Conferences in Research and Practice in
Information Technology

Volume 142

Health Informatics and Knowledge Management 2013

Australian Computer Science Communications, Volume 35, Number 7

Health Informatics and Knowledge Management 2013

Proceedings of the Sixth Australasian Workshop on Health Informatics and Knowledge Management (HIKM 2013), Adelaide, Australia,
29 January – 1 February 2013

Kathleen Gray and Andy Koronios, Eds.

Volume 142 in the Conferences in Research and Practice in Information Technology Series.
Published by the Australian Computer Society Inc.

Published in association with the ACM Digital Library.

Health Informatics and Knowledge Management 2013. Proceedings of the Sixth Australasian Workshop on Health Informatics and Knowledge Management (HIKM 2013), Adelaide, Australia, 29 January – 1 February 2013

Conferences in Research and Practice in Information Technology, Volume 142.

Editors:

Kathleen Gray
Health and Biomedical Informatics Research Unit
Melbourne Medical School and Department of Information Systems
University of Melbourne
Melbourne, VIC 3010
Australia
Email: kgray@unimelb.edu.au

Andy Koronios
School of Information Technology and Mathematical Sciences
Division of Information Technology, Engineering and the Environment
University of South Australia
Adelaide, South Australia 5001
Australia
Email: Andy.Koronios@unisa.edu.au

Series Editors:
Vladimir Estivill-Castro, Griffith University, Queensland
Simeon J. Simoff, University of Western Sydney, NSW
Email: crpit@scm.uws.edu.au

Publisher: Australian Computer Society Inc.
PO Box Q534, QVB Post Office
Sydney 1230
New South Wales
Australia.

Conferences in Research and Practice in Information Technology, Volume 142.
ISSN 1445-1336.
ISBN 978-1-921770-27-2.

Document engineering, January 2013 by CRPIT
On-line proceedings, January 2013 by the University of Western Sydney
Electronic media production, January 2013 by the University of South Australia

The *Conferences in Research and Practice in Information Technology* series disseminates the results of peer-reviewed research in all areas of Information Technology. Further details can be found at http://crpit.com/.

Table of Contents

Invited Papers

Contributed Papers

Preface

We are pleased to present the papers from the Australasian Health Informatics and Knowledge Management (HIKM) Workshop held 31 January 2013 in Adelaide.

A highlight of this years HIKM Workshop was the opportunity to host invited speaker Riccardo Bellazzi from the University of Pavia, Italy, whose research interests include intelligent data analysis, biomedical data mining, bioinformatics, information technology infrastructures to support biomedical research, secondary use of clinical data in health and advanced telemedicine systems.

As well, this years Workshop featured two invited papers showcasing activities in the Workshop state, South Australia.

A surprisingly small number of papers was received for peer review this year seven in all, of which three were selected for the proceedings. We acknowledge all of the authors who submitted papers, we congratulate those whose papers were chosen and we hope that the others found the feedback helpful for refining their work.

This years papers are a microcosm of the breadth of research and development in the field of health informatics and health knowledge management in Australia:

Allan Baird heralds the arrival of the digital hospital in South Australia, with the advanced ICT system integrated in the new Royal Adelaide Hospital.

Julie Harris presents the next generation health-related data linkage management system being used in South Australia and the Northern Territory.

Shima Ghassem Pour and her colleagues present an approach incorporating quantitative methods for comparison between original and synthetic versions of longitudinal health datasets, performing clustering on synthetic data derived from the 45 and Up Study baseline data from New South Wales.

Lua Perimal-Lewis and her colleagues explore the quality of care relative to the ward outlier or inlier status of patients at a large general medicine service at a busy public hospital, presenting new findings about length of stay, discharge summaries, readmission rates and in-hospital mortality.

The team of Patel, Warren, Kennelly and Wai show how querying electronic medical records in a general practice can improve patient management, by identifying patients with persistently high risk of adverse outcomes and concurrent unchanged therapy during successive visits.

We thank our colleague Anthony Maeder for his assistance with reviewing, and our colleagues at the University of South Australia which hosted this years Workshop. We hope to see HIKM receive stronger support from the Australasian research community in the coming year. It provides an excellent opportunity for early-career and established researchers to work together to examine new research topics and methods in detail, and thereby strengthen the scientific foundations of the field.

The proceedings of HIKM are included in the ACM Digital Library and in Elseviers Scopus database.

Kathleen Gray
University of Melbourne

Andy Koronios
University of South Australia

HIKM 2013 Programme Chairs
January 2013

Programme Committee

Chairs

Kathleen Gray, University of Melbourne, Australia Andy Koronios, University of South Australia

Members

Vicki Bennett, Australian Institute of Health and Welfare, Australia
Yulong (Helen) Gu, Auckland University, New Zealand
David Hansen, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia
Anthony Maeder, University of Western Sydney, Australia
Tony Sahama, Queensland University of Technology, Australia
Louise Schaper, Health Informatics Society of Australia, Australia
Andrew Stranieri, University of Ballarat, Australia
Jim Warren, University of Auckland, New Zealand
Gerald Webster, Auckland University, New Zealand
Sallyanne Wissmann, Health Information Management Association of Australia, Australia
Dennis Wollersheim, La Trobe University, Australia

Organising Committee

Chair

Dr. Ivan Lee

Finance Chair

Dr. Wolfgang Mayer

Publication Chair

Dr. Raymond Choo

Local Arrangement Chair

Dr. Grant Wigley

Registration Chair

Dr. Jinhai Cai

Welcome from the Organising Committee

On behalf of the Organising Committee, it is our pleasure to welcome you to Adelaide and to the 2013 Australasian Computer Science Week (ACSW 2013). Adelaide is the capital city of South Australia, and it is one of the most liveable cities in the world. ACSW 2013 will be hosted in the City West Campus of University of South Australia (UniSA), which is situated at the north-west corner of the Adelaide city centre.

ACSW is the premier event for Computer Science researchers in Australasia. ACSW2013 consists of conferences covering a wide range of topics in Computer Science and related area, including:

- Australasian Computer Science Conference (ACSC) (Chaired by Bruce Thomas)
- Australasian Database Conference (ADC) (Chaired by Hua Wang and Rui Zhang)
- Australasian Computing Education Conference (ACE) (Chaired by Angela Carbone and Jacqueline Whalley)
- Australasian Information Security Conference (AISC) (Chaired by Clark Thomborson and Udaya Parampalli)
- Australasian User Interface Conference (AUIC) (Chaired by Ross T. Smith and Burkhard C. Wünsche)
- Computing: Australasian Theory Symposium (CATS) (Chaired by Tony Wirth)
- Australasian Symposium on Parallel and Distributed Computing (AusPDC) (Chaired by Bahman Javadi and Saurabh Kumar Garg)
- Australasian Workshop on Health Informatics and Knowledge Management (HIKM) (Chaired by Kathleen Gray and Andy Koronios)
- Asia-Pacific Conference on Conceptual Modelling (APCCM) (Chaired by Flavio Ferrarotti and Georg Grossmann)
- Australasian Web Conference (AWC2013) (Chaired by Helen Ashman, Michael Sheng and Andrew Trotman)

In additional to the technical program, we also put together social activities for further interactions among our participants. A welcome reception will be held at Rockford Hotel's Rooftop Pool area, to enjoy the fresh air and panoramic views of the cityscape during Adelaide's dry summer season. The conference banquet will be held in Adelaide Convention Centre's Panorama Suite, to experience an expansive view of Adelaide's serene riverside parklands through the suite's seamless floor to ceiling windows.

Organising a conference is an enormous amount of work even with many hands and a very smooth cooperation, and this year has been no exception. We would like to share with you our gratitude towards all members of the organising committee for their dedication to the success of ACSW2013. Working like one person for a common goal in the demanding task of ACSW organisation made us proud that we got involved in this effort. We also thank all conference co-chairs and reviewers, for putting together conference programs which is the heart of ACSW. Special thanks goes to Alex Potanin, who shared valuable experiences in organising ACSW and provided endless help as the steering committee chair. We'd also like to thank Elyse Perin from UniSA, for her true dedication and tireless work in conference registration and event organisation. Last, but not least, we would like to thank all speakers and attendees, and we look forward to several stimulating discussions.

We hope your stay here will be both rewarding and memorable.

Ivan Lee
School of Information Technology & Mathematical Sciences

ACSW2013 General Chair
January, 2013

CORE - Computing Research & Education

CORE welcomes all delegates to ACSW2013 in Adelaide. CORE, the peak body representing academic computer science in Australia and New Zealand, is responsible for the annual ACSW series of meetings, which are a unique opportunity for our community to network and to discuss research and topics of mutual interest. The original component conferences - ACSC, ADC, and CATS, which formed the basis of ACSW in the mid 1990s - now share this week with eight other events - ACE, AISC, AUIC, AusPDC, HIKM, ACDC, APCCM and AWC which build on the diversity of the Australasian computing community.

In 2013, we have again chosen to feature a small number of keynote speakers from across the discipline: Riccardo Bellazzi (HIKM), and Divyakant Agrawal (ADC), Maki Sugimoto (AUIC), and Wen Gao. I thank them for their contributions to ACSW2013. I also thank invited speakers in some of the individual conferences, and the CORE award winner Michael Sheng (CORE Chris Wallace Award). The efforts of the conference chairs and their program committees have led to strong programs in all the conferences, thanks very much for all your efforts. Thanks are particularly due to Ivan Lee and his colleagues for organising what promises to be a strong event.

The past year has been turbulent for our disciplines. ERA2012 included conferences as we had pushed for, but as a peer review discipline. This turned out to be good for our disciplines, with many more Universities being assessed and an overall improvement in the visibility of research in our disciplines. The next step must be to improve our relative success rates in ARC grant schemes, the most likely hypothesis for our low rates of success is how harshly we assess each others' proposals, a phenomenon which demonstrably occurs in the US NFS. As a US Head of Dept explained to me, ”in CS we circle the wagons and shoot within”.

Beyond research issues, in 2013 CORE will also need to focus on education issues, including in Schools. The likelihood that the future will have less computers is small, yet where are the numbers of students we need? In the US there has been massive growth in undergraduate CS numbers of 25 to 40% in many places, which we should aim to replicate. ACSW will feature a joint CORE, ACDICT, NICTA and ACS discussion on ICT Skills, which will inform our future directions.

CORE's existence is due to the support of the member departments in Australia and New Zealand, and I thank them for their ongoing contributions, in commitment and in financial support. Finally, I am grateful to all those who gave their time to CORE in 2012; in particular, I thank Alex Potanin, Alan Fekete, Aditya Ghose, Justin Zobel, John Grundy, and those of you who contribute to the discussions on the CORE mailing lists. There are three main lists: csprofs, cshods and members. You are all eligible for the members list if your department is a member. Please do sign up via http://lists.core.edu.au/mailman/listinfo - we try to keep the volume low but relevance high in the mailing lists.

I am standing down as President at this ACSW. I have enjoyed the role, and am pleased to have had some positive impact on ERA2012 during my time. Thank you all for the opportunity to represent you for the last 3 years.

Tom Gedeon

President, CORE
January, 2013

ACSW Conferences and the Australian Computer Science Communications

The Australasian Computer Science Week of conferences has been running in some form continuously since 1978. This makes it one of the longest running conferences in computer science. The proceedings of the week have been published as the *Australian Computer Science Communications* since 1979 (with the 1978 proceedings often referred to as *Volume 0*). Thus the sequence number of the Australasian Computer Science Conference is always one greater than the volume of the Communications. Below is a list of the conferences, their locations and hosts.

2014. Volume 36. Host and Venue - AUT University, Auckland, New Zealand.

2013. Volume 35. Host and Venue - University of South Australia, Adelaide, SA.

2012. Volume 34. Host and Venue - RMIT University, Melbourne, VIC.
2011. Volume 33. Host and Venue - Curtin University of Technology, Perth, WA.
2010. Volume 32. Host and Venue - Queensland University of Technology, Brisbane, QLD.
2009. Volume 31. Host and Venue - Victoria University, Wellington, New Zealand.
2008. Volume 30. Host and Venue - University of Wollongong, NSW.
2007. Volume 29. Host and Venue - University of Ballarat, VIC. First running of HDKM.
2006. Volume 28. Host and Venue - University of Tasmania, TAS.
2005. Volume 27. Host - University of Newcastle, NSW. APBC held separately from 2005.
2004. Volume 26. Host and Venue - University of Otago, Dunedin, New Zealand. First running of APCCM.
2003. Volume 25. Hosts - Flinders University, University of Adelaide and University of South Australia. Venue - Adelaide Convention Centre, Adelaide, SA. First running of APBC. Incorporation of ACE. ACSAC held separately from 2003.
2002. Volume 24. Host and Venue - Monash University, Melbourne, VIC.
2001. Volume 23. Hosts - Bond University and Griffith University (Gold Coast). Venue - Gold Coast, QLD.
2000. Volume 22. Hosts - Australian National University and University of Canberra. Venue - ANU, Canberra, ACT. First running of AUIC.
1999. Volume 21. Host and Venue - University of Auckland, New Zealand.
1998. Volume 20. Hosts - University of Western Australia, Murdoch University, Edith Cowan University and Curtin University. Venue - Perth, WA.
1997. Volume 19. Hosts - Macquarie University and University of Technology, Sydney. Venue - Sydney, NSW. ADC held with DASFAA (rather than ACSW) in 1997.
1996. Volume 18. Host - University of Melbourne and RMIT University. Venue - Melbourne, Australia. CATS joins ACSW.
1995. Volume 17. Hosts - Flinders University, University of Adelaide and University of South Australia. Venue - Glenelg, SA.
1994. Volume 16. Host and Venue - University of Canterbury, Christchurch, New Zealand. CATS run for the first time separately in Sydney.
1993. Volume 15. Hosts - Griffith University and Queensland University of Technology. Venue - Nathan, QLD.
1992. Volume 14. Host and Venue - University of Tasmania, TAS. (ADC held separately at La Trobe University).
1991. Volume 13. Host and Venue - University of New South Wales, NSW.
1990. Volume 12. Host and Venue - Monash University, Melbourne, VIC. Joined by Database and Information Systems Conference which in 1992 became ADC (which stayed with ACSW) and ACIS (which now operates independently).
1989. Volume 11. Host and Venue - University of Wollongong, NSW.
1988. Volume 10. Host and Venue - University of Queensland, QLD.
1987. Volume 9. Host and Venue - Deakin University, VIC.
1986. Volume 8. Host and Venue - Australian National University, Canberra, ACT.
1985. Volume 7. Hosts - University of Melbourne and Monash University. Venue - Melbourne, VIC.
1984. Volume 6. Host and Venue - University of Adelaide, SA.
1983. Volume 5. Host and Venue - University of Sydney, NSW.
1982. Volume 4. Host and Venue - University of Western Australia, WA.
1981. Volume 3. Host and Venue - University of Queensland, QLD.
1980. Volume 2. Host and Venue - Australian National University, Canberra, ACT.
1979. Volume 1. Host and Venue - University of Tasmania, TAS.
1978. Volume 0. Host and Venue - University of New South Wales, NSW.

Conference Acronyms

ACDC	Australasian Computing Doctoral Consortium
ACE	Australasian Computer Education Conference
ACSC	Australasian Computer Science Conference
ACSW	Australasian Computer Science Week
ADC	Australasian Database Conference
AISC	Australasian Information Security Conference
APCCM	Asia-Pacific Conference on Conceptual Modelling
AUIC	Australasian User Interface Conference
AusPDC	Australasian Symposium on Parallel and Distributed Computing (replaces AusGrid)
AWC	Australasian Web Conference
CATS	Computing: Australasian Theory Symposium
HIKM	Australasian Workshop on Health Informatics and Knowledge Management

Note that various name changes have occurred, which have been indicated in the Conference Acronyms sections in respective CRPIT volumes.

ACSW and HIKM 2013 Sponsors

We wish to thank the following sponsors for their contribution towards this conference.

CORE - Computing Research and Education,
www.core.edu.au

Australian Computer Society,
www.acs.org.au

University of South Australia,
www.unisa.edu.au/

University of Melbourne
www.unimelb.edu.au

University of Western Sydney
www.uws.edu.au

Invited Papers

The new Royal Adelaide Hospital – The Age of the Digital Hospital Dawns in South Australia

Allan H. Baird
ICT Consultant, New Royal Adelaide Hospital
Major Projects Office
Central Adelaide Local Health Network, Department of Health and Ageing, South Australia

AllanH.Baird@health.sa.gov.au

Abstract

The South Australian Government developed the Health Care Plan 2007-2016 to meet the health challenges of an ageing population, increasing incidence of chronic diseases, international workforce shortages and ageing infrastructure. The plan included an outline of the most significant single investment in health care in South Australia's history - the new Royal Adelaide Hospital. Efficient and effective application of the new Royal Adelaide Hospital Model of Care is reliant upon a robust ICT system which is fully integrated throughout the Facility and with primary and secondary health providers. The ICT element of the hospital is critical in ensuring that the return on the investment in such a new and complex facility will be achieved.

Keywords: new Royal Adelaide Hospital, Model of Care, Digital Hospital, Patient Systems.

1 Introduction

In May 2002 the then South Australian Minister for Health, the Hon Lea Stevens, announced the Generational Health Review (GHR) (Government of South Australia 2003) and appointment of the Review Committee, chaired by Mr John Menadue AO.

The aim of GHR was to develop a framework to guide the South Australian health care system over the next 20 years. The objectives are that the health care system, in partnerships with governments and stakeholders will:

- strive to maintain and improve the health of the population with an emphasis on addressing health inequalities
- ensure safe, accessible, efficient and effective health care

The principles and values underpinning GHR take into account the South Australian Government's health and social agenda commitment:

- improving the quality and safety of services
- greater opportunities for inclusion and community participation
- strengthening and reorienting services towards prevention and primary health care
- developing service integration and coordination
- whole-of-government approaches to advance and improve health status
- sustainability in delivery through ensuring efficiency and evaluation

Guided by the GHR, the State Government developed South Australia's Health Care Plan 2007-2016 (Government of South Australia 2007) to meet the health challenges of an ageing population, increasing incidence of chronic diseases, international workforce shortages and ageing infrastructure.

The plan included an outline of the most significant single investment in health care in South Australia's history - the new Royal Adelaide Hospital (RAH).

The new Royal Adelaide Hospital will provide world-class health care and facilities for South Australians.

Located in the CBD, construction of the 175,000 square meter facility started in September 2011 and is scheduled to be completed in 2016. The new hospital will replace the existing Royal Adelaide Hospital which opened in 1840.

