knowledge Management

1 Introduction

This paper describes the use of a new technology (GeneRic HeAlth Network Information Technology

for the Enterprise – GRHANITE™) in the collection and utilisation of primary care data for health surveillance. It describes the application of GRHANITE™ in extracting patient-level (pseudonymous) records from primary care in a manner that permits central, systematic analysis of semi-structured narrative-type pathology results – in this case Chlamydia. Chlamydia notifications have increased by approximately 60% in the past five years from 36,222 in 2004 to 58,520 in 2008 with the majority among 15-29 year olds [1]. However, currently there remains limited information on changes in testing uptake over time and predictors of Chlamydia prevalence at GPs using the current passive surveillance systems. The hypothesis is that GRHANITE™ and our parsing techniques can resolve the problem of parsing highly variable pathology data in general practice and can deliver such data in accordance with ethics requirements for health surveillance, research and audit.

2 Background

In Australia, electronic clinical data is increasingly being requested from general practice for the purposes of audit, governance and research. This is in response to increasing requests for accountability and conformance to evidence-based medicine – the UK Quality and Outcomes Framework [2], [3] is a good example of this type of activity and in Australia, practices supply aggregated data for National Performance Indicators [4] and for the activities of the Australian Primary Care Collaboratives project [5]. In most cases, data is extracted using tools that aggregate patient information to minimise ethical issues. Whilst this is the correct approach where aggregated data are sufficient, in many cases, data complexity (e.g. narrative pathology test results) prevents automatic interpretation for aggregation and on occasion, linkage to other external clinical databases is a requirement. Where record linkage is required mechanisms for ethically managing patient confidentiality and patient consent must be found.

Since 2006, the University of Melbourne has been developing a tool (GeneRic HeAlth Network Information Technology for the Enterprise – GRHANITE™) specifically to address the issues of ethical access and linkage of data from health providers with a specific focus on primary care. This software (but not the parsing mechanism) has been described elsewhere [6].

In 2007, the University of Melbourne, School of Rural Health became partners in the Australian Collaboration for Chlamydia Enhanced Sentinel Surveillance (ACCESS) project [7] to deliver many aspects of the technical infrastructure. The following sub-components of this project are important in relation to the use of GRHANITE™:

1. Patient information regarding Chlamydia was required from laboratories, general practices and family planning clinics

2. Although practices utilising electronic pathology test result messaging could be selected, Chlamydia test results are held in narrative text with the pathologist frequently designing the text structure and layout

3. For comparison purposes, record linkage between laboratory and primary care records is required

4. An overarching requirement for health surveillance is that all patients are counted and as a result, obtaining consent is not practicable. To deal with this, very stringent mechanisms need to be in-place to guarantee that patient confidentiality cannot be compromised.

5. The technical model needs to be extensible beyond the initial pilot phase

In relation to item 2 above, utilising practice-based pathology results is not a trivial issue. In many cases, test results arrive at a general practice using the Pathology Information Transfer (PIT) format – a standard unique to Australia that has been reported as hindering adoption of more international standards such as HL7 [8]. The resulting test results (if stored electronically) are stored as formatted text. Also, although many laboratory suppliers and primary care systems now support HL7 pathology result messaging, test results can still be narrative in nature and may be formatted using the PIT formatting structure, plain formatted text, Rich Text Format (RTF) or ASCII text in various combinations [9]. Work is being undertaken by the National eHealth Transition Authority (NeHTA) to address these issues but this work is on-going [9].

Because they are semi-structured text records, a systematic mechanism cannot be developed to extract aggregated totals of positive and negative results for tests like Chlamydia without extracting the full text of the test result for central analysis.

Based on pilot work undertaken by the University of Melbourne, we ascertained that GRHANITE™ had the capacity to meet the ethical, data acquisition and record linkage requirements of the project and that we would have the means of parsing Chlamydia test results from primary care databases.

3 Methods

Using knowledge of the security, ethics, automated record linkage and consent provisions of GRHA-NITE™, ethics approval for the project was sought and obtained: RACGP National Research and Evaluation Ethics Committee (NREEC 07/017: Australian Collaboration for Chlamydia Enhanced Sentinel Surveillance (ACCESS) General Practices (GP Network). Interfaces to Medical Director V2 and Medical Director V3 had previously been developed. HCN (Health Communication Network) who own MD recently claimed 80% of GPs in Australia use MD – a 2006 paper found 73.1% [10]. Initial analysis with two pilot practices narrowed-down the data extraction of pathology results to those indicative of containing Chlamydia test results namely:

Test Name contains: 'CHLA' or 'TRACH' or 'NG AND CT' or 'CT AND NG' or 'CTPCR' or 'PCR CT' or 'NGCT' or 'PCR-CT' or the test result text contains: 'CHLA' or 'TRACH'. Once the data extraction protocols had been refined and tested, further practices utilising MD2 or MD3 were recruited. GRHANITE™ was installed as an automated data collection tool at these practices and the data securely forwarded to the project SQL data repository on a weekly basis. Regardless of whether the data was from MD2 or MD3, it was merged into one repository storing consultation, test request and test results for tests indicative of Chlamydia.