The new Royal Adelaide Hospital will be the State's flagship public hospital and will be the cornerstone of the reformed South Australian health system,

The Facility will attract some 6,000 staff, have the capacity to treat over 400,000 outpatients per year and will provide overnight care to approximately 85,000 inpatient admissions per annum.

The hospital will have 800 beds (700 multi-day beds and 100 same-day beds) and all inpatient rooms will be single bedrooms. There will be more operating theatres, intensive care beds and emergency care capacity.

The new hospital will harness the latest in architectural design to create a healing environment for patients and a positive working environment for staff, all while minimising the building's environmental footprint. This will be achieved by ensuring all rooms have access to natural light and easy access to internal gardens on balconies and the roof.

The new Royal Adelaide Hospital will remain a major teaching hospital and will also be co-located with the new South Australian Health and Medical Research Institute, making the health precinct the hub of medical research in the State.

2 Technology Vision for the new Royal Adelaide Hospital

Efficient and effective application of the new RAH Model of Care is reliant upon a robust ICT system which is fully integrated throughout the Facility and with primary and secondary health providers.

The facility is being designed to incorporate a sophisticated and fully integrated ICT system including:

- incorporating an integrated ICT platform that facilitates the delivery of Clinical Services in accordance with the NRAH Model of Care;
- incorporating fully integrated ICT systems and platforms so that all departments, and each of the discrete components of the Clinical Services, Clinical Support Services and Non Clinical Support Services interface with each other
- including integrated booking systems, E-health record, clinical data and information transfer, bed management and supplies
- incorporating an ICT system that supports high levels of seamless access and use;
- incorporating ICT infrastructure which includes:
    - community interface capability (for example, communications with GPs and other service providers and hospitals);
    - a platform to interface between Clinical Services and Non Clinical Services;
    - accessibility of information, close to point of service;
    - equipment and technology that maximise care delivery and safety;
    - the ability to accommodate ongoing growth and change in the use of ICT, whilst maintaining full operation of services;
    - a platform for a range of ICT technologies such as wireless coverage for all technologies in all areas of the Facility;
    - technology throughout the Clinical Areas including at the bedside providing access to a number of systems including clinical information, communication, education and entertainment; digital health and hospital technologies, telemedicine and audiovisual systems locally and remotely;
    - integrated fire and life safety systems, data, paging, staff / patient /public management systems, building services, engineering management and security systems; and
    - provision of electronic signage, interactive information sites, options for biometric authentication, RFID and other tracking technologies

3 Integration with State-wide Patient Systems and Applications

At the same time as the design and construction of the new RAH is underway, the SA Department of Health and Ageing is developing and starting the roll-out of a number of ambitious and highly complex enterprise systems, which will embrace most of the State's hospitals and primary care facilities, including the new RAH. These systems include:

- **EPAS – Enterprise Patient Administration System**: will provide the foundation for delivering South Australia's state-wide electronic health record (EHR), which will:
    - standardise and consolidate the majority of patient information system data into a state-wide system
    - lay significant foundations for the EHR, a requirement under the national eHealth strategy
    - link relevant administrative and clinical processes, and
    - replace a number of SA Health's complex network of incompatible and outdated systems.
- **ESMI – Enterprise State Medical Imaging**: will support the objectives of the SA Medical Imaging objectives which are to achieve greater efficiencies and improve medical imaging services through the introduction of uniform enterprise picture archiving and communication (PACS) enterprise radiology information (RIS) and Voice Recognition systems across public hospitals.
- **State-wide Pathology**: over time, South Australia's government run Pathology services have been consolidated into a single service, and a single system will be rolled out to support the new organization.
- **State-wide Pharmacy**: already commissioned, this system is designed to provide a platform for the delivery of medication related services in a consistent manner across most State hospitals.

The simultaneous development and roll-out of these State-wide systems provide a challenging backdrop in order to ensure that the innovation in the design of the new RAH can be supported by systems that also need to be able to effectively operate in more traditional hospital settings.

4 The Dawn of Digital Hospitals in South Australia

Legendary American football coach, Tom Landry (1924-2000) is credited with the saying: "Setting a goal is not the main thing. It is deciding how you will go about achieving it and staying with that plan."

This mirrors the experience in SA Health circles, and specifically that of the new RAH.

Whilst there are highly aspirational goals, it is the detailed planning that ensures a facility will be able to fully realise the goals of the GHR and new Model of Care.

Therefore, not all of the end goals of the physical design and supporting ICT systems, either in house of State-wide, for the new RAH below will be implemented on Day 1, but may be progressively rolled out in line with the realities of funding and available resources, and when the supporting functionality is available from the respective systems:

- **Patient Arrivals**: For outpatients and day patients arriving at the hospital, direction to their appointments and where they will be admitted will be provided from wayfinding/patient information kiosks. In the case of outpatients, should the patient arrive too early for an appointment or the clinic is running late, the patient will be advised to delay presentation.
- **Patient Admissions**: There will be no central patient admissions office. Admissions will be undertaken at the patient bedside on a multi-purpose monitor that will serve as a patient entertainment system (PES) with free-to-air and pay TV, and internet access during the patient visit.

- **Patient Meals**: On admission, the patient meal requirements will be recorded, and subsequent to admission, the patient will be offered on the PES in their room a menu tailored to their dietary requirements. If the patient is transferred during their visit to the hospital, the delivery instructions for the patient meals will be automatically updated. On discharge, any outstanding meal orders will be cancelled.
- **Patient Tracking**: Initially it is planned to only track patients for whom the State has very high duty of care, e.g. those with Mental Health conditions and orders, and patients known or discovered to wander or abscond. RFID tags will be attached to these patients and details of the tags will be added to the patients' respective electronic records. On discharge, the records will be updated to indicate the removal of the tag. Asset Tracking and Management: RFID will also be used to track valuable assets in the hospital to ensure that the nearest asset to a required location can be retrieved as quickly as required. This is a widely used technology from the logistics industry. However, taken further, RFID technology can provide vital information on the hours of operation of a biomedical device providing early advice on the need to withdraw it from operation for periodic servicing and recalibration.
- **Video Conferencing/Telehealth**: The hospital will have state-of-the-art flexible learning and teaching facilities with extensive audio visual capabilities. The audio visual capabilities will also be implemented in operating theatres, selected procedural areas, consulting clinics and other areas like the Mortuary. These facilities will have the ability to engage health professionals within and external to the hospital for teaching and learning, or for patient consultations.
- **Biomedical Equipment**: Adequate provisions are being made in the hospital to accommodate outputs from newer biomedical equipment being fed directly into the EPAS.
- **Wireless**: The hospital will be saturated with wireless capabilities to take advantage of the new and emerging devices that support wireless in a hospital setting. These include a wide range of biomedical devices and imaging equipment. Patient data from these devices, whether wireless enabled or hardwired, will ultimately be uploaded to the electronic patient record.
- **Imaging**: in the new RAH, the concept of medical imaging has been extended beyond traditional radiology to also include any digital photography, such as that taken during plastic surgery procedures, or moving pictures such as videos capturing a patient's gait before and after therapy. These are for all intents and purposes part of the patient's record from a medico-legal perspective, as well as providing important information in subsequent patient visits for continuing treatment. Consequently, over time non-radiology images will be linked with the electronic record to provide the most complete visual as well as written record of patient treatment in South Australia.
- **Pharmacy**: the new RAH will in all likelihood be the first hospital or one of the first hospitals in SA to implement what are commonly known as Pharmacy Robots, which have become one of the preferred methods of providing unit doses (as described by Deloitte Touche Tohmatsu 2010). The process of prescribing medication for patients, including an automated check for contraindications and over/under dosing, will be handled by an EPAS front end, and then handed off to the backend dispensing systems, including the robots. Other medication will be provided from ward based automated dispensing cabinets. The administration of medication by clinical staff to the patient will recorded in the electronic patient record.
- **Pathology**: Patient samples will be labelled and tracked through the hospital to Pathology for testing, with the results being automatically written into the patient record from the Laboratory Information System.
- **Wayfinding/Public Information Systems**: the Wayfinding and Public Information systems will overlap in functionality in the direction of people to events, such as lectures, seminars, exams etc, in the hospital as well as providing general health information in targeted areas and alerts to patients, staff and visitors about evacuations. Wayfinding within the hospital for staff and visitors on mobile devices such as smartphones is under consideration.
- **Personal Devices**: There is an expectation that there will be a significant increase in the use of personal computing devices, such as iPads, tablets and the like, to source information about or in patient interactions. As such, EPAS will form the basis for clinical staff to create their own views of patients assigned to their care, to receive alerts, and to develop and execute patient care plans.
- **Large Patient Information Displays**: The old whiteboard in the nurses' station will be a thing of the past. Rather, large displays showing the patients in each pod (collection of n beds, where n can be 4, 8, or 16 beds) and the status of various tests, imaging etc can be seen through the area, and replicated on their personal devices if necessary will ensure that clinical care is accurate and up to date at all times.

5 Challenges in a Private Public Partnership (PPP) Environment

The hospital is being built and will be operated under a PPP arrangement. In terms of the new RAH this means that the design, development and construction of the hospital, which includes significant elements of the ICT is the responsibility of the private partner to deliver on the basis that it meets the functionality specified in the Project Agreement.

Once the hospital is complete, the private partner's operating organisation will take responsibility for the non-clinical aspects of the hospital, such as food services, cleaning, porters, logistics, physical security and building and engineering maintenance.

In terms of ICT, the State will take responsibility for the running of the network which the private partner has designed, procured and commissioned. The network will be virtually separated and the private partner will share the same physical network to support all of its systems.

Accordingly, this has resulted in some very detailed work to accommodate a private organisation running its ICT over the same physical network as the health information.

In the main most of the information that will cross from the State’s virtual network to the private partner’s operating organisation will comprise standard HL7 messages to ensure that services that they provide are delivered in a timely manner. The time stamps on these messages will also be used to determine whether the private partner’s operating organisation is meeting its agreed service levels.

6 Summary

The new RAH will be the beginning of new and innovative ways to improve the delivery of health care in South Australia.

It will bring innovations across the whole range of elements that make for an efficient and new age tertiary hospital in design, services and patient flow.

The ICT element of the hospital is critical in ensuring that the return on the investment in such a new and complex facility will be achieved.

7 References

Deloitte Touche Tohmatsu (2010): Robotic dispensing Automation in pharmacy. http://www.deloitte.com/assets/DcomAustralia/Local%20Assets/Documents/Industries/LSHC/1011_Automation%20in%20Pharmacy_National%203.PDF. Accessed 30 Oct 2012.

Government of South Australia (2003): Better Choices Better Health. http://www.publications.health.sa.gov.au/cgi/viewcontent.cgi?article=1003&context=spp&sei-redir=1&referer=http%3A%2F%2Fwww.google.com.au%2Furl%3Fsa%3Dt%26rct%3Dj%26q%3Dbetter%2520choices%2520better%2520health%253A%2520final%2520report%2520of%2520the%2520south%2520australian%2520generational%2520health%2520review%26source%3Dweb%26cd%3D1%26cad%3Drja%26ved%3D0CCwQFjAA%26url%3Dhttp%253A%252F%252Fwww.publications.health.sa.gov.au%252Fcgi%252Fviewcontent.cgi%253Farticle%253D1003%2526context%253Dspp%26ei%3D34-nUK77DMasrAeZ5oCIBQ%26usg%3DAFQjCNE7T1v3BDPixYR0iGvm74MSPD_NCw#search=%22better%20choices%20better%20health%3A%20final%20report%20south%20australian%20generational%20health%20review%22. Accessed 30 Oct 2012.

Government of South Australia (2007): South Australia's health care plan 2007-2016 : the South Australian Government's plan for health care over the next 10 years. http://www.publications.health.sa.gov.au/cgi/viewcontent.cgi?article=1002&context=spp&sei-redir=1&referer=http%3A%2F%2Fwww.google.com.au%2Fsearch%3Fq%3DSouth%2BAustralia%25E2%2580%2599s%2BHealth%2Bcare%2BPlan%2B2007-2016%252C%2BGovernment%2Bof%2BSouth%2BAustralia%252C%2BUndated%26oq%3DSouth%2BAustralia%25E2%2580%2599s%2BHealth%2Bcare%2BPlan%2B2007-2016%252C%2BGovernment%2Bof%2BSouth%2BAustralia%252C%2BUndated%26sugexp%3Dchrome%2Cmod%3D4%26sourceid%3Dchrome%26ie%3DUTF-8#search=%22South%20Australia%E2%80%99s%20Health%20care%20Plan%202007-2016%2C%20Government%20South%20Australia%2C%20Undated%22. Accessed 30 Oct 2012.

Next Generation Linkage Management System

Julie Harris
Manager, Data Linkage, SA NT Datalink
University of South Australia
UniSA House Level 3, 195 North Terrace, City East Campus, Adelaide 5000, South Australia

julie.harris@unisa.edu.au

Abstract

SA NT Datalink is a consortium of government departments, universities and other parties that are committed to providing high quality data linkage to support research. The backroom technologies that provide linked data to researchers will be discussed in detail in this paper. Data linkage is commonly known as a process utilising computer data matching technology to compare similar records from within and across multiple datasets.

The Next Generation Linkage Management System has been developed using open source technologies to manage disparate source data files coming in to the organisation, cleansing and standardisation, then the analysis of the data which will determine blocking parameters and linkage weights. The open source linkage engine called FEBRL (Freely Extensible Biomedical Record Linkage) is used to link the datasets using probabilistic methods. For storage of the linked records SA NT Datalink has employed a graph database which allows us to keep and reuse the rich comparison vectors. The data structures within a graph database are more aligned with the native formats of linked data. The graph database also provides a repository that is very fast for the retrieval of data, as unlike relational database there are no indexes or joins which are computationally expensive. The benefits of using both deterministic and probabilistic linkages will be discussed, and the analysis that is required on a dataset to assist in selecting the best linkage strategy. Graph databases are based on graph theory, and are used by some of the largest organisations on the web to deliver a very fast service to their customers. Some quality tools have been implemented by SA NT Datalink to ensure a reduction in the number of false positives and false negatives. Some mention will be given to what the Next Generation Linkage Management System does not provide will be touched upon. SA NT Datalink have developed a loosely coupled, open source system of managing, linking and extracting the linked data which will form the corner stone of their offerings to researchers for the coming decade.

Keywords: Data linkage, data matching, graph database, FEBRL, probabilistic linkage, deterministic linkage.

1 Introduction

SA NT Datalink is a consortium of government departments, universities and other parties that are committed to providing high quality data linkage to support research.

There is increasing recognition that administrative data, collected and held within public and private organisations, is a valuable resource for population research that underpins important program evaluation and policy making. SA NT Datalink was established to create linkages between data relating to individuals across multiple datasets and captured across many sectors, including publically funded health care, education and social services. Once linked, data describing the health and experiences of many thousands of individuals can be supplied to a researcher in a completely de-identified format. In effect, this intelligent linkage process strengthens privacy protection while giving researchers access to true population-based data, maximising the value derived from this often dormant resource.

SA NT Datalink was launched in November 2009 and has been progressing towards providing a true representation of the population – far beyond the usual sample population studies. The South Australian linked population recently exceeded 1.6 million people, and although this includes deceased individuals, it demonstrates that SA NT Datalink is nearing full population capture, as SA’s current estimated resident population is proximately 1.65 million.

SA NT Datalink have designed and built a holistic management system to analyse, store and extract linked data for researchers. At the heart of this system is a graph database which provides the ability to store data in a format true to its natural state.

2 What is data linkage?

At its simplest, data linkage aims to identify the same entities (people, events, object) across different databases. Each organisation has at least one and often many databases where information on people is stored and used for their own purposes. Because of this, unique identifiers for individuals are not shared across organisational borders. SA NT Datalink use data linkage to probabilistically link individuals across many different datasets from state and commonwealth government departments and other bodies, then provide the de-identified data back to researchers.

The linkage process is computationally complex because if we were looking for the same individual in two different datasets, potentially we would have to take the first record in the first dataset and compare it with every record in the second dataset to find a pair. Strategies and techniques have been developed to reduce this complexity.

Blocking or indexing techniques are used to reduce the number of record pairs to be compared by removing pairs that are unlikely to match. A common blocking technique is to alphabetically sort the surname of the records into blocks and compare like with like. The selection of the blocking key is an outcome of the raw data analysis and will change according to its characteristics. For example in some datasets the postcode may be a reliable and well populated field so would make a strong candidate for the blocking key.

3 What it isn't

Data linkage is not the same as data warehousing or data mining. The size of the stored data often pushes it into the category of Big Data. Staff at SA NT Datalink do not do any research, we limit ourselves to analysis of the data for hr purposes of data linkage.

We are not building a large database of individuals and their service data. Due to our adherence of the Separation Principle, SA NT Datalink only ever get to handle the demographic fields in a dataset which are required for linkage. Inside our stand-alone secure facility we effectively have an electronic copy of the white pages for SA and NT.

We do not have a researchable dataset, as stated no service data is kept inside the Master Linkage File but always resides and is under the control of the data custodians.

4 The high level process

4.1 Receive the data

Data custodians such as SA Health, the Department of Education and Children's Development, pathology organisation, etc provide SA NT Datalink with datasets from their own collections. It is in most cases provided as a raw data extraction from their databases of the unique identifier and demographic data and in most cases it is delivered safe hand to us. The data is loaded onto our secure stand-alone server and the original media is destroyed.

4.2 Analysis of the raw data

A critical success factor in successful data linkage is in understanding the raw datasets. We seek to understand how the data is collected and any idiosyncrasies of the collection. We look at the frequency of variables in the raw data and how well populated fields are. Meta data is provided by the data custodians to assist in our understanding of the data. This analysis also feeds into the guidelines for clerical review.

4.3 Selection of a linkage strategy

As touched on before the blocking key is selected and linkage weights are chosen. Some analysis is completed on possible results arising from using the strategy. Usually with a new dataset, several of these 'test' linkages would occur before a satisfactory linkage strategy was decided upon.

4.4 Pre-processing (ETL)

The raw data is mapped to database variables in our Master Linkage File. Unwanted characters are removed, and standardisation of some fields occurs, i.e. street and st would be changed to Street. Addresses would be segmented from one long field into address line 1, address line 2, suburb and postcode. The data is then loaded into a staging area ready for linkage.

4.5 Record pair comparison

This is where the datasets are processed by the linkage engine. At SA NT Datalink we use FEBRL, which is an open source product developed by staff at Australian National University. There are many linkages engines on the market and they all do basically the same thing. They compare each candidate record based on several attributes i.e. surname, first name, date of birth, suburb etc and generate a vector of numerical similarities – the 'comparison vector'. At SA NT Datalink we are able to store the comparison vector in its native format in the graph database. In previous data linkage models the comparison vector had to be summed into a single value so it could be stored in a relational database. By only storing the summation value, a severe loss of information occurs. Effectively the linkage engine is doing some complex calculations and similarity scores which are being simplified into one number when stored in the relational database. SA NT Datalink's innovative use of a graph database stores and re-uses the advanced classification calculations that occur inside FEBRL.

4.6 Classification

We use threshold based classification to decide of the upper and lower limits of the linked data. Above a set threshold the records are considered true positives and below a threshold they are true negatives.

SA NT Datalink uses a stratified sampling approach to determine where the threshold should be placed to maximize the number of True Positives and minimize the number of False Positives.

Each record pair can be classified as a True Positive, False Positive, True Negative or False Negative. The ideal is to find a high number of true positives with a low false positive and false negative score. Analysis is conducted of the record pairs to determine the precision and recall of the linkage. We document this analysis in our linkage specification using a precision-recall graph, F-measure graph and a ROC curve.

4.7 Evaluation

For the records which fall in between the upper and lower thresholds we conduct clerical review to try to classify as many records as possible. Only the clusters of records that are most difficult are sent to clerical review. Each cluster is manually inspected by trained clerical reviews and classified according to guidelines. A cluster contains a variable number of records (or nodes) that refer to the same entity. See Figure 3 - linked records in a graph database. A clerical review officer can either 'break' a link or 'force' a link. Inside the graph database they are represented differently to the weighted links created by FEBRL. The clerical review links are unweighted.

This increases the quality of the data in the Master Linkage File that is provided to researchers. The downside of clerical review is that it is time consuming and unless quality evaluation is conducted and guidelines are provided, variable quality could result.

4.8 Extraction of project specific linkage keys

This process is the raison d'être of SA NT Datalink. It is where we provide the de-identified data for researchers.

At this stage in the process, we attach a randomly generated project specific linkage key to each cluster of records; a cluster being a number of records (or nodes) that refer to the same entity i.e. one probable person. We then remove all demographic variables just leaving the data custodian's own unique identifier (i.e. the customer number). When the data file is proved back to the data custodian they are able to extract the service data from their database using their unique identifier (i.e. enrolment number), remove the demographic data and attach the project specific linkage key before sending this data to the researcher.

Because the project specific linkage key is the same across all datasets in the researcher's cohort, they can identify the same individual across all datasets.

5 What happens inside the linkage engine

5.1 Deterministic Linkage

Deterministic or rules-based record linkage is the simplest type of linkage. Two records are compared on one or more fields in both records match exactly. For example, deterministic linkage will create a record pair if the surname, date of birth, and Medicare number are identical. This simpler type of linkage works well when there are similar entities in the datasets i.e. hospital separation data and emergency room data. At SA NT Datalink we use this method when appropriate, usually as a forerunner to probabilistic linkage.

5.2 Probabilistic Linkage

Often called fuzzy matching, probabilistic linkage looks at a wider range of fields and calculates weights for each field similarity. This is the 'comparison vector'. Following on from the last example, each surname field would be compared and a weight calculated. So for the surnames "Smith" and "Smyth", a relatively high score would be given using probabilistic methods, while it would not match at all using deterministic methods. The next field containing the Medicare number would be compared and a high score could still be calculated even if one number was transposed in the field. The linkage weights selected during the analysis phase will determine which fields provide a greater discrimination for record pairs.

SA NT Datalink use probabilistic linkage as the cornerstone of their linkage system.

It is interesting to note that in other jurisdictions such as the United Kingdom, deterministic linkage is the main linkage methodology as all individuals are identified by their NHS number.

6 Issues with Linked Data

False positives and false negatives will inevitably be included in the researcher's dataset. In many cases twins and triplets are brought together by the linkage engine and can be difficult to detect. We have one category of twins we have called 'super twins', where the date of birth, surname, gender, address and even birth weight are exactly the same.

The Black Box syndrome has long been an issue. Our researchers are increasingly demanding that we release information on the analysis of raw data, the reasoning behind the selection of a linkage strategy and weights, the analysis and results of test linkages and the outcomes of clerical review. They want to know what is going on inside the linkage engine and what changes are being made by the clerical review officers. We are responding to this need by documenting all stages of the linkage process and providing it to researchers.

Low quality of the underlying datasets can result in poor linkage quality. It is an unfortunate fact of life that data is not always collected in ideal conditions using modern field validation techniques. We find that some datasets have no surname or first name, or a low reliability ATSI indicator. Analysis conducted on the raw dataset and intelligence on the meta data provided by the data custodian informs the linkage strategy used to optimise results using low quality datasets.

7 Privacy and confidentiality

Privacy and confidentiality are a major concern for both the data custodian and the linkage unit. SA NT Datalink's model for data linkage is superior to ad-hoc data linking performed by the researchers themselves. Self-linking by researcher requires full disclosure of identifying data. Our model and specialised technology improves linkage quality, while saving the researcher time, allowing them to focus on analysis, and publication of results.

The process allows for total anonymity of study populations, removing this knowledge burden from researchers, respecting the privacy of the population, and providing additional assurance to ethics committees. A follow on effect of this increased information and privacy protection had been greater access to data, which has hitherto been tightly guarded.

Rigorous security processes, a layered security model and adherence to the 'Separation Principle' ensure that only those staff with a need to know work with the identified data.

SA Health staff are embedded within SA Datalink and are the only ones allowed to handle the demographic data (names, addresses, etc) that is to be used for linkage purposes. Researchers only have access to the de-indentified service data that has been approved by an ethics committee.

8 The underpinning architecture

SA NT Datalink steering committee made a strong directive to use open source technologies wherever

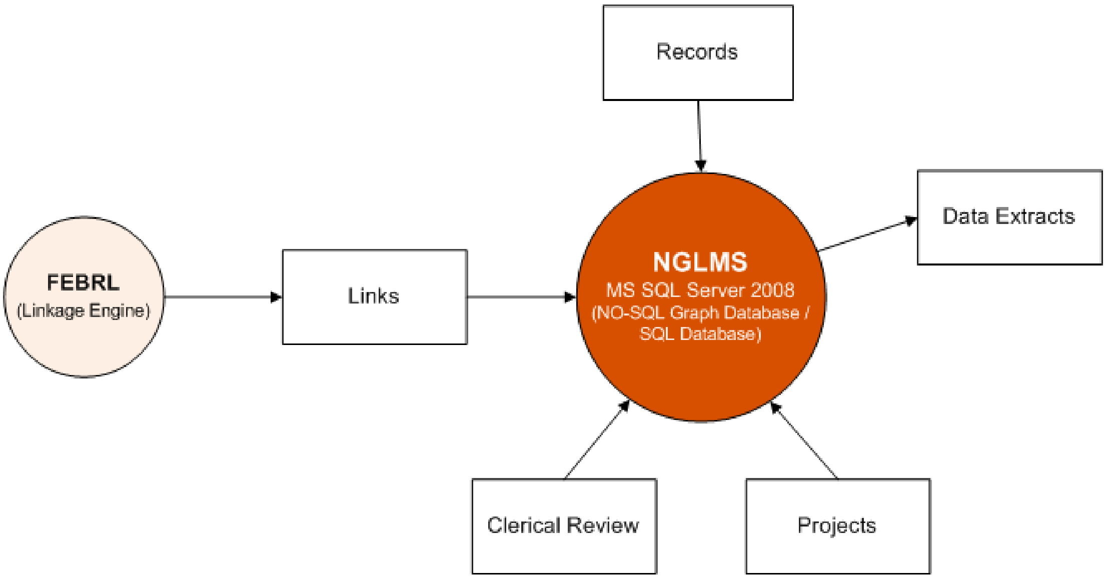


Figure 1: Next Generation Linkage Management System functional diagram

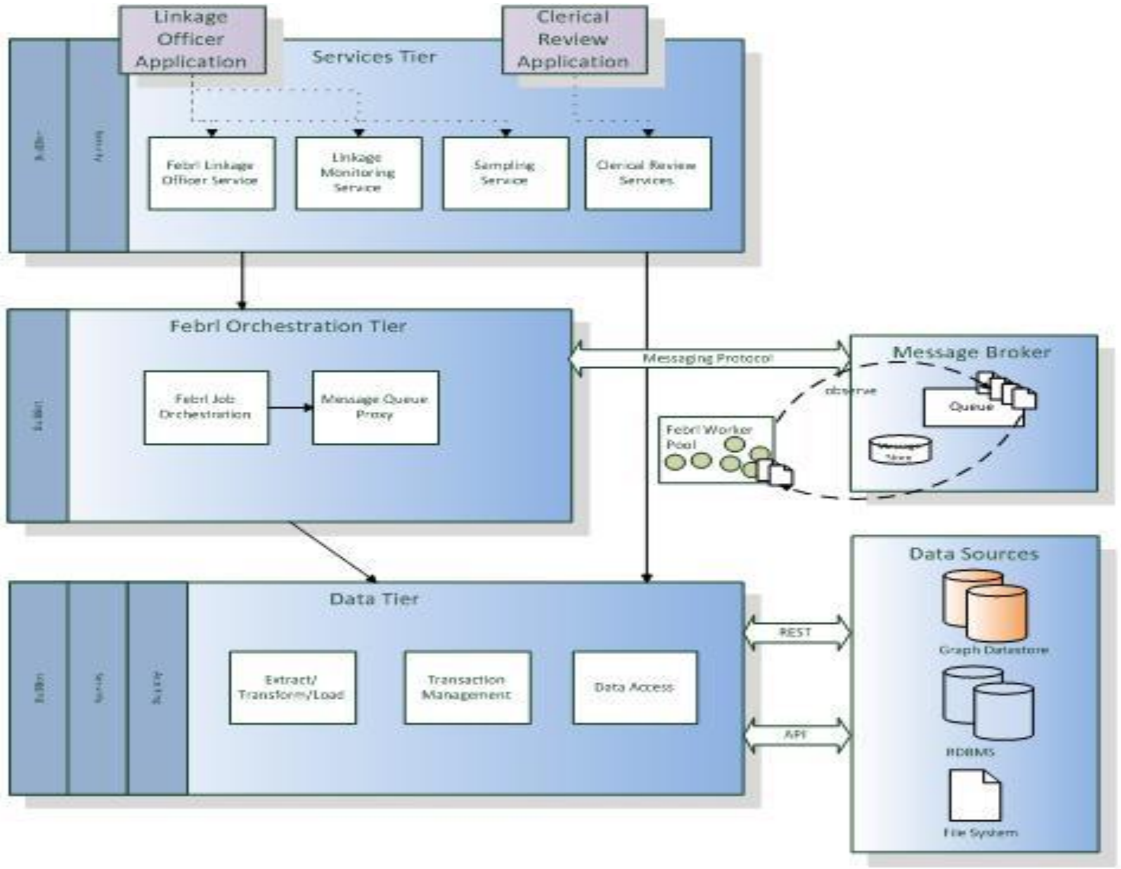


Figure 2: Next Generation Linkage Management System - Technical Architecture

possible. This will allow the Next Generation Linkage Management System to be shared within the Public Health Research Network to other linkage nodes. We already have close links to other state jurisdictions that may choose to take advantage of the intellectual property that has been developed by SA NT Datalink.

To this end we selected FEBRL as the linkage engine which is at the core of the Next Generation Linkage Management system. We also use Jython code, a Postgresql database, a Neo4jgraph database, html screens and csv files

9 The graph database – a new approach to storing linked data

One of the design decisions that has been made for the Next Generation Linkage Management System is the use of a graph database for the storage the Master Linkage File. Previously we (and every other Australian data

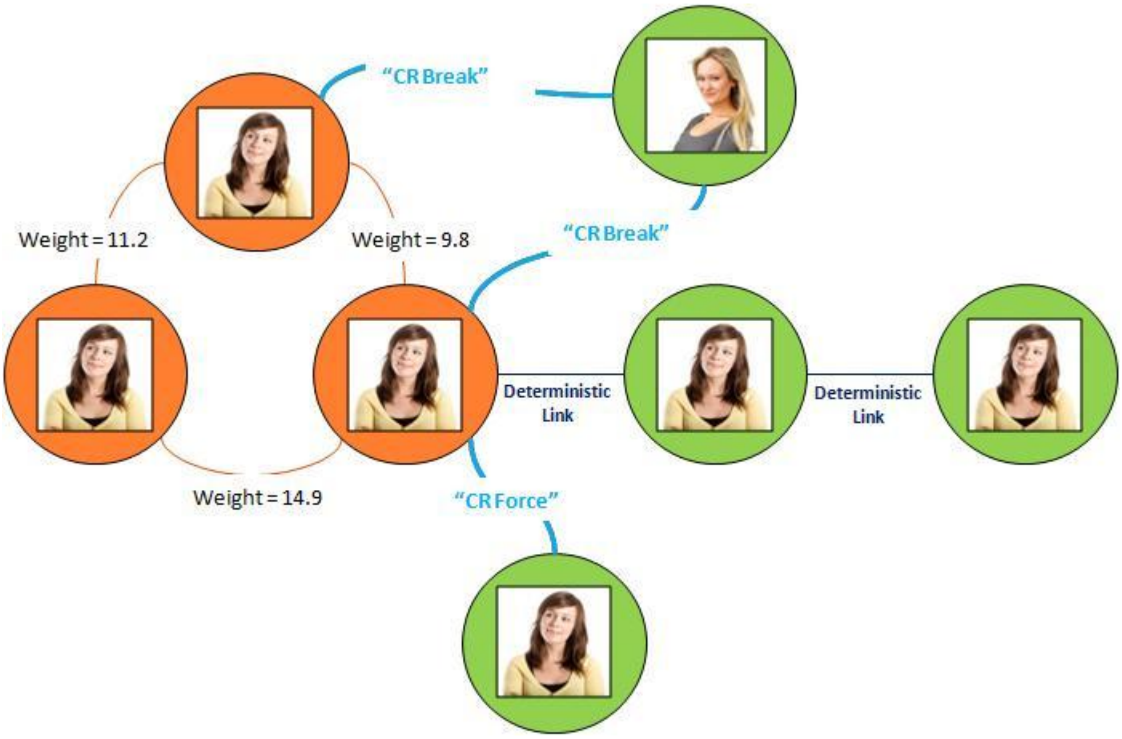


Figure 3: Linked records in a graph database

linkage node) used the more traditional relational database model which has been in use for over 20 years. We have chosen to use Neo4j which is a market leader in this space.

Graph databases are based on graph theory which uses mathematical structures to model pair wise relations between objects from a certain collection. So the use of this model lends itself to data linkage. It allows us to store the comparison vector in its native state. Figure 3 - linked records in a graph database has been simplified to show just one score for similarity between records (or nodes). The thresholds set for a research study's linked data will determine where the clusters are drawn. This allows for extractions of clusters at the required level of specificity and sensitivity. Specificity is the True Positive rate and sensitivity is the True Negative rate. We have found that traversals of the graph database occur very fast as there is no need for computationally costly indexes and joins.

Graph databases have been around for the last 5 years and are being used in 24/7 business. One of the heaviest users of graph database is the social networking market. Twitter uses its FlockDB graph database to store the social graph that lets the site determine who's connected to whom, and how.

Amazon's recommendation engine is housed in a graph database, with each product being represented as a node and the similarity score represented in the link or edge.

Neo4j is the leader in terms of usage in the graph database market. In line with SA NT Datalink's philosophy it is an open source product.

The reason graph databases have not been used before in data linkage nodes is that within the Australian network all the nodes are fairly well established and have designed their system 10 or more years ago. SA NT Datalink has the opportunity to take advantages of the developments in this field.

10 Master linkage file

SA NT Datalink has invested heavily in building the Master Linkage File. This ensures that the datasets that are collected, linked and clerically reviewed and made available (with the proper approvals) to many research studies.

Over our four years of existence we have collected data from -

- SA Cancer Registry
- SA Public Hospital Emergency Department Presentations
- SA Health Public Hospital Inpatient Data (ISAAC)
- SA Public School Enrolments Census
- SA Public School Student, Years 1 to 3 Reading Assessments
- SA Public School Students English as a Second Language Scale
- SA Women's & Children's Health Universal Neonatal Hearing Screening Program
- Families SA Child Protection
- Families SA Care & Protection Orders
- Housing SA Public Housing Program
- Housing SA Aboriginal Rental Housing Program
- SA Perinatal Outcome Unit
- SA Births Registry
- SA Deaths Registry
- SA Mental Health and Substance Abuse
- SA Drugs of Dependence Registry
- SA Disability Services
- SA Private Pathology Services
- NT Health Department Client Master Index database (Health)
- NT Perinatal Outcomes Registry
- NT Immunisation Registry

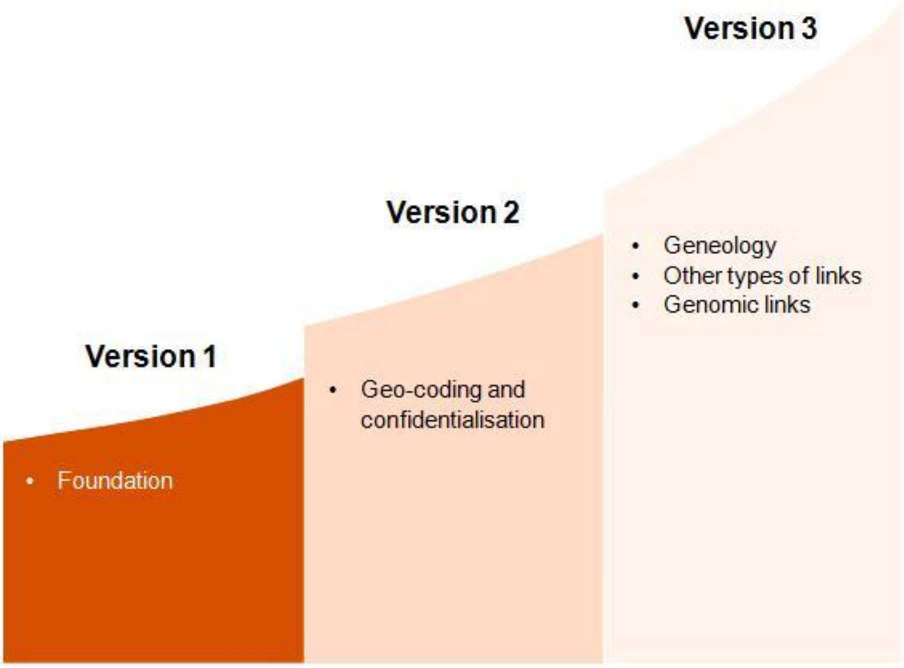


Figure 4: The future releases of the NGLMS

- NT Deaths Registry
- NT Births Registry
- NT Department of Education and Training National Assessment Program Literacy and Numeracy NAPLAN
- NT Department of Education and Training Enrolment School enrolments
- Australian Early Development Index (AEDI) – SA and NT

Currently we have over 4 million records within the Master Linkage File and we expect significant growth over the next 10 years. Our technology decision to use a graph database in the Next Generation Linkage Management System means that we can accommodate this growth and still have a system responsive to queries.