Initial analysis of the data showed that the textual nature of the test results would require a means of ensuring all possible Chlamydia tests and results were correctly interpreted. As more data arrived over time, more laboratories and different tests would require parsing. A systematic means of adding rules to deal with this was required.

Separate rules managing tests for each laboratory, each test name, each specimen type and each test result would need to be employed.

Figure 1 illustrates the negative-feedback loop mechanism employed to ensure all possible textual combinations indicative of Chlamydia testing were accounted for.
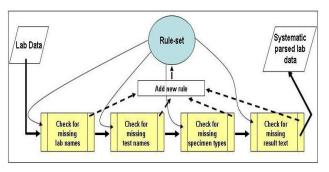

Figure 1: Using negative feedback to iteratively test the completeness of the rule-set

Because the rule-set had the potential to grow fast, a C# program was developed to automatically generate SQL query code embodying the four completeness tests and the code that ultimately generates the definitive set of Chlamydia test results. When programmers develop SQL or other code for parsing ever-increasing numbers of rules, there is great potential for cut-and-paste and other related human transcription errors to affect the final result. In this case, code that interprets an individual rule for each of the four tests was extensively tested and validated to ensure it definitely produced the desired results. Once the definitive SQL code for the four completeness tests had been generated, this code was embedded in the C# code generator so that definitive SQL code could be generated for each of the individual rules that needed applied. Further testing and validation were employed later to ensure that the specific implementation within C# produced the desired results. Because the recording of the specific rules is defined within a SQL table, the automated code generator is able to create the final SQL code without any human intervention beyond this initial set-up.

To clarify the process, the example of searching for missing laboratory names is described:

1. The set of defined rules hold the names of all laboratories that have previously supplied data to participating practices. The code generator uses this list to create a SQL query that searches for any laboratory name not present in the current rule set list.

2. The SQL Query for identifying missing laboratories is run

3. Any laboratories listed in the resulting query are manually inspected and rules are added into the rule-set to start searching for Chlamydia tests provided by these new laboratories

4. The C# code generator is re-run to generate a new SQL query to search for any laboratories still not

present in the current rule set list.

5. In a recursive manner, the above process is repeated until there are no laboratories not included.

6. Because the testing is recursive, even if there are human errors in recording the rules in the ruleset list, these are identified when the test is re-run.

7. This process is repeated for each of the four completeness tests and by undertaking this process, we determine by a process of elimination that all laboratories, all test names, all specimen types and all test results have been accounted for in our rule set.

4 Results

GRHANITE™ gathered comprehensive consultation and Chlamydia test data from ten sites used to validate the parsing technique. Automation means that new data arrives on a weekly basis supplying on-going result data conforming to the project study protocol and ethics approval. After consolidation of the test results from MD2 and MD3, 7,072 test results were found representing test results from the study population where they had a corresponding Chlamydia test request or the textual content of a test result was indicative of Chlamydia. A typical excerpt from the data is shown in Fig 2 below:

| Spec.Date Sample I.D Spec Type/Site Organ- |
| ism   Result   \ul0 \cf0 \par |
| 09/07/08   08040331   **Urine**\par |
| \ul \cf1 **Chlamydia trachomatis**\ul0   \cf0 |
| **Negative**\par |

Figure 2: A sample of text relating to a negative urine test result for Chlamydia

Using the C# rule-set code generator, rules corresponding to each laboratory and test result combination were compiled. This was an iterative process with the negative feedback loops ensuring that if we had missed any possible combination of tests and results, this was easily identified and added to our rule-set. By the end of the iterative process, we identified 236 rules covering 16 distinct laboratories. The rule-set generator automatically generated:

- 63 lines of SQL code that determines all laboratories are accounted for

- 352 lines of SQL code that checks all test names for each lab are accounted for

- 473 lines of SQL code that checks all specimen types are accounted for

- 1953 lines of SQL code that checks all possible test outcomes are covered

- 9135 lines of SQL code that uses the final definitive rule-set to generate the definitive result data

By the end of the iterative rule development process, the rule testing code confirmed that rules had been allocated to account for every single possible combination of tests and results encountered. Running the rules to generate outcomes data takes less than 3 minutes on a standard PC.