A key benefit of our system is the ability to link in a dataset, extract the project specific linkage keys and then completely remove the dataset. We believe this ability will provide us with opportunities to have access to datasets which are deemed highly sensitive such as offenders' data, IVF etc.

11 Project linkage files

SA NT Datalink is also geared up to do linkage on a project basis. We can take data files, analyse, link and extract keys for a specific purpose then return the linked dataset to the custodian and remove it from our system.

12 What are the benefits to researchers

The Next Generation Linkage System was designed to 'open the black box' for researchers. SA NT Datalink is committed to providing information to researchers on how their data was linked, what the raw data looked like, the analysis that occurred and the impact of clerical review. We can provide custom extractions with different parameters with different characteristics. Sensitivity and specificity of data can be manipulated to suit the research study.

13 The Future for the Next Generation Linkage Management System

We are only in stage one of this ambitious project, and have built and are using a basic system. As funds allow we plan to move into the areas of geocoding data, then onto building the capability to represent genomic links and 'community' links (i.e. people who live in the same public house).

14 References

Christen, P. (2012): *Data Matching*- Data-Centric Systems and Applications. Berlin Heidelberg , Springer-Verlag.

Contributed Papers

Validating Synthetic Health Datasets for Longitudinal Clustering

SHIMA GHASSEM POUR[1], **ANTHONY MAEDER**[1]
and LOUISA JORM[2]

[1] School of Computing, Engineering and Mathematics
University of Western Sydney
Campbelltown, Australia
Email: `A.maeder@uws.edu.au`

[2]School of Medicine
University of Western Sydney
Campbelltown, Australia
Email: `L.jorm@uws.edu.au`

Abstract

Clustering methods partition datasets into subgroups with some homogeneous properties, with information about the number and particular characteristics of each subgroup unknown a priori. The problem of predicting the number of clusters and quality of each cluster might be overcome by using cluster validation methods. This paper presents such an approach incorporating quantitative methods for comparison between original and synthetic versions of longitudinal health datasets. The use of the methods is demonstrated by using two different clustering algorithms, K-means and Latent Class Analysis, to perform clustering on synthetic data derived from the 45 and Up Study baseline data, from NSW in Australia.

Keywords : Cluster analysis; longitudinal synthetic data; Cluster validation

1 Introduction

Unsupervised learning methods (such as clustering) are based on discovering statistically reliable, unknown previously, and actionable insights from datasets, with information about structure of the datasets (such as cluster number and size) unknown a priori. On the other hand, some clustering algorithms seek to determine the number of clusters in advance. In data that are not clearly separated into groups, identifying the number of clusters becomes difficult. Various validity indexes are available to measure the quality of each cluster, such as Silhouette index (Rousseeuw 1987), Dunn's index (Dunn 1973) and Davies-Bouldin index (Davies & Bouldin 1979). In addition, the BIC index (Schwarz 1978) has been used because it is closely associated with the Latent Class Analysis method.

The organization of the paper is as follows: in Section2 we describe two different clustering methods; in Section 3 we introduce relevant validation methods and Section 4 presents construction of a synthetic dataset. Comparison of the validation methods is presented in section5 and a conclusion in section 6.

2 Two different clustering methods

Cluster analysis refers to partitioning the data into meaningful subgroups, when the information about their composition and the number of subgroups are unknown (Jain et al. 1999). In this paper we use two different kinds of clustering algorithms to cluster our datasets, in order to also investigate the effect of the algorithm choice.

The K-means algorithm is a point-based clustering method which places cluster centers in an arbitrary position and relocates them at each step to optimize the clustering error. Despite being widely used in many clustering applications, this method suffers from sensitivity to initial position of the cluster centers (Likas et al. 2003).

Latent Class Analysis (LCA) is a statistical clustering approach that attempts data reduction by classifying objects into one of K homogeneous clusters, where within-group-objects similarity is minimized and the between-group-objects dissimilarity is maximized, and where K is fixed and known. LCA applies a probabilistic clustering approach: this means that although each object is assigned to belong to one cluster, it is taken into account that there is uncertainty about an object's class membership (Magidson & Vermunt 2002, Lanza et al. 2003).

3 Validation methods

There are several validation methods available to validate the quality of clusters resulting from a given clustering method. One approach consists of running a clustering algorithm several times for different numbers of clusters and computing validity indexes to assess the quality of each cluster. Validation indexes can be divided into two categories: external index and internal indexes. External index techniques use a dataset with known cluster configurations and measure how well clustering methods perform with respect to these known clusters. Internal indexes techniques are used to evaluate the goodness of a cluster configuration without any prior knowledge of the nature of the clusters (Rendón et al. 2011). In practice, external information such as class labels is often not available in many application scenarios. Therefore, in the situation where there is no external information available, internal validation indexes are the only option for cluster validation. This section presents four widely used and well-known internal validation indexes: Silhouette index (Rousseeuw 1987), Davies-Bouldin index, Dunn's index (Dunn 1974) and BIC index (Schwarz 1978); used to assess the ideal number of clusters and the quality of clusters (Liu et al. 2010). Useful reviews of available validation techniques have been presented elsewhere (Halkidi et al. 2002, Datta & Datta 2003, Kryszczuk & Hurley 2010).

3.1 Silhouette index

For a given cluster, $X_j(j = 1..c)$, a quality measure assigned to the ith sample of X_j which known as silhouette width. This value is a confidence indicator on the membership of ith sample in the cluster X_j and defined as:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{1}$$

where $a(i)$ is the average distance between the ithsample and all of samples belong to X_j, $b(i)$ is the minimum distance between the ithsample and all of samples clustered in $X_k(k = 1...c; k \neq j)$. Thus for a given cluster $X_j(j = 1..c)$, it is possible to calculate a cluster silhouette S_j, which characterizes the heterogeneity and isolation properties of such a cluster:

$$S_j = \frac{1}{m}\sum_{i=1}^{m} s(i) \tag{2}$$

where m is number of samples in X_j. It has been shown that for any partition $U \longleftrightarrow X : X_1 \bigcup ...X_i \bigcup ...X_c$, Global Silhouette value can be used as an effective validity index for U.

$$GS_u = \frac{1}{c}\sum_{j=1}^{c} S_j \tag{3}$$

if c is the number of clusters for partition U: a maximum value of GS_u indicates the better cluster configuration for a given dataset.

3.2 Dunn's index

The idea of this index (Dunn 1974) is based on clustering compactness and good separation:

$$D(U) = \min_{1 \leq i \leq c} \left\{ \min_{1 \leq j \leq c_{j \neq i}} \left\{ \frac{\delta(X_i, X_j)}{\max \Delta(X_k)} \right\} \right\} \tag{4}$$

if $U = X_i \bigcup ...X_j \bigcup ...X_c$, the $\delta(X_i, X_j)$ is the inter-cluster distance between clusters i and j and $\Delta(X_k)$ is the intra-cluster distance for cluster K. The main goal of this measure is to minimize intra-cluster distance and maximize the inter-cluster distance.

3.3 Davies-Bouldin index

To express how far clusters are located from each other and how compact they are, the Davies-Bouldin index can be used. The Davies-Bouldin index can be defined as:

$$DB(U) = \frac{1}{c}\sum_{i=1}^{c} \max_{i \neq j} \left\{ \frac{\Delta(X_i) + \Delta(X_j)}{\delta(X_i, X_j)} \right\} \tag{5}$$

if $U = X_i \bigcup ...X_j \bigcup ...X_c$, $\Delta(X_i)$ and $\Delta(X_j)$ represent the intra-cluster distance and $\delta(X_i, X_j)$ define as the inter-cluster distance. Therefore, the number of clusters that minimizes DB is chosen as the optimal number of clusters.

3.4 BIC index

This index is presented to avoid over-fitting in a dataset and is defined as:

$$BIC = -ln(L) + vln(n) \tag{6}$$

where n is the number of objects, L is the likelihood of the parameters to generate the data in the model and v is the number of free parameters in the Gaussian model. The BIC index takes into account both the fit of the model to the data and the complexity of the model: a model that has a smaller BIC is better.

There are different methods available to calculate the intra-cluster and inter-cluster distances based on functions defined on the set of all sample pairs (Azuaje 2002). Here we present two examples for each of these distances to show their diversity. Various well known metrics are used to calculate the distance between two samples, $d(x, y)$, such as Euclidean and Manhattan metrics (Salzberg 1991). In this paper we use the Euclidean metric as it is computationally simple.

3.4.1 Inter-cluster Distance

Single linkage is defined as the closest distance between two samples which belong to different clusters.

$$\delta(S, T) = \min d(x, y)_{x \in S, y \in T} \tag{7}$$

Complete linkage is represented by the distance between two remote samples in different clusters.

$$\delta(S, T) = \max d(x, y)_{x \in S, y \in T} \tag{8}$$

If S and T represent clusters from partition U and $d(x, y)$ defines the pair-wise distance between samples in S, T.

3.4.2 Intra-cluster distance

Complete diameter is defined as maximum distance between two samples belonging to the same cluster:

$$\Delta(S) = \max d(x, y)_{x, y \in S} \tag{9}$$

Average diameter is defined as the average distance between all of the samples in same cluster.

$$\Delta(S) = \frac{1}{|S|.(|S| - 1)} \sum_{x, y \in S, x \neq y} d(x, y) \tag{10}$$

Where $|S|$ represents the number of samples in the cluster S and $d(x, y)$ is the distance between two samples $x, y \in S$.

In this paper we use complete linkage for the intra-cluster distance and complete diameter for inter-cluster distance to calculate the Dunn's index, Davies-Bouldin index and Silhouette index.

We used the Cluster Validity Analysis Platform (CVAP) (Wang et al. 2009) to run K means and compute Davies-Bouldin, Dunn's index and Silhouette index. Latent Gold software(*Welcome to Statistical Innovations Inc.* 2011) was used to run Latent Class Analysis and compute BIC. Also, we used Matlab to compute Dunn's index and Silhouette index for LCA .

4 Constructing a synthetic dataset

In our data mining research, we are using the 45 and Up Study baseline dataset (*Study Overview* 2011). The 45 and Up Study is a large-scale cohort involving 266, 848 men and women aged 45 years and over from New South Wales (NSW), Australia. Participants in the 45 and Up Study were randomly sampled from the database of Australia′s universal health insurance provider, Medicare Australia, which provides virtually complete coverage of the general population. Participants joined the Study by completing a baseline questionnaire (between February 2006 and April 2009) and giving signed consent for follow-up and linkage of their information to a range of health databases. The baseline questionnaire (available at http://www.45andup.org.au) collected measures of general health, health related behaviors and demographic and social characteristics.The overall response rate was 18%. The Study is described in detail elsewhere (Banks et al. 2008). In addition, it is planned to follow up the cohort every five years (Banks et al. 2009). Thus this study is of interest for many researchers to evaluate and develop longitudinal data mining methods. However, this study has finished only its first stage of collecting data and is currently entering the second phase to provide the first time step after baseline. It is necessary to have longitudinal datasets to test and evaluate longitudinal clustering methods; therefore as part of our project we are interested to create a synthetic longitudinal dataset based on this study. This section aims to explain our procedure for creating of a synthetic dataset. For the sake of simplicity we chose two variables, Body Mass Index (BMI) and the amount of Physical Activity (PA) for each case. Body Mass Index (BMI) was calculated from weight and height as self-reported on the baseline survey. After excluding people with a reported BMI of <15 or >50 kg/m2 or unknown BMI, BMI was categorized using the following cut-points: 15 (underweight), 18.5, 20 and 22.5 (normal weight), 25 and 27.5 (overweight), 30 (obese). Participants overall level of physical activity was classified according to their responses to elements of the Active Australia Questionnaire (of Health & Welfare AIHW), comprising information on number of weekly sessions (of any duration) of moderate and vigorous physical activity and episodes of walking for longer than 10 min. A weighted weekly average for number of sessions was calculated for each participant by adding the total number of sessions, with vigorous activity sessions receiving twice the weighting of moderate activity or walking sessions, and was categorized as $0 - 3, 4 - 9, 10 - 17 and 18$or more sessions per week. From about 160, 000 cases we randomly chose about 1, 000 cases from the first stage of the 45 and Up Study dataset, to represent the first time step of data. The LCA method was used to cluster our baseline data and based on the established BIC (Schwarz 1978) we determined the number of clusters which minimized the BIC index. The 45 and Up Study has primary ethical approval from the University of New South Wales Human Research Ethics Committee (HREC 05035). The main goal in creating a longitudinal synthetic dataset is to explain cluster behaviour in the next time step, as exhibited by either merging (clusters vanishing) or splitting (creating new clusters). Therefore we seek a parameter point in our sequence of different synthetic datasets at which the number of clusters change in situation, as a means to explain where the characteristics of the dataset changes. What proportion of data and by how much our data should change, is a fundamental question in this step. To address this question we investigated two different scenarios.

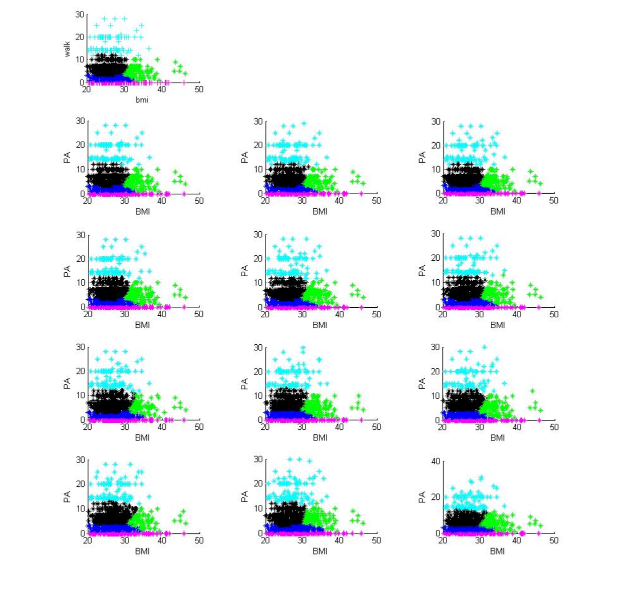


Figure 1: sequence of changing elements of larger cluster (BMI and PA increased by random number up to variance)

Scenario 1: we decided to follow a systematic change pattern for successively every five percent of the larger cluster (in terms of number of elements), for each element, adding a normally distributed random number in the range(-variance, +variance) of the targeted cluster. At each step we ran LCA to cluster the new time step data and we determined how many elements moved from one cluster to another cluster. The results in Figure 1 show that by applying this amount of change we observed cluster boundary positions changing somewhat, as might be expected.

Senario2 : we decided to change successively the element values of every five percent of the larger cluster (in terms of number of elements), for each element adding a normally distributed random number in the range(-2*variance, +2*variance) of the targeted cluster. At each step we ran LCA to cluster the new time step data and we determined how many elements actually moved from one cluster to another cluster. With this amount of change we might expect the targeted cluster to split as well as experiencing element movements. The results in Figure 2 show that after changing 40% of elements, our targeted cluster was splitting.

Based on this approach, three different datasets were chosen, one being the baseline dataset and the other two from the synthetic datasets, to compare LCA and K-means. The first synthetic dataset was created by increasing Physical Activity and Body Mass index for 40% of samples, in the range of a random number with normal distribution up to twice the variance. Finally, we chose the second synthetic dataset with only 20% of samples changed in the range of a random number with normal distribution up to the variance. With these three datasets we investigated the stability of each clustering method in the situations resulting from the changes in datasets.

5 Comparison of validation techniques

As discussed before in section 4, we chose three different datasets to compare LCA and K-means methods

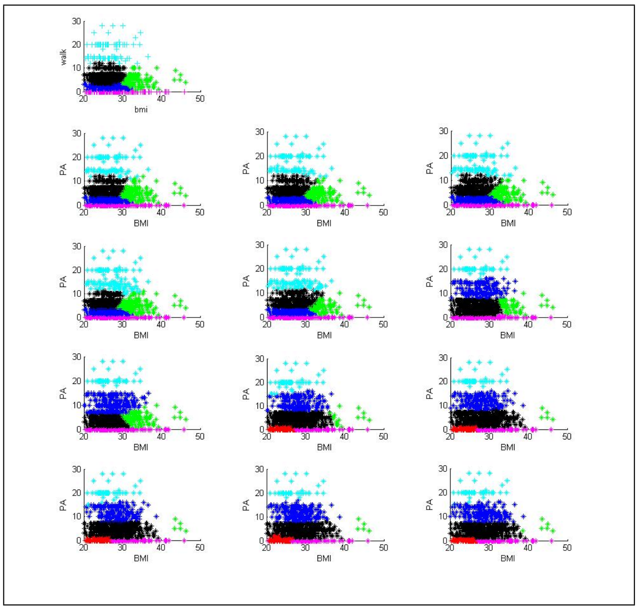


Figure 2: sequence of changing elements of larger cluster (BMI and PA increased by random number up to twice the variance)

for clustering our datasets (K-means method is chosen due to wide use). Table 1 shows the different range of validation methods for LCA and K-means clustering methods to cluster our baseline dataset. There are four different validation indexes computed for both LCA and K-means methods, while the number of clusters in each method varied from $K = 2...8$. The bold entries correspond to the optimal value predicted by each validation index.