Subsequent work has permitted us to record link data from this analysis to the same data provided from pathology suppliers. Using this information, we have been able to confirm that specimen types and test results were correctly allocated during our parsing process in all cases – a description of this work will be the subject of a later paper.

5 Discussion

This paper is a short description of a pathology parsing mechanism employed in the context of health surveillance utilising general practice data. It describes some of the features that are important to a Chlamydia Surveillance project in particular but it is not in the scope of this paper to report clinical results or other dimensions of that project. GRHANITE™ is described here to report the conditions under which the laboratory data can be interpreted within the bounds of normal research ethics.

The results above make some strong assertions that the testing rules work. It is important to understand how such assertions can be made. If you consider a normal parsing technique where specific combinations of text are searched for there is room for error depending on the textual content of the test result. In the course of undertaking this work, Chlamydia test results that are combined with other tests have been seen frequently. Because of this, you cannot search for 'Chlamydia' and the word 'Negative' in a test result because although you may identify a test result as including the text 'Chlamydia' it may also include a result for 'Gonorrhea' and the 'Negative' result may be for 'Gonorrhea' not 'Chlamydia'. The test result may also include text such as: 'Chlamydia PCR test result to follow'. A successful interpretation of a test result must therefore use a means to specifically search for definitive textual strings such as 'Chlamydia trachomatis\ul0 \cf0 Negative\par' for a specific test from a specific laboratory.

Generating specific rules to interpret the results is not the end of the story. In this case, we have used automation to generate SQL code that systematically checks to see if any combination of laboratory, test, specimen wording or result wording has not been accounted for. Because of the automated nature of the code generation, human error is removed from the process of iteratively checking that we have not missed anything. There is still room for human error in the compilation of the rules themselves. It is possible for a combination of text to be wrongly classified as a 'Positive' result rather than a 'Negative' result. This is very simple to validate though. An independent researcher can manually scan the rule set looking for rules that are not logical. Because the rule set is in one table and is an ordered list of rules with a relatively small number of rows, this is very easily achieved.

The number of lines of code generated by the rule-set generator may sound like an issue in-terms of scalability. Instead, this is an example where a computer is the ideal tool for the job. As the number of practices later doubled, the number of lines of code generated increased by 56%. Should these rules be manually applied, time taken would be prohibitive. Instead, the rule-set runs in under 4 minutes. Also, the number of laboratories in Australia is finite. As a result, the methodology lends itself to computer management and is fully scalable.

Updates to study protocols can be (and have been) applied without returning to a practice. This is important in a number of situations for example where additional data needs collected or where a system supplier changes table structures or other recording methods. Further studies and data extractions can be applied without re-engineering GRHANITE™. Once installed, subject to practice permission, GRHANITE™ can be used by different health organisations and each organisation can have access to its own agreed data. This flexibility is hugely important – having to request software changes is a major inhibitor and cost limiter for research and audit.

As mentioned in the results section, a further analysis of the results of the parsing was done comparing the parsed data with definitive data from pathology suppliers. This work involved innovations surrounding privacy protecting record linkage and hence the details are the subject of a subsequent paper, but the very close matches found between the lab and GP results confirmed that the human step of generating and checking the rule set was a viable proposition. The logistics of performing record linkage between general practice and laboratories

normally precludes such activity, so having confidence that good results can be obtained without such validation is important.

It has been estimated by the National Serology Reference Laboratory that there are approximately 50 laboratories in Australia that conduct Chlamydia testing. Since the original work was undertaken, the GRHANITE™ software has now been connected to 38 clinics utilising 4 different computer systems distributed across all Australian states and territories. A further 140 installations are planned by July 2011. Because of the wide geographical distribution, we have had to generate parsing rules for over 40 laboratories. At this point (December 2010), we have 768 rules. Given an estimate that 20% of laboratories are not represented we estimate a full rule set covering all laboratories in Australia to require around 1,000 rules. This gives us confidence that the number of parsing rules is not going to escalate indefinitely and the technique is viable on a large scale.

In Australia, pathology test results are passed back to General Practices in the form of PIP or HL7 structured messages. Unfortunately, at this time the structured messages do not contain sufficient detail to extract tests like Chlamydia in an atomic manner as you would a Blood Glucose result. The result arrives as a formatted text string. This means that there is no way within a GP computer system to systematically identify Chlamydia test results. Some investigations have been made to parse the data for example by Pen Computing, Australia but due to the variability in pathology lab reporting, accuracy in determining positivity cannot be guaranteed (for the reasons reported above). Depending on the laboratory in question, this type of parsing has the potential to be very inaccurate. The technique employed by GRHANITE™ may not be required in countries where tighter application of interfacing standards exist, but the technique is applicable in any area where textual information is semi-structured and accurate parsing is a necessity.