Table 1: validation index- LCA and K means clustering for baseline dataset

Latent Class Analysis			
k	BIC	Silhouette	Dunn
2-Cluster	12099.3657	**0.5999**	1.3078
3-Cluster	11673.6888	0.4348	1.3251
4-Cluster	11606.441	0.4458	1.3723
5-Cluster	**11590.8389**	0.2919	1.3724
6-Cluster	11599.435	0.3435	1.9116
7-Cluster	11608.1122	0.4257	1.9115
8-Cluster	11634.5377	0.4538	**1.9979**
K means			
K	Silhouette	Davies-Bouldin	Dunn
2-Cluster	**0.46296**	0.88306	**2.0997**
3-Cluster	0.39821	0.85408	1.4749
4-Cluster	0.36353	0.82917	1.1644
5-Cluster	0.34664	0.8032	1.1575
6-Cluster	0.35132	0.73965	1.0911
7-Cluster	0.37077	**0.68475**	1.2532
8-Cluster	0.36513	0.69344	1.0437

In Table 1, clustering the baseline dataset using LCA shows the Silhouette index suggests that $K = 2$ has the best cluster configuration and may also suggest $K = 8$ be considered as a second option because it has second highest value for this index. Dunn's index is maximized at $K = 8$ and it might be of interest to consider $K = 6, 7$ as other options for choosing the number of clusters as they have the highest values. The BIC is minimized at $K = 5$ and based on Silhouette index and, Dunn's index at least $K = 5$ would be a reasonable choice for the number of clusters by using LCA clustering. Using K-means method, the Silhouette and Dunn's indexes are maximized at $k = 2$ ($K = 3$ has the second highest value for both of these indexes). However, the Davies-Bouldin index indicates that $K = 7$ has the best cluster configuration.

Table 2 shows the result for a synthetic dataset with 20% change of elements; Silhouette index for LCA clustering is maximized in the 2 cluster solution, however, the second highest value is the 7 cluster solution that one may infer as a second option for number of clusters. BIC suggests the 5 cluster solution and Dunn's index indicates 8 cluster solution for this synthetic dataset. With K-means clustering algorithm, Silhouette and Dunn's indexes suggest the 2 cluster solution while Davies-Bouldin index indicates that the 8 cluster solution is the optimal number of clusters.

Table 2: validation index- LCA and K means clustering for synthetic dataset with 20 percent change

Latent Class Analysis			
K	BIC	Silhouette	Dunn
2-Cluster	11251.4222	**0.5915**	1.3158
3-Cluster	10848.3937	0.4319	1.3336
4-Cluster	10805.5082	0.4472	1.3723
5-Cluster	**10790.0733**	0.3058	1.3723
6-Cluster	10800.7002	0.3137	1.3723
7-Cluster	10808.5815	0.4747	1.3723
8-Cluster	10826.1286	0.3669	**1.9167**
K means			
K	Silhouette	Davies-Bouldin	Dunn
2-Cluster	**0.4490**	0.9025	**2.0737**
3-Cluster	0.3968	0.8388	1.4866
4-Cluster	0.3632	0.8230	1.2061
5-Cluster	0.3417	0.7306	1.1751
6-Cluster	0.3527	0.8009	1.1495
7-Cluster	0.3749	0.7407	1.2887
8-Cluster	0.3669	**0.6597**	1.0308

Table 3: validation index- LCA and K means clustering for synthetic dataset with 40 percent change

Latent Class Analysis			
K	BIC	Silhouette	Dunn
2-Cluster	12215.7611	0.5023	1.3349
3-Cluster	11891.1496	0.3667	1.3185
4-Cluster	11873.3579	0.3752	1.3368
5-Cluster	11859.5514	0.3792	1.3723
6-Cluster	**11857.5536**	**0.5044**	1.86
7-Cluster	11870.1069	0.4038	1.9478
8-Cluster	11885.8412	0.3941	**1.9765**
K means			
K	Silhouette	Davies-Bouldin	Dunn
2-Cluster	**0.45064**	0.88824	**2.1344**
3-Cluster	0.38521	0.84444	1.4926
4-Cluster	0.35882	0.89291	1.3874
5-Cluster	0.34747	0.79671	1.2
6-Cluster	0.36214	0.81538	1.2627
7-Cluster	0.36771	**0.7499**	1.3059
8-Cluster	0.36395	0.80687	1.2128

Based on reported results in Table 3, using K-means for that synthetic dataset, the Silhouette index and Dunn's index suggest the 2 cluster solution and Davies-Bouldin index suggests the 7 cluster solution. Validation indexes for Latent class analysis show a more promising result, as presented in Table 3, with the optimal number of clusters based on BIC and Silhouette index a 6 cluster solution and the highest range of Dunn's index for $K = 6, 7, 8$.

6 Conclusions

The fundamental problem in unsupervised learning using clustering methods is to determine the number of clusters. Some methods like K-means try to assign each case to a cluster based on distance from each cluster center while others work based on the posterior probability of each case. Either way, using validation methods would give some insight for better understanding the quality of each cluster, and then based on specific domain knowledge decide on which clustering method is used and which model best explains the characteristics of our data. In this paper, we have used the K-means and LCA algorithms to cluster our data, and to explain a clustering solution for a synthetic dataset. Results of this work indicate that with a small change in data, LCA would still discover almost the same clusters as in the baseline dataset.

7 Acknowledgment

This research was completed using data collected through the 45 and Up Study (www.45andUp.org.au). The 45 and Up Study is managed by the Sax Institute in collaboration with major partner Cancer Council New South Wales; and partners the National Heart Foundation (NSW Division); NSW Health; Beyondblue: the national depression initiative; Ageing, Disability and Home Care, NSW Department of Human Services; and Uniting Care Ageing. We thank the many thousands of people participating in the 45 and Up Study.

References

Azuaje, F. (2002), 'A cluster validity framework for genome expression data', *Bioinformatics* **18**(2), 319–320.

Banks, E., Jorm, L., Lujic, S. & Rogers, K. (2009), 'Health, ageing and private health insurance: baseline results from the 45 and up study cohort', *Australia and New Zealand health policy* **6**(1), 16.

Banks, E., Redman, S., Jorm, L., Armstrong, B., Bauman, A., Beard, J., Beral, V., Byles, J., Corbett, S., Cumming, R. et al. (2008), 'Cohort profile: the 45 and up study', *Int J Epidemiol* **37**(5), 941–947.

Datta, S. & Datta, S. (2003), 'Comparisons and validation of statistical clustering techniques for microarray gene expression data', *Bioinformatics* **19**(4), 459.

Davies, D. & Bouldin, D. (1979), 'A cluster separation measure', *Pattern Analysis and Machine Intelligence, IEEE Transactions on* (2), 224–227.

Dunn, J. (1973), 'A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters'.

Dunn, J. (1974), 'Well-separated clusters and optimal fuzzy partitions', *Journal of cybernetics* **4**(1), 95–104.

Halkidi, M., Batistakis, Y. & Vazirgiannis, M. (2002), 'Cluster validity methods: part i', *ACM Sigmod Record* **31**(2), 40–45.

Jain, A., Murty, M. & Flynn, P. (1999), 'Data clustering: a review', *ACM computing surveys (CSUR)* **31**(3), 264–323.

Kryszczuk, K. & Hurley, P. (2010), 'Estimation of the number of clusters using multiple clustering validity indices', *Multiple Classifier Systems* pp. 114–123.

Lanza, S., Flaherty, B. & Collins, L. (2003), 'Latent class and latent transition analysis'.

Likas, A., Vlassis, N. & J Verbeek, J. (2003), 'The global k-means clustering algorithm', *Pattern recognition* **36**(2), 451–461.

Liu, Y., Li, Z., Xiong, H., Gao, X. & Wu, J. (2010), Understanding of internal clustering validation measures, *in* 'Data Mining (ICDM), 2010 IEEE 10th International Conference on', IEEE, pp. 911–916.

Magidson, J. & Vermunt, J. (2002), 'Latent class models for clustering: A comparison with k-means', *Canadian Journal of Marketing Research* **20**(1), 36–43.

of Health, A. I. & Welfare(AIHW) (2003), 'The active australia survey: A guide and manual for implementation, analysis and reporting', **Catalogue no. CVD 22. Canberra: AIHW**.

Rendón, E., Abundez, I., Gutierrez, C. & DÍAZ, S. (2011), A comparison of internal and external cluster validation indexes, *in* 'Proceedings of the 2011 American conference', pp. 158–163.

Rousseeuw, P. (1987), 'Silhouettes: a graphical aid to the interpretation and validation of cluster analysis', *Journal of computational and applied mathematics* **20**, 53–65.

Salzberg, S. (1991), 'Distance metrics for instance-based learning', *Methodologies for Intelligent Systems* pp. 399–408.

Schwarz, G. (1978), 'Estimating the dimension of a model', *The Annals of Statistics* **6**(2), 461–464.

Study Overview (2011).
URL: *http://www.45andup.org.au/*

Wang, K., Wang, B. & Peng, L. (2009), 'Cvap: Validation for cluster analyses', *Data Science Journal* (0), 904220071.

Welcome to Statistical Innovations Inc. (2011).
URL: *http://www.statisticalinnovations.com/*

The Role of Electronic Medical Records in the Identification of Suboptimal Prescribing for Hypertension Management: An Opportunity in Unchanged Therapy

DEPAK PATEL[1], JIM WARREN[2,3], JOHN KENNELLY[3]
[1]School of Medicine [2]Department of Computer Science [3]School of Population Health
University of Auckland
PO Box 92019, Auckland, New Zealand

jim@cs.auckland.ac.nz

KUINILETI CHANG WAI
West Fono Health Trust
411 Great North Road, Henderson, Auckland 0650, New Zealand

Abstract

A Participatory Action Research (PAR) approach was taken to identify electronic medical record (EMR) queries for hypertension management quality review in the context of a Pacific-led New Zealand general practice. In each PAR cycle, queries to identify patients with prescribing at variance from evidence-based practice were formulated and run, relevant patient notes were retrieved, and a quality audit of the medication decisions was carried out by a medical practitioner working in the practice. 764 enrolled and funded patients with current antihypertensive prescriptions were queried regarding adherence to national treatment guidelines. Queries based on drug classes indicated by specific comorbidities (e.g. hypertension complicated by diabetes) retrieved few cases, and with almost none having a compelling case for change in therapy upon review. A query on unchanged therapy while cardiovascular risk (CVR) and systolic blood pressure remained high, however, yielded 30 cases for review, and 10 of these were deemed as warranting further investigation. We conclude that a promising area for the use of EMR queries to improve long-term condition management is in identification of patients with persistently high risk of adverse outcomes and concurrent unchanged therapy during successive general practice visits.

Keywords: clinical quality improvement; computer-based patient records; electronic prescribing; hypertension management.

1 Introduction

Hypertension is a significant health burden due to its strong association with CVR and chronic renal disease (Kearney et al., 2005). Hypertension has a high prevalence, with 73 million (34%) individuals over 20 diagnosed with the condition in the US in 2005 (Rosamond et al., 2008). In terms of cardiovascular disease (CVD), an individual's risk doubles for each rise of 20/10mmHg, beginning at 115/75mmHg (Chobanian et al., 2003b).

Research has shown impressive efficacy rates of blood pressure (BP) lowering medications for reduction in cardiovascular risk and renal disease (Strippoli et al., 2005, Law et al., 2009). Although these drugs are effective when taken as directed, even when diagnosed and treated, patients frequently fail to achieve BP control to recommended levels (Chobanian et al., 2003a, National Committee for Quality Assurance, 2009). Part of the problem is certainly poor adherence; in fact, it is thought that poor adherence to antihypertensive medication contributes to inadequate BP control in more than two-thirds of hypertensive patients (Miller et al., 1997). Beyond adherence per se, however, are there other similarly large opportunities for improved BP control in better alignment of treatment decisions to evidence based best practice?

New Zealand is recognised in the top tier of nations with respect to information technology use in General Practice medicine (Schoen et al., 2009). Our research indicates that the electronic medical records (EMRs) held in General Practice systems are reasonably sensitive and specific for detecting patients whose hypertension management is suboptimal (Warren et al., 2008). We find that there is a large cohort of high-needs patients in New Zealand with poor adherence identifiable through General Practice EMRs (Mabotuwana et al., 2009a), and that poor adherence observable through these EMRs is associated with significantly reduced odds of BP control (Mabotuwana et al., 2009b).

The present study focuses on identifying queries to the General Practice EMR that, rather than being related to adherence issues, can function to identify other substantial case cohorts with specific opportunities to improve their hypertension management. We take a particular interest in management of hypertension for the Pacific population. The Pacific population in New Zealand (NZ) has grown dramatically since World War II, from 2,200 people in 1945 to 266,000 in 2006, with 66% living in the Auckland metropolitan area and Samoan being the largest Pacific ethnic group (Statistics New Zealand and Ministry for

Pacific Island Affairs, 2010). This Pacific population has a greater cardiovascular disease (CVD) risk than European New Zealanders (Sundborn et al., 2008).

2 Methodology

Data and Algorithms – EMRs of 5454 enrolled and funded patients of a Pacific-led New Zealand metropolitan general practice, having largely Pacific caseload, were extracted under protocol NTX/09/100/EXP of the Northern X Regional Ethics Committee. Data extraction included prescriptions, diagnosis codes, laboratory test results, BPs and CVRs (the practice made extensive use of PREDICT (Riddell et al.), which provides this value into the EMR) up to 15 May 2009. 764 patients had at least one antihypertensive prescription since January 1 2008. Medication Possession Ratio (MPR, percent of days covered by a prescription – a supply-based measure of adherence) was computed from the prescriptions in the EMRs through methods we have previously documented (Mabotuwana et al., 2009b, Mabotuwana and Warren, 2010). In brief, MPR and other EMR statistics were computed on data extracted from Medtech32 using its interactive reporting function and then imported into a Microsoft SQL Server database for further processing with a system of stored procedures and supporting data files (including lists of drug names and diagnosis codes) called the ChronoMedIt framework (Mabotuwana and Warren, 2010). Additional processing was conducted using study-specific SQL (structured query language) queries formulated as part of the procedure.

Procedure – Participatory Action Research (PAR) methodology has been endorsed and promoted internationally as a format for primary health care research, particularly in communities with high needs (Macaulay et al., 1999). Key elements of PAR were considered: (a) the use of an iterative plan, act, observe, reflect cycle; and (b) collaborative, collective and self-reflective enquiry (Baum et al., 2006, Kemmis and McTaggart, 1988). These elements were adapted for the development of evidence-based EMR queries, yielding a three-stage cycle (Figure 1). Firstly (A, Figure 1), hypertension prescribing quality improvement opportunities were considered in light of relevant guidelines and research literature, with reflection on what is known about the local cases (especially after the first cycle) and consideration of what criteria are amenable to automated assessment from the EMR. This led to an hypothesized opportunity to identify discrepancies between actual and evidence-based best practice in terms of specific criteria. Secondly (B, Figure 1), one or more EMR queries were formulated and run to identify cases at variance to the criteria. Finally (C, Figure 1), patient notes were retrieved, and a quality audit of the medication decisions in the patient notes carried out by a medical staff member working in the practice. The findings of this audit were considered in depth to assess the relevance of the query to actual practice and to inform the next PAR cycle. We executed PAR cycles over the period November 2008 to February 2009.

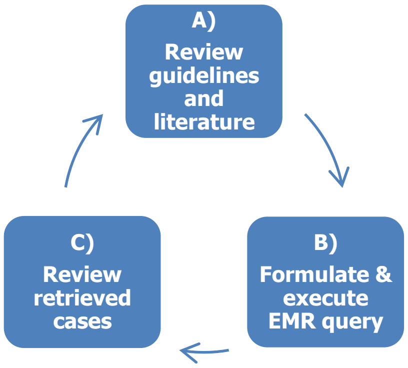

A) Review guidelines and literature
B) Formulate & execute EMR query
C) Review retrieved cases

Figure 1: PAR cycle for EMR query development

Due to the iterative nature of the PAR cycles, we report the details of our specific queries, along with the results of those queries, in the Results section below.

3 Results

A total of four queries were formulated, taking the NZ Guidelines Group *Cardiovascular Guidelines Handbook* (New Zealand Guidelines Group, 2009) as our primary guide, but supplementing with research literature around specific known deficiencies in hypertension management. Key concepts from the guidelines concern the indications for prescribing specific antihypertensive agents including angiotensin converting enzyme inhibitors (ACEi), angiotensin II receptor blockers (ARBs), beta-blockers and thiazide diuretics, as well as key comorbidities (other conditions that complicate treatment) of diabetes, previous myocardial infarction (MI, i.e. heart attack) and microalbuminuria (small amounts of protein in the urine).

Query 1. ACEi/ARB in diabetes – Our first query was founded on the recommendation of ACEi (or ARB if ACEi not tolerated) as preferred therapy in diabetes, and aggressive BP control if microalbuminuria is also present. Our query identified patients whose EMR data indicated:

- Diagnosis with diabetes with neither ACEi nor ARB (MPR=0, i.e. no ACEi or ARB prescription records from 1 January 2008 to 15 May 2009)

  and
- Diagnosis with hypertension or either:
  - A diagnosis of microalbuminuria and at least three systolic BPs ≥ 135mmHg, or
  - At least three systolic BPs ≥ 140mmHg

  and
- CVR≥15% recorded or diagnosis with microalbuminurea.

Figure 2 provides a breakdown of the results from this query; of the six cases retrieved, none presented a compelling case for change of therapy at the time of review.

Query 2. Beta-blocker and ACEi post-MI – Our second query was based on the recommendation to treat all people post-MI with beta-blocker and to consider adding an ACEi regardless of BP, especially if there is any significant left ventricular impairment. Lack of adherence to this recommendation had been observed for a US cohort, where only 64% had any beta-blocker and 52% any ACEi/ARB in the first three months after hospital discharge for acute coronary syndrome (Lee et al., 2008). Our EMR query required:

- Patients with recorded MI who have not received both a beta-blocker and an ACEi/ARB (i.e., MPR=0 for one, the other or both).
- Exclusion of patients with recorded diagnoses of asthma, heart block, peripheral vascular disease, sinus bradycardia, acute decompensated heart failure, hypotension, end stage renal failure or primary renal artery stenosis.

This query retrieved only two cases (see figure 3); clinicians characterised these as logistical issues in GP / cardiologist communications, rather than suboptimal prescribing patterns by the GPs per se.

The modest yield from the first two queries suggested that there were few identifiable opportunities for improvement where patients were indicated towards certain therapies due to comorbidities, such as diabetes and post-MI. This motivated a third query to investigate prescribing of first-line antihypertensives in patients not otherwise indicated towards therapies due to comorbidities.