In many countries, legislation and standard unique health identifiers make record linkage between clinical providers for research a routine operation. Privacy legislation, the current lack of a single health identifier (available in Australia but not yet widely implemented [11]) and the private nature of General Practice make this extremely difficult. At this time (December 2010) GRHANITE™ is the only tool that can systematically record link data across primary, secondary and tertiary care jurisdictions in Australia. Some linkages have occurred historically, but not on a wide-scale where linkage may be

required (and needs maintained) across tens or hundreds of clinical sites. The generic interfacing capability of GRHANITE™ has allowed it to interface to the following Australian GP computer systems and to systematically parse Chlamydia data from them: Medical Director V2, Medical Director V3, GENIE, Best Practice, Practix, Zedmed, MedTech32, Communicare. GRHANITE™ can interface to any database system able to communicate using the following database technologies: Oracle, SQL Server, Firebird, Interbase (all generations), ODBC, JDBC, OleDb. JDBC, OleDB and ODBC include text files, 4D databases, spreadsheets, Access databases, Visual FoxPro, dbase etcetera.

6 Conclusion

In many cases, aggregated patient data cannot provide answers to audit or research questions. Reasons for this include:

1. The structure of the required data may be indeterminate and only by accessing data centrally can rules for its interpretation be developed and applied

2. Linkage to other databases may be required

3. Where linkage is required, the sensitivity of data can be a major impediment to collaboration and participation. Even with appropriate governance and oversight, this may still be an issue.

Acknowledgements

References

1 Australian Government Department of Health and Ageing 'National Diseases Surveillance System. Number of notifications of Chlamydia infection, Australia, by age group and sex' Canberra, 2008 Available from http://www9.health.gov.au/cda/Source/Rpt_5.cfm Accessed 23/12/2010

2 Shekelle P. 'New contract for general practitioners' BMJ 2003; 326: 457 – 458

3 Lester H, Sharp DJ, Hobbs FDR, Lakhani M. 'The Quality and Outcomes Framework of the GMS contract: a quiet evolution for 2006' British Journal of General Practice 56 (525): 244-246 APR 2006

4 RACGP National Steering Committee – Quality Care, National Standing Committee 'Clinical indicators and the RACGP' Accessed 23/12/2010 http://www.racgp.org.au/policy/Clinical_indicators_and_the_RACGP_Policy.pdf

5 'Australian Primary Care Collaboratives (APCC) Monthly Measures' Available from http://www.apcc.org.au/about_the_APCC/program_results/program_measures/

6 Liaw ST, Boyle DIR 'Secure Data Linkage and Information Sharing with GRHANITE™ ™' HIC 2008 Australia's Health Informatics Conference ISBN 978 0 9805520 0 3 2008. Available from http://search.informit.com.au/home

7 'The Australian Collaboration for Chlamydia Enhanced Sentinel Surveillance (ACCESS) project'. http://www.access-study.org/

8 Bowden T. 'A Connected Health Sector – How Can We Get There From Here?' (16th January 2007) Guest Blogger Article, www.aushealthit.blogspot.com

9 'NeHTA Pathology Result Reporting Package (v1.0 Draft)' (1st September 2008) http://www.nehta.gov.au/component/docman/doc_download/535-pathology-result-reporting-package-v10-draft-readers-guide

10 McInnes, D., Saltman, D., & Kidd, M. (2006). 'General Practitioners' use of computers for prescribing and electronic health records: results from a national survey.' Medical Journal of Australia 185(2), 88-91

11 Medicare Australia – Healthcare Identifiers Service http://www.medicareaustralia.gov.au/provider/health-identifier/index.jsp

Correspondence

GRHANITE™:

Dr Douglas Boyle

Rural Health Academic Centre, Melbourne Medical School
PO Box 6500, Shepparton Vic 3632 Australia

Phone: +61 3 5823 4521
Fax: +61 3 5823 4555

http://www.ruralhealth.unimelb.edu.au
http://www.GRHANITE.com

dboyle@unimelb.edu.au

ACCESS Project:

Fabian Kong
Research Officer
Centre for Population Health

Macfarlane Burnet Institute for Medical Research and Public Health
85 Commercial Road, Melbourne
Victoria 3001, Australia

Phone: +61 3 8506 2327,
Fax: +61 3 9282 2138

http://www.burnet.edu.au
http://www.access-study.org

fabian.kong@burnet.edu.au