Query 3. Thiazides as first-line antihypertensive – This query looked for those who, in the absence of other indicators, did not have a thiazide or thiazide-like diuretic as a first line antihypertensive agent at the time of a hypertension diagnosis. This query was motivated by calls internationally (Sweileh, 2009) and locally (van der Merwe, 2008) to keep thiazides in the therapeutic mix when treating hypertension. The query retrieved patients whose EMR data indicated:

- A hypertension diagnosis since 1 Jan 2007 (and none earlier than that).
- No prescription of a thiazide/thiazide-like diuretic prior to or 90 days following their hypertension diagnosis
- A CVR≥15% recorded or a diagnosis of microalbuminurea
- Never diagnosed with diabetes
- Never diagnosed with gout

Figure 4 shows the breakdown of the results of this query; two of the 13 cases retrieved were deemed in need of follow-up action upon clinical review.

The results of the first three queries led us to consider backing off to a more fundamental concept: 'clinical inertia' (Phillips et al., 2001). Rather than looking for the absence of a particular therapeutic agent, we would look for cases of poor control where nothing further had been tried.

Query 4. Non-intensification of antihypertensive therapy – Our fourth and final query identified patients who were receiving the same set of antihypertensive medications and at the same dose in the face of consistently high BP readings and high CVR. The criteria were:

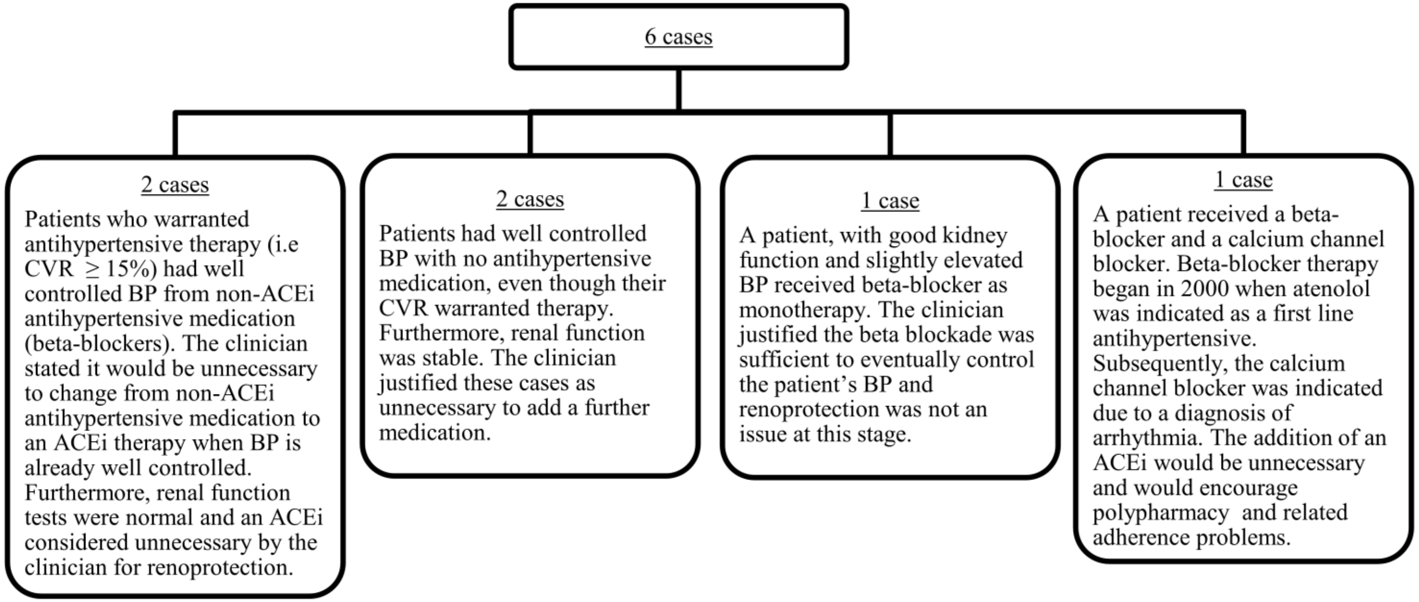
- Risk: recorded CVR ≥15%
- Identical antihypertensive prescription: same set and dose(s) of antihypertensives prescribed on two occasions, with the second coinciding to end of supply from the first (90 days ± 2 weeks)
- Sustained high BP: identical antihypertensive prescription subsequent to the recording of 3 high BP readings (above 'target' – see point below) and no BP not-high in the prior 200 days
- Target: for diabetic patients a systolic BP of ≥130mmHg was considered above target; for non-diabetics ≥140mmHg was considered above target
- No obvious adherence problem: have MPR ≥80% (for at least some antihypertensive) in the 6 month period prior to their second visit with identical antihypertensive prescription.

This query retrieved 30 cases, 10 of which were deemed in need of follow-up upon clinical review (see figure 5).

4 Discussion

We undertook an iterative programme of EMR query development with a Pacific-led general practice to find ways to identify automatically patients whose antihypertensive prescribing was at variance with evidence-based guidelines. Our initial two queries focused on drug/co-morbidity combinations (ACEi/ARB in diabetes; beta-blocker and ACEi/ARB post-MI), and a third on greater use of a drug (thiazide diuretic). These queries yielded few cases: 21 retrieved in total, with two warranting further investigation, from 764 patients with active antihypertensive therapy. A fourth query targeting unchanged therapy in light of established risk and sustained high BP had a much better yield: 30 retrieved with 10 warranting further investigation.

The low yield from our first three queries should not be taken as indicating that there is a lack of opportunity for better use of ACEi/ARB, beta-blocker and thiazide diuretics in management of hypertension and co-morbidities. It does, however, indicate two barriers to better use. First, review of Figures 2-4 reveals factors inhibiting a simple progression to the maximal guideline-indicated therapy, including patients who are stable on therapy they have had for some time, concern that the polypharmacy effect of adding another agent may outweigh the benefits, and logistic issues of a patient seeing different doctors. Second, there is the challenge of case identification. To identify patients via the EMR, the appropriate data must be present, such as BPs, diagnoses (and/or in some cases laboratory test results) and, ideally, CVRs. The practice we worked with has a dedication to CVD management for its largely Pacific Island community of patients, and with this a dedication to use of PREDICT, which provides CVRs, and may account for the low number of cases at variance from guidelines. Nonetheless, we have probably erred on the conservative side in query formulation and thus have missed cases because the data for inclusion is not encoded in the EMR.



6 cases

2 cases

Patients who warranted antihypertensive therapy (i.e CVR ≥ 15%) had well controlled BP from non-ACEi antihypertensive medication (beta-blockers). The clinician stated it would be unnecessary to change from non-ACEi antihypertensive medication to an ACEi therapy when BP is already well controlled. Furthermore, renal function tests were normal and an ACEi considered unnecessary by the clinician for renoprotection.

2 cases

Patients had well controlled BP with no antihypertensive medication, even though their CVR warranted therapy. Furthermore, renal function was stable. The clinician justified these cases as unnecessary to add a further medication.

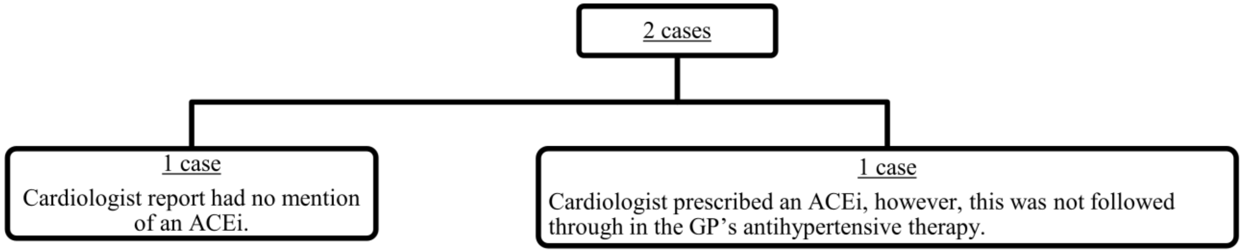
1 case

A patient, with good kidney function and slightly elevated BP received beta-blocker as monotherapy. The clinician justified the beta blockade was sufficient to eventually control the patient's BP and renoprotection was not an issue at this stage.

1 case

A patient received a beta-blocker and a calcium channel blocker. Beta-blocker therapy began in 2000 when atenolol was indicated as a first line antihypertensive. Subsequently, the calcium channel blocker was indicated due to a diagnosis of arrhythmia. The addition of an ACEi would be unnecessary and would encourage polypharmacy and related adherence problems.

Figure 2: Results of ACEi/ARB in diabetes query



2 cases

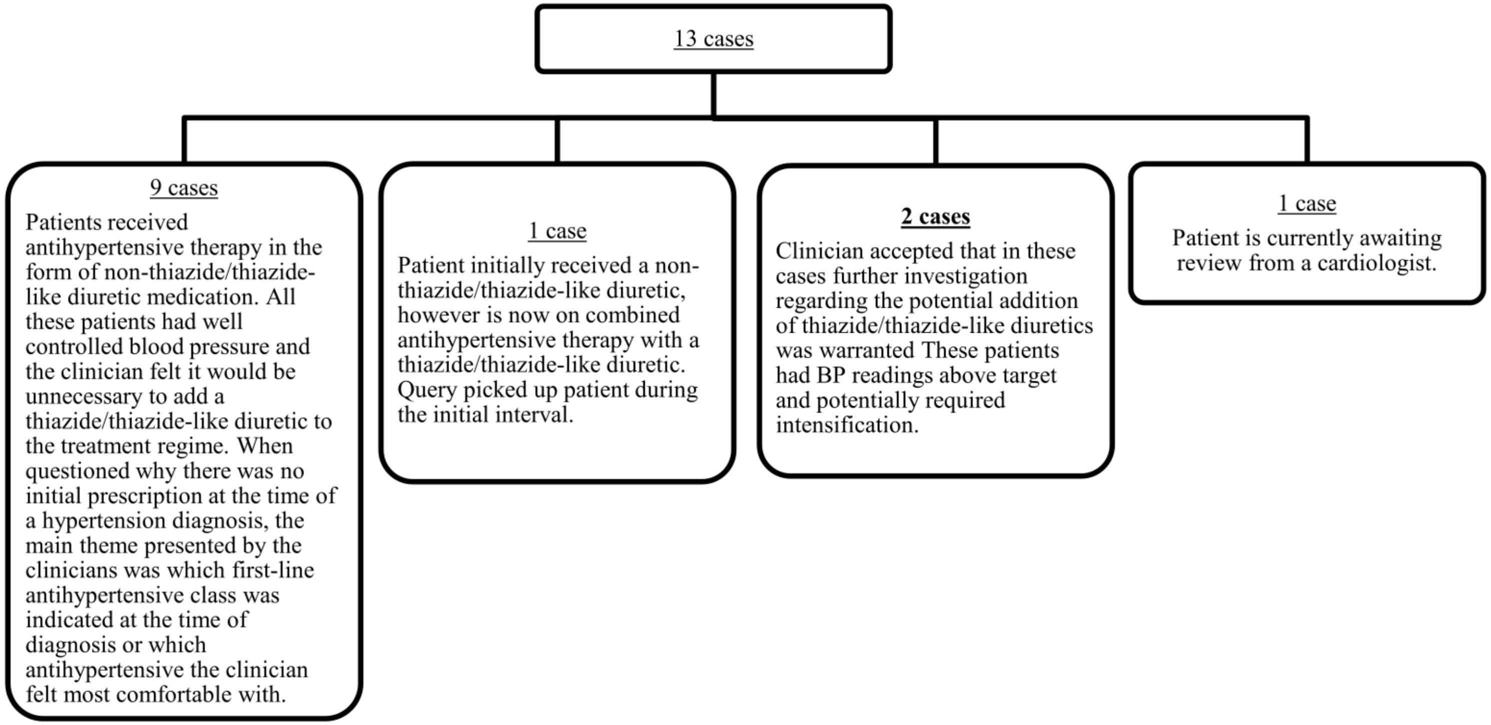
1 case

Cardiologist report had no mention of an ACEi.

1 case

Cardiologist prescribed an ACEi, however, this was not followed through in the GP's antihypertensive therapy.

Figure 3: Result of beta-blocker and ACEi post-MI query



13 cases

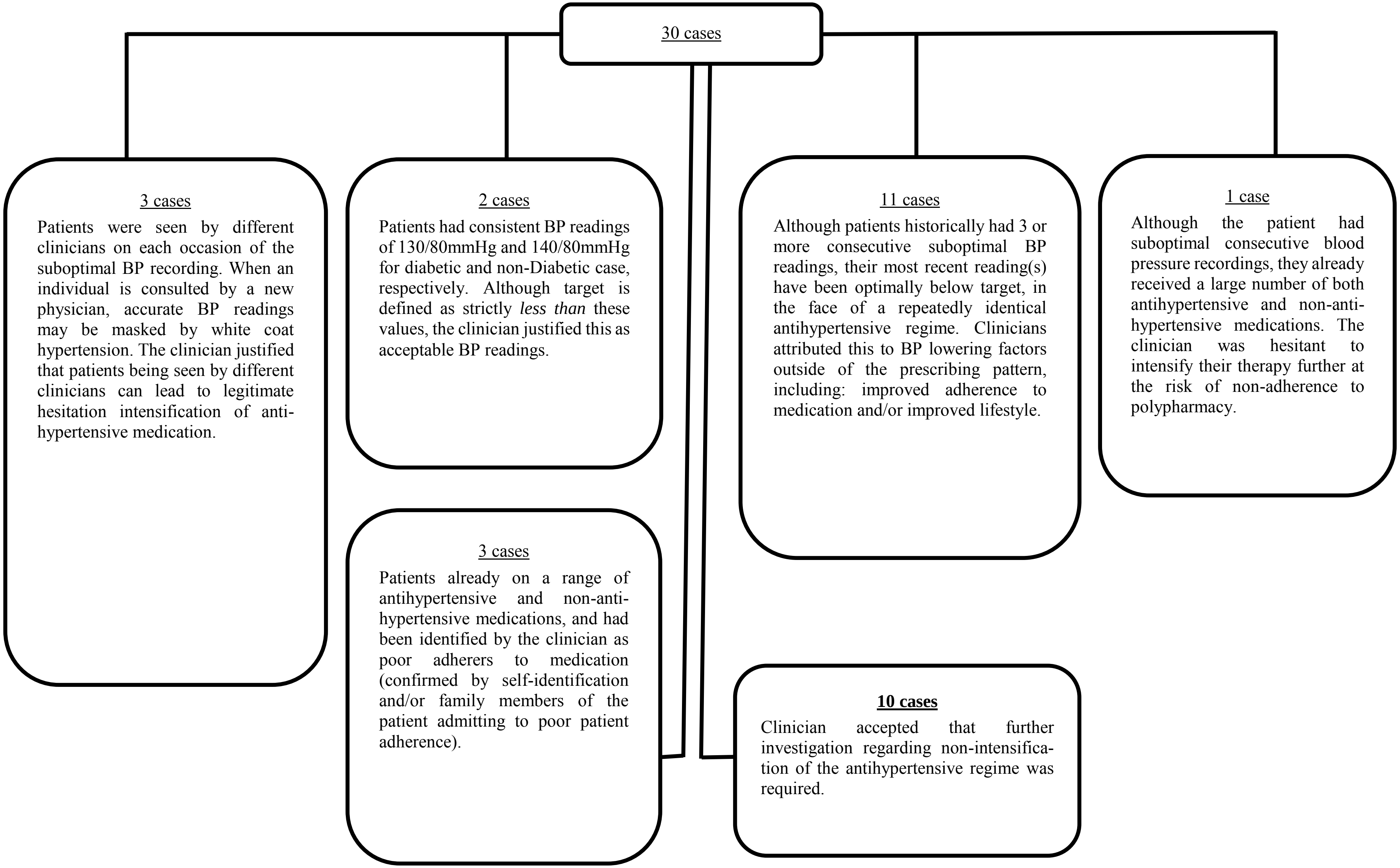
9 cases
Patients received antihypertensive therapy in the form of non-thiazide/thiazide-like diuretic medication. All these patients had well controlled blood pressure and the clinician felt it would be unnecessary to add a thiazide/thiazide-like diuretic to the treatment regime. When questioned why there was no initial prescription at the time of a hypertension diagnosis, the main theme presented by the clinicians was which first-line antihypertensive class was indicated at the time of diagnosis or which antihypertensive the clinician felt most comfortable with.

1 case
Patient initially received a non-thiazide/thiazide-like diuretic, however is now on combined antihypertensive therapy with a thiazide/thiazide-like diuretic. Query picked up patient during the initial interval.

2 cases
Clinician accepted that in these cases further investigation regarding the potential addition of thiazide/thiazide-like diuretics was warranted These patients had BP readings above target and potentially required intensification.

1 case
Patient is currently awaiting review from a cardiologist.

Figure 4: Results of query for lack of thiazide/thiazide-like diuretic as first-line therapy in absence of diabetes or gout.



30 cases

3 cases

Patients were seen by different clinicians on each occasion of the suboptimal BP recording. When an individual is consulted by a new physician, accurate BP readings may be masked by white coat hypertension. The clinician justified that patients being seen by different clinicians can lead to legitimate hesitation intensification of anti-hypertensive medication.

2 cases

Patients had consistent BP readings of 130/80mmHg and 140/80mmHg for diabetic and non-Diabetic case, respectively. Although target is defined as strictly *less than* these values, the clinician justified this as acceptable BP readings.

3 cases

Patients already on a range of antihypertensive and non-anti-hypertensive medications, and had been identified by the clinician as poor adherers to medication (confirmed by self-identification and/or family members of the patient admitting to poor patient adherence).

11 cases

Although patients historically had 3 or more consecutive suboptimal BP readings, their most recent reading(s) have been optimally below target, in the face of a repeatedly identical antihypertensive regime. Clinicians attributed this to BP lowering factors outside of the prescribing pattern, including: improved adherence to medication and/or improved lifestyle.

1 case

Although the patient had suboptimal consecutive blood pressure recordings, they already received a large number of both antihypertensive and non-anti-hypertensive medications. The clinician was hesitant to intensify their therapy further at the risk of non-adherence to polypharmacy.

10 cases

Clinician accepted that further investigation regarding non-intensification of the antihypertensive regime was required.

Figure 5: Results of query for unchanged therapy in light of persistently high BP.

The fourth query reveals an opportunity in unchanged therapy. ‘Clinical inertia’ has been examined in some depth in the context of diabetes, with indications that it is a major problem (Ziemer et al., 2005) but may be reduced with appropriate computerized alerts (Ziemer et al., 2006). Schmittdiel et al. (2008) found that lack of therapy intensification was a more common problem than lack of adherence for reaching risk targets in diabetes patients. This is at variance with our findings where adherence problems to long-term medications appear to be present in 50% of cases (Mabotuwana et al., 2009a). That said, our query was highly conservative in that it required multiple BPs and a CVR to be present in the EMR; many cases of unchanged therapy would be missed by this query due to absence of these data. Moreover, we only looked at exactly identical drugs and doses on two visits spaced close to 90 days apart; there are many more prescribing patterns that could be classed as inertia or lack of intensification.

Caution is warranted around the interaction of adherence and unchanged therapy. Heisler et al. (2008) note a lack of influence of adherence on dose intensification which they point out as worrying – a patient suddenly moving into compliance could suffer hypotension. Particularly with the elderly, falls risk must also be considered as a reason for moderation in antihypertensive treatment. Review of Figure 5 shows known poor adherence as one of the reasons for unchanged therapy in our cohort – there would be further cases of unknown poor adherence. Nonetheless, the unchanged therapy cohort represents an area of opportunity to manage down patient CVR within the scope of normal general practice activities.

General practice EMRs provide a rich resource to target quality improvement in treatment of long-term conditions, including hypertension. Automatic case identification from the EMR, however, is an area still in need of further study to assess its sensitivity, specificity and overall value to effectiveness of healthcare delivery. Searching out patients with unchanged antihypertensive therapy in the face of significant risk and sustained high blood pressure is one promising direction for further development and evaluation. Such queries are different in technical form in terms of looking for absence of change as compared to simply looking for the co-occurrence of the presence of conditions or actions as is done for a drug-drug or drug-problem interaction. We are some way from having a visual query builder or other easy tool for end users to explore unchanged therapy scenarios on their caseloads.

Decision support of the type illustrated herein is dependent on EMR data that is structured for automated interpretation. In environments such as New Zealand (or Australian) general practice medicine, electronic prescribing is virtually universal and forms the foundation to identify unchanged therapy and medication possession ratios consistent with medication adherence. The other requirement is an outcome measure. In the case of blood pressures their recording into the EMR is potentially ad hoc – the software supports their entry as structured observations, but they may also simply be entered as text in the practice notes. Fortunately, blood pressures are relatively easy to detect in an automated scan of the notes. Increasing use of devices that interoperate with the EMR to log blood pressures automatically (as home monitoring devices, or in the practice), have the potential to make the tracking of outcome more reliable. Our case was further facilitated by the use of CVR computations that interoperate with the EMR system, providing additional support that these high blood pressures were indeed important to manage down.

We restricted our analysis to blood pressure management (which is, in and of itself, a huge area of opportunity). The work should extend readily for related cardiovascular risk factors around management of cholesterol and blood sugar, where outcome measures are typically tracked routinely and would be expected to be recorded as structured observations (the relevant laboratory tests for these conditions are automatically transmitted to the general practice system in New Zealand). In theory, the approach may extend to other long-term conditions such as in the mental health domain, but would be dependent on regular tracking of outcomes (e.g. as a PHQ-9 depression score (Spitzer et al., 1999)). It is worth noting that a further and related decision support opportunity is for systems to prompt for the requisite outcome measures when they are absent.

5 Conclusions

A query on unchanged therapy while cardiovascular risk (CVR) and systolic blood pressure remained high yielded 30 cases for review, 10 of which were deemed as warranting further investigation. In contrast, EMR queries based narrowly on drug class or comorbidity criteria yielded small numbers of cases and mostly patients with reasons for maintaining present therapy that are at least as compelling as the reasons for change. On this basis we see queries for unchanged therapy in the presence of concurrent high risk of adverse event (such as the CVD events like heart attacks, as well as kidney damage, in the case of hypertension) as a promising direction for use of the EMR to guide quality improvement efforts with respect to prescribing to manage long-term conditions.

6 Acknowledgments

Our thanks are extended to the medical director of the participating practice, to A/Prof Tim Kenealy for additional clinical guidance and Debra Warren and Thusitha Mabotuwana for technical support in query formulation, as well as to Yulong Gu for support in manuscript preparation. This research was supported in part by a National Heart Foundation Summer Studentship award and a Partnerships for Excellence grant from the New Zealand Tertiary Education Commission.

7 References

Baum, F., Macdougall, C. & Smith, D. (2006) Participatory action research. *J Epidemiol Community Health,* 60**,** 854-7.

Chobanian, A.V., Bakris, G.L., Black, H.R., Cushman, W.C., Green, L.A., Izzo, J.L., Jr., Jones, D.W., Materson, B.J., Oparil, S., Wright, J.T., Jr. & Roccella, E.J. (2003a) The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure: the JNC 7 report. *JAMA,* 289**,** 2560-72.

Chobanian, A.V., Bakris, G.L., Black, H.R., Cushman, W.C., Green, L.A., Izzo Jr, J.L., Jones, D.W., Materson, B.J., Oparil, S. & Wright Jr, J.T. (2003b) Seventh report of the Joint National Committee on prevention, detection, evaluation, and treatment of high blood pressure. *Hypertension,* 42**,** 1206.

Heisler, M., Hogan, M.M., Hofer, T.P., Schmittdiel, J.A., Pladevall, M. & Kerr, E.A. (2008) When more is not better: treatment intensification among hypertensive patients with poor medication adherence. *Circulation,* 117**,** 2884-92.

Kearney, P.M., Whelton, M., Reynolds, K., Muntner, P., Whelton, P.K. & He, J. (2005) Global burden of hypertension: analysis of worldwide data. *The Lancet,* 365**,** 217-223.

Kemmis, S. & Mctaggart, R. (1988) *The Action Research Planner, 3rd ed.,* Geelong, Deakin University.

Law, M.R., Morris, J.K. & Wald, N.J. (2009) Use of blood pressure lowering drugs in the prevention of cardiovascular disease: meta-analysis of 147 randomised trials in the context of expectations from prospective epidemiological studies. *BMJ,* 338**,** b1665.

Lee, H.Y., Cooke, C.E. & Robertson, T.A. (2008) Use of secondary prevention drug therapy in patients with acute coronary syndrome after hospital discharge. *J Manag Care Pharm,* 14**,** 271-80.

Mabotuwana, T. & Warren, J. (2010) ChronoMedIt--a computational quality audit framework for better management of patients with chronic conditions. *J Biomed Inform,* 43**,** 144-58.

Mabotuwana, T., Warren, J., Harrison, J. & Kenealy, T. (2009a) What can primary care prescribing data tell us about individual adherence to long-term medication?-comparison to pharmacy dispensing data. *Pharmacoepidemiol Drug Saf,* 18**,** 956-64.

Mabotuwana, T., Warren, J. & Kennelly, J. (2009b) A computational framework to identify patients with poor adherence to blood pressure lowering medication. *Int J Med Inform,* 78**,** 745-56.

Macaulay, A.C., Commanda, L.E., Freeman, W.L., Gibson, N., Mccabe, M.L., Robbins, C.M. & Twohig, P.L. (1999) Participatory research maximises community and lay involvement. North American Primary Care Research Group. *BMJ,* 319**,** 774-8.

Miller, N.H., Hill, M., Kottke, T. & Ockene, I.S. (1997) The multilevel compliance challenge: recommendations for a call to action. A statement for healthcare professionals. *Circulation,* 95**,** 1085-90.

National Committee for Quality Assurance (2009) The State of Health Care Quality 2009. Washtington, D.C.

New Zealand Guidelines Group (2009) New Zealand Cardiovascular Guidelines Handbook: A summary resource for primary care practitioners. 2nd edition.

Phillips, L.S., Branch, W.T., Cook, C.B., Doyle, J.P., El-Kebbi, I.M., Gallina, D.L., Miller, C.D., Ziemer, D.C. & Barnes, C.S. (2001) Clinical inertia. *Ann Intern Med,* 135**,** 825-34.

Riddell, T., Wells, S., Jackson, R., Lee, A.W., Crengle, S., Bramley, D., Ameratunga, S., Pylypchuk, R., Broad, J., Marshall, R. & Kerr, A. Performance of Framingham cardiovascular risk scores by ethnic groups in New Zealand: PREDICT CVD-10. *N Z Med J,* 123**,** 50-61.

Rosamond, W., Flegal, K., Furie, K., Go, A., Greenlund, K., Haase, N., Hailpern, S.M., Ho, M., Howard, V. & Kissela, B. (2008) Heart disease and stroke statistics--2008 update: a report from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. *Circulation,* 117**,** e25.

Schmittdiel, J.A., Uratsu, C.S., Karter, A.J., Heisler, M., Subramanian, U., Mangione, C.M. & Selby, J.V. (2008) Why don't diabetes patients achieve recommended risk factor targets? Poor adherence versus lack of treatment intensification. *Journal of General Internal Medicine,* 23**,** 588-594.

Schoen, C., Osborn, R., Doty, M.M., Squires, D., Peugh, J. & Applebaum, S. (2009) A survey of primary care physicians in eleven countries, 2009: perspectives on care, costs, and experiences. *Health Aff (Millwood),* 28**,** w1171-83.

Spitzer, R.L., Kroenke, K. & Williams, J.B. (1999) Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. *JAMA,* 282**,** 1737-44.

Statistics New Zealand and Ministry for Pacific Island Affairs (2010) Demographics of New Zealand's Pacific Population. Wellington.

Strippoli, G.F., Craig, M. & Craig, J.C. (2005) Antihypertensive agents for preventing diabetic kidney disease. *Cochrane Database Syst Rev,* CD004136.

Sundborn, G., Metcalf, P.A., Gentles, D., Scragg, R.K., Schaaf, D., Dyall, L., Black, P. & Jackson, R. (2008) Ethnic differences in cardiovascular disease risk factors and diabetes status for Pacific ethnic groups and Europeans in the Diabetes Heart and Health Survey (DHAH) 2002-2003, Auckland New Zealand. *N Z Med J,* 121**,** 28-39.

Sweileh, W.M. (2009) Target blood pressure attainment in diabetic hypertensive patients: need for more diuretics? *International journal of clinical pharmacology and therapeutics,* 47**,** 434.

Van Der Merwe, W. (2008) Establishment of a Difficult Hypertension Clinic in Whangarei, New Zealand: the first 18 months. *N Z Med J,* 121**,** 63-72.

Warren, J., Gaikwad, R., Mabotuwana, T., Kennelly, J. & Kenealy, T. (2008) Utilising practice management system data for quality improvement in use of blood pressure lowering medications in general practice. *N Z Med J,* 121**,** 53-62.

Ziemer, D.C., Doyle, J.P., Barnes, C.S., Branch, W.T., Jr., Cook, C.B., El-Kebbi, I.M., Gallina, D.L., Kolm, P., Rhee, M.K. & Phillips, L.S. (2006) An intervention to overcome clinical inertia and improve diabetes mellitus control in a primary care setting: Improving Primary Care of African Americans with Diabetes (IPCAAD) 8. *Arch Intern Med,* 166**,** 507-13.

Ziemer, D.C., Miller, C.D., Rhee, M.K., Doyle, J.P., Watkins, C., Jr., Cook, C.B., Gallina, D.L., El-Kebbi, I.M., Barnes, C.S., Dunbar, V.G., Branch, W.T., Jr. & Phillips, L.S. (2005) Clinical inertia contributes to poor diabetes control in a primary care setting. *Diabetes Educ,* 31**,** 564-71.

Analysing homogenous patient journeys to assess quality of care for patients admitted outside of their 'home-ward'

LUA PERIMAL-LEWIS[1], SHAOWEN QIN[1], CAMPBELL H THOMPSON[2],

PAUL HAKENDORF[3]

[1] School of Computer Science, Engineering and Mathematics
Flinders University of South Australia

[2] Medicine
The University of Adelaide, South Australia

[3] Clinical Epidemiology Unit
Flinders Medical Centre, South Australia

lua.perimal-lewis@flinders.edu.au, shaowen.qin@flinders.edu.au, campbell.thompson@adelaide.edu.au,
Paul.Hakendorf@health.sa.gov.au

Abstract

This study is the first to explore the quality of care based on the outlier or the inlier status of patients for a large heterogeneous General Medicine (GM) service at a busy public hospital. The study compared the quality of care between ward outliers and ward inliers based on a homogenous group of patients using Two-step clustering method. Contrary to common perception, ward outliers had overall shorter Length of Stay (LOS) than ward inliers. The study also was unable to support the perception of shorter LOS in the outlier group being associated with higher in-hospital mortality. The study confirmed that overall the outliers received inferior quality of care as discharge summaries for the outliers were delayed and more outliers were re-admitted within 7 days of discharge in comparison to the inliers.

Keywords: *homogenous in-patient journey analysis, process mining, 'home-ward', outliers, inliers, quality of care, cluster analysis*

1 Introduction

Australian Public Hospitals are faced with increasing demands for hospital services. This is largely due to the aging Australian population and its associated demand for the usage of acute care facilities. The demand on Australian Emergency Department (ED) has been consistently increasing at an average of 1.8% per annum (FitzGerald, Toloo et al., 2012). Over the last 2 decades Australians' median age has increased by 4.8 years and population projections suggest increase in the proportion of population over the age of 65 thereby indicating that the demand on ED services will be an ongoing issue (FitzGerald, Toloo et al., 2012).

As with any organisations, hospitals too have to comply with strict measure of operational efficiency and effectiveness by conforming to Key Performance Indicators (KPI). Hospitals have mature processes that collect data on various quality measures in order to report and adhere to these KPIs. One such KPI is the improvement in ED throughput measures such as reducing the time first seen by a doctor, reducing did-not-wait rates and the reduction in ED Length of Stay (LOS) (Shetty, Gunja et al., 2012). Patient LOS is one of the criteria used to measure ED performance and a hospital's performance in general. Performance is measured as percentage of patients who stayed beyond the established LOS target (Kolker, 2008). ED LOS measurement, although a functional performance indicator, could possibly contribute to the streaming of patients to any available wards regardless of whether the ward is an appropriate ward for the condition of the patient.

The complexity and diversity of hospital processes means that there are also diverse ways to measure the quality of patient care which varies based on the characteristics of the process area being studied. This study investigated the quality of care received by patients who were admitted to their 'home-ward' referred to as inliers and patients who were admitted outside of their 'home-ward' referred to as outliers. It is a common perception amongst clinicians that outliers have longer overall in-hospital LOS compared to inliers. It is also perceived that quality of care received by outliers is inferior to that of inliers. At Flinders Medical Centre (FMC) where this study was undertaken, percentage of outlier patients was a regularly reported hospital performance indicator and therefore substantial effort is taken to collect the appropriate data needed in regards to the 'home-ward' status of the admitted patient.

The study indentified common variables or attributes used to measure quality of care and assessed how these attributes affect quality of care according to whether a patient was admitted to their 'home-ward' or outside of their 'home-ward'. The variables identified were *'discharge summary sent within 2 days of discharge'*, *'in-hospital mortality'*, *'re-admitted within 7 days'*, *'total in-hospital LOS'* and *'time spent in the ED'*.

Discharge summary contains relevant information pertinent to a patient's care during a hospital admission which is important to be communicated to primary health professionals who will continue a patient's care or provide future care for a patient after discharge (Li, Yong et al., 2011). The same authors established an association between delayed dissemination or the absence of discharge summary and re-admission rate thus encouraging health professionals to complete discharge summary promptly. Prompt discharge summary dissemination has also been associated with decreased hospital re-admission. Re-admission rate within 3 months decreased when a patient followed-up on continuity of care by seeing a physician who had received the discharge summary (Van Walraven, Seth et al., 2002). The hypothesis was that patients admitted outside of their 'home-ward'; the outliers will have higher re-admission rate because the discharge summaries for these patients were either not processed or delayed.

Inpatient LOS has become one of the many ways used to measure performance of a hospital. Patient mean LOS has been used to measure quality of care and hospital efficiency in terms of resource usage (Thomas, Guire et al., 1997). Lower than normal LOS could indicate that hospitals are discharging patients early possibly sacrificing quality of care (Thomas, Guire et al., 1997). The hypothesis was that outliers have longer overall LOS as their stay were probably prolonged as a consequence of being admitted outside of their 'home-ward' therefore not receiving the required level of care.

There are various studies establishing an association between ED overcrowding and in-hospital mortality. Richardson (2006) reported increased in-hospital mortality at 10 days amongst patients presenting at the ED during high ED occupancy. The hypothesis was that more patients would end up in an outlier ward during ED overcrowding due to the pressure to reduce ED congestion. As a consequence of inferior quality of care received by outliers, it was perceived that this group might have higher in-hospital mortality rate due to the delay in receiving treatment.

Health Care data analysis is traditionally done using various statistical techniques in order to report and hopefully forecast health care performances. New approaches in health care modelling and data analysis are emerging where more than one technique and approach are used to discover hidden information that might not be easily discovered from one approach. Combinations of techniques are used to complement each other. The use of Decision Support System (DSS) in health care is wide spread. DSS in Health Care industry could be divided into 2 broad categories. One category is used to help physicians with their day-to-day decision-makings. An example is a DSS based on clinical practice guideline in the management of diabetic patients (Lobach and Hammond, 1997). The other category of DSS is used by hospital management to make decisions for better hospital resource management. The fundamental information needed for such a system is based on the outcomes of some sophisticated methods of data analysis and modelling. The closer the outcome is in depicting the real scenario the better the DSS output.

Improving operational efficiency based on average bed occupancy alone is too weak to predict a complex hospital system and the dynamic nature of patient flow (Braitberg, 2007). A study that uses a combination of techniques to complement the strength in each technique will give a better in-sight. This study aims to investigate the relationship between the quality of care attributes in regards to the patient's inlier or outlier status by applying cluster analysis combined with statistical techniques. An in-depth evaluation of the patient flow processes using data from the Patient Journey Database was used to aid in identifying the relationships hidden within statistics alone.

2 Study Setting and Data

The analysis was undertaken on in-patient records for patients admitted to and discharged by the General Medicine (GM) service at Flinders Medical Centre (FMC). FMC is a public teaching hospital in South Australia and it attends to approximately 62,000 patients per annum. The GM service controlled about 100 in-patient beds out of about 500 beds in FMC as a whole. The analysis was carried out on in-patient records of the GM service only; that is, on those patients whose in-patient care had been allocated to a GM team. The wards that were 'home-wards' for this service were clearly defined. A home-ward is a ward that is equipped with the appropriate medical team and specialised equipment to treat the patient's primary disease. Patients who were not allocated a 'home-ward' of the GM unit responsible for their care were defined as being an outlier and staying in an outlier ward.

The Patient Journey Database from FMC contains information on in-patients or officially admitted patients only and records detailed information on the journey or movements of a patient from the time of admission to the time of discharge. An individual patient could have multiple admissions at different points in time and each admission will be allocated with a unique journey number that remains the same until discharge. Each movement of the patient from one ward to another ward is recorded with a timestamp, so at any point the "start time" in a ward and the "end time" in a ward are known together with the name of the ward. Each ward occupied by a patient is appropriately marked to reflect whether the ward occupied was an inlier or an outlier ward. Patient admitted to an inlier ward is admitted to their 'home-ward'. Timestamp for Admission is the combination the "Date" field and the "Admission Time" field. Timestamp for Discharge is the combination of "Date" field and "Discharge Time" field. Timestamp is a derived field. The individual patients are not identifiable at any point.

The original data set contained about 1.9 million records spanning from January 2003 to September 2009. To reduce the heterogeneous nature of the types of patients, various levels of record filtering were applied to reduce the dimension of the data set. The final record set which was used for the analysis only consisted of patient journeys that had been exclusively cared by the GM service from admission to discharge. If a patient's journey was under the care of a combination of GM service and non GM service, the journey was excluded. This level of filtering reduced the record set to about 24, 439 patient journeys.

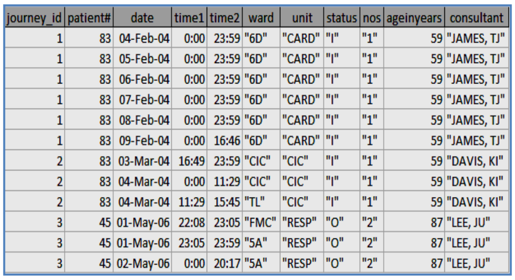
Ethics approval for the use of data from the patient journey database was granted by the Southern Adelaide Health Service / Flinders University Human Research Ethics Committee.

3 Methodology – Process Mining – Case Perspective

Process Mining uses event logs to discover organisational processes, control data, social and organisation structure (van der Aalst, Reijers et al., 2007). According to the same authors, processes could be analysed from the process perspective, the organisational perspective and the case perspective. The use of process mining techniques in the healthcare industry is becoming increasingly widespread. The complex nature of healthcare industry and varied processes makes the use of process mining techniques a viable method to gain insights into these processes (Perimal-Lewis, Qin et al., 2012). Mans, Schonenberg et al., (2008) used process mining techniques to identify bottleneck and to better understand the different clinical pathways taken by various groups of patients. Rebuge and Ferreira (2012) concluded that despite the proven success of process mining techniques, the complexity and the ad hoc nature of health data calls for the identification of right algorithm to handle noise in the data. It is common knowledge that healthcare industry is rich in data which presents a challenging task for researchers trying to discover knowledge using data from this domain. As with any knowledge discovery, gaining meaningful insight from data has to be accompanied with the knowledge rendered by domain experts to understand the intricacies behind complex health care decision making processes. The notion of efficient patient care providing patient-centred approach has seen the emergence of various Health Information Systems (Vezyridis, Timmons et al., 2011). Electronic Patient Management or Tracking Systems have all become not only common but essential systems for any hospital. These information systems store invaluable information that can be used for knowledge discoveries.

Process mining enables the discovery of knowledge regarding a process. Process mining uses event or process logs to extract information regarding a process as it has taken place (van der Aalst, Reijers et al., 2007). These process / event logs do not have to necessarily originate from a Workflow Management System. A process log could be derived from a dataset that contains an order of events which could be used to construct a process model that portrays the activity of the subject matter (van der Aalst, Reijers et al., 2007). In this study the event log was constructed from information collated from the Patient Journey Database. The process mining activities discussed in this paper is from the case perspective. The concept of event log as introduced by van der Aalst, Reijers et al., (2007) referred to as "history", "audit trail" and "transaction log" shapes the foundation of the event log used in this study. The individual patient journey is comparable to the concept of process instance introduced by the same authors. The information collated from the patient journey database is a derived event log. After constructing the event log, the event log was analysed as described and discussed in the following sections. The focus of the study is on gaining insight into the process of streaming patients to an outlier or an inlier ward and its effect on quality of care received by these patients.

Each patient journey is the process instance or the case being studied in relation to the activities on the patient journey. An activity is equivalent to a ward occupied by the patient which correlate to either an inlier or an outlier ward. Each case which in this study is the patient journey can be characterised by the values of the corresponding data elements (van der Aalst, Reijers et al., 2007). Data elements are the quality of care variables/attributes. The patient journeys are analysed to establish the relationship between the quality of care attributes of the journey against the amount of time a patient stayed in an outlier or an inlier ward. Table 1 shows a snippet of the dataset from the patient journey database.



journey_id	patient#	date	time1	time2	ward	unit	status	nos	ageinyears	consultant
1	83	04-Feb-04	0:00	23:59	"6D"	"CARD"	"I"	"1"	59	"JAMES, TJ"
1	83	05-Feb-04	0:00	23:59	"6D"	"CARD"	"I"	"1"	59	"JAMES, TJ"
1	83	06-Feb-04	0:00	23:59	"6D"	"CARD"	"I"	"1"	59	"JAMES, TJ"
1	83	07-Feb-04	0:00	23:59	"6D"	"CARD"	"I"	"1"	59	"JAMES, TJ"
1	83	08-Feb-04	0:00	23:59	"6D"	"CARD"	"I"	"1"	59	"JAMES, TJ"
1	83	09-Feb-04	0:00	16:46	"6D"	"CARD"	"I"	"1"	59	"JAMES, TJ"
2	83	03-Mar-04	16:49	23:59	"CIC"	"CIC"	"I"	"1"	59	"DAVIS, KI"
2	83	04-Mar-04	0:00	11:29	"CIC"	"CIC"	"I"	"1"	59	"DAVIS, KI"
2	83	04-Mar-04	11:29	15:45	"TL"	"CIC"	"I"	"1"	59	"DAVIS, KI"
3	45	01-May-06	22:08	23:05	"FMC"	"RESP"	"O"	"2"	87	"LEE, JU"
3	45	01-May-06	23:05	23:59	"5A"	"RESP"	"O"	"2"	87	"LEE, JU"
3	45	02-May-06	0:00	20:17	"5A"	"RESP"	"O"	"2"	87	"LEE, JU"

Table 1: Snippet of data used for process mining

The pre-processed dataset from the patient journey database as discussed in this section forms the source of data for the rest of the analysis.

3.1 Data analysis to define outlier patients and inlier patients

The next task was an explorative analysis to discover a meaningful way to categorise the population of the GM patients into 2 distinct categories of outlier patients and inlier patients. At any given time, it is possible to establish from the original dataset whether a patient stayed in an outlier or inlier ward. Majority of patients had stayed in a combination of outlier and inlier wards. The patient journeys were categorised according to those who had overall in-hospital LOS of "*0-3 Days*", "*4-7 Days*", "*8-30 Days*" and "*> 30 Days*". The percentage of time spent in an outlier ward and the percentage of time spent in an inlier ward were derived for each journey to show the distribution of the percentage of outlier time versus the percentage of inlier time.

It was discovered that the distribution of percentage of time spent in an outlier ward was very similar across all 4 categories of LOS. Figure 1 shows the distribution of outlier hours for the overall GM patient journeys.

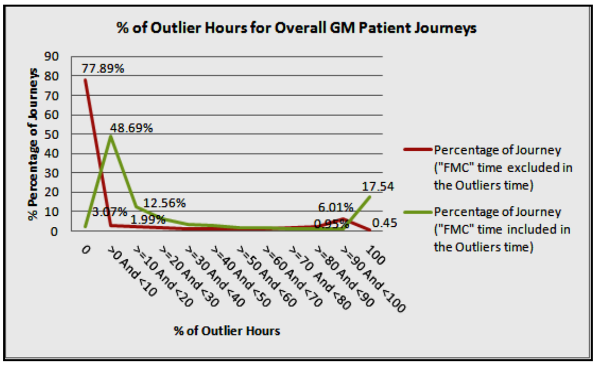


Figure 1: Distribution of the percentage of outlier time for the overall GM patient journeys

Figure 1 also shows the distribution of patient journeys with and without including the time spent waiting in the ED or ward "FMC" in the outlier hour calculation. This study was carried out on in-patient journeys, and in theory ward "FMC" or ED time should be zero however, this was not the case for many patient journeys. This indicates that many in-patients were spending time in the ED waiting for an in-patient bed to become available after decision to admit. Naturally, as far as the data recorded, the time in ED after decision to admit is considered as outlier time. The presentation of the above information assisted the domain experts to further deliberate on how the ED time should be classified in regards to the overall definition of the outlier and inlier status for the overall patient journeys, which was important to address as this might confound the findings.

The distribution for the different LOS categories is not presented here because the trend across the groups was similar. Based on the information discovered from this analysis together with the insight from the domain experts, it was decided that the best classification of inlier patient journeys will be those journeys that spent "≥ 70% Inlier Hours" in their 'home-ward' and the best classification for outlier patient journeys will be those journeys that spent "≥ 70% Outlier Hours" outside their 'home-ward'.

LOS	Number of journeys and (%)	Number and % of journeys with "≥ 70% Outlier Hours"	Number and % of journeys with "≥ 70% Inlier" Hours
0 - 3 Days	12012 (51.25)	3542 (29.49)	7123 (42.74)
4 - 7 Days	5637 (24.05)	758 (13.45)	4456 (122.85)
8 - 30 Days	5154 (21.99)	522 (10.13)	4223 (324.87)
Above 30 Days	636 (2.71)	54 (8.49)	531 (1091.53)

Table 2: Breakdown of patient journeys according to LOS and time in outlier and inlier ward

Table 2 shows the breakdown of patient journeys with "≥ 70% Inlier Hours" and "≥ 70% Outlier Hours" for each LOS category. This classification of the outlier and inlier group captured about 90% of the GM patient journeys.

The dataset was further filtered to remove patient journeys with 100% ED time as these patient journeys might confound LOS analysis for the outlier patient group. According to the domain experts these patients' health might have improved while waiting for a bed to become available and discharged from the ED as an outlier patient with short LOS. The other set of patient journeys that could also confound the outcome were those patient journeys who had stayed more than 30 days. According to the domain experts the longer a patient stays in the hospital, the more likely these patients would eventually end up in a 'home-ward'. Prolonged LOS for these patients are normally not related to medical issues but more likely related to finding appropriate care outside of the hospital.

After excluding patient journeys that were discharged from the ED, those staying more than 30 days and patient journeys with missing attributes the final sample size derived for the outlier group was 2592 records and for the inlier group was 15213 records.

The rest of the analysis is based on investigating the quality of care received by these 2 groups of patient journeys. The patients who stayed "<70%" of their in-hospital stay in an outlier or an inlier ward were not included in this analysis as the aim of this study was focussing on the outliers and the inliers.

3.2 Cluster analysis

Acknowledging the diversity of GM patients, as well as the complexity and variability embedded in each patient journey, it was important to reduce the heterogeneity of the patient journeys to gain better insight from the data. Disregarding patient heterogeneity can mask the discovery of meaningful patterns in patient characteristics which can lead to misleading results (Armstrong, Zhu et al., 2011). The aim of cluster analysis is to group cases, which in this study are the patient journeys, into homogenous groups based on the natural structure of data (Tan, Steinbach et al., 2005). Cluster analysis is an exploratory technique which aims to group cases into clusters based on their similarities and dissimilarities (Luke 2005). Cases in the same cluster share similar characteristics and very dissimilar to cases belonging to other clusters (Mooi and Sarstedt 2011). Applying statistical methods to a homogenous cluster of patients would be much more meaningful in revealing in-sights that are otherwise hidden due to heterogeneity.

In this study, the patient journeys were clustered using the two-step cluster analysis in SPSS. Two-step cluster analysis was chosen because of its ability to handle both continuous and categorical variables (SPSS 2001). From the automatic number of clusters derived by this clustering procedure an optimal number of clusters were derived using exploratory method while taking into consideration of the practicality of having large or small number of clusters against the ratio between clusters and the goodness of fit for the derived model. In the 1st step, the automatic number of clusters is determined using Clustering Criterion by choosing either Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC). The number of clusters derived for this data set using BIC and AIC were similar. SPSS computes the BIC and AIC for J clusters respectively as per equation (1) and equation (2) below (IBM 2011).

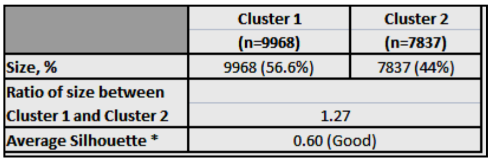
$$BIC(J) = -2\sum_{j=1}^{J}\xi_j + m_J \log(N), \tag{1}$$

$$AIC(J) = -2\sum_{j=1}^{J}\xi_j + 2m_J \tag{2}$$

In equation (1), N stands for total number of records in the data set. In equation (2), m_J is calculated as shown in equation (3). K^A, K^B and L_k in equation (3) stands for '*total number of continuous variables used in the procedure*', '*total number of categorical variables used in the procedure'* and *'number of categories for the kth categorical variable'* respectively (IBM 2011). In the 2nd step, the initial number of clusters derived in the 1st step is further refined. This is done by finding the largest increase in the distances between the 2 closest clusters (IBM 2011). The distance between 2 clusters is calculated by using log-likelihood distance measure, which is the decrease in log-likelihood as the clusters are combined into 1 cluster.

$$m_J = J\{2K^A + \sum_{k=1}^{K^B}(L_k - 1)\} \tag{3}$$

The patient journeys were clustered based on the quality of care variables and their outlier or inlier status. Patient journeys with outlier status were journeys with "≥ 70% Outlier Hours" and patient journeys with inlier status were journeys with "≥ 70% Inlier Hours". The variables chosen have been assessed for collinearity between variables to ensure that they were unique in identifying distinct clusters. Table 3 shows the clustering results of the 2 homogenous clusters.



	Cluster 1 (n=9968)	Cluster 2 (n=7837)
Size, %	9968 (56.6%)	7837 (44%)
Ratio of size between Cluster 1 and Cluster 2	1.27	
Average Silhouette *	0.60 (Good)	

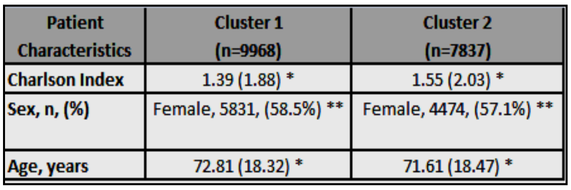
* Measure of cluster cohesion and separation

Table 3: Patient journey composition in the 2 clusters

Cluster cohesion measures how closely the cases in the cluster are related to other cases within the cluster and cluster separation measures how well a cluster is separated or different from other clusters (Tan, Steinbach et al., 2005). Silhouette Coefficient is the combination of both cohesion and separation for individual cases and clusters (Tan, Steinbach et al., 2005). Average Silhouette Coefficient ranges from -1 for very poor model and 1 for excellent model (Kaufman and Rousseeuw 2005). Based on the measurement as defined by Kaufman and Rousseeuw (2005), the model indicates a reasonable partitioning of data. The average Silhouette coefficient is calculated as per Equation (4), where '*A is the distance from the case to the centroid of every other cluster which the case belongs to*' and '*B is the minimal distance from the case to the centroid of every other cluster*' (IBM 2012).

$$(B-A) / \max(A,B) \tag{4}$$

The ratio between the smallest and largest cluster is 1.27 which is a good ratio as the larger cluster is less than 2 times larger than the smaller cluster. Between the 0.60 average Silhouette and the ratio, the model is a good fit for the purpose of this study where all the quality of care variable identified had to be included in the model to give the insight required.



Patient Characteristics	Cluster 1 (n=9968)	Cluster 2 (n=7837)
Charlson Index	1.39 (1.88) *	1.55 (2.03) *
Sex, n, (%)	Female, 5831, (58.5%) **	Female, 4474, (57.1%) **
Age, years	72.81 (18.32) *	71.61 (18.47) *

* Mean (SD) for continuous variables; ** Mode, n, (%) for dichotomous variables (indicating the most frequent category)

Table 4: Patient characteristics

The characteristics of patients in both clusters are listed in Table 4 above. The number of female patients is higher in both the clusters. Charlson co-morbidity Index (CI) is the most widely used clinical index for the evaluation of co-morbidities (Simon, Beland et al., 2012). CI is a pre-calculated variable for every patient admission and was supplied with the data set. Patients in cluster 2 had a higher Charlson co-morbidity Index (CI) score suggesting that these patients were sicker than those in cluster 1. Age differences between patients in both clusters were small and the difference is not clinically significant.

The table below (see Table 5) summarises the quality of care attributes and their relative importance in deriving the 2 clusters. The predictor importance for each quality of care attribute is calculated as per Equation (5) where *' Ω is the set of predictor and evaluation fields'* and *'sig $_j$ is the p-value*' (IBM 2012). The values are relative; therefore the sum of values for all attributes is 1. An attribute with a value close to 1 is the most important attribute in deriving the cluster and a value close to 0 is the least important attribute.

$$VI_j = \frac{-\log_{10}(sig_j)}{\max\limits_{j\in\Omega}(-\log_{10}(sig_j))} \tag{5}$$

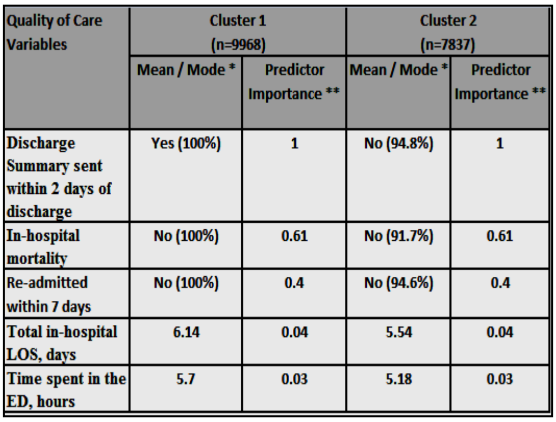
The most significant quality of care attribute for deriving the 2 homogenous clusters was *'discharge summary sent within 2 days of discharge'* with the relative importance of 1.0. Patient journeys in cluster 1 consists of patients where the discharge summaries were sent within 2 days for the entire, 100% of the cluster population as opposed to 94.8% of patient journeys who did not have their discharge summaries sent within 2 days suggesting inferior quality of care received by patients in Cluster 2.

The next quality of care attribute used to derive the 2 clusters was *'in-hospital mortality'* with relative importance value of 0.61. None of the patients in cluster 1 died during their hospital admission. 8.3% of patients in cluster 2 died.

The next quality of care attribute in order of importance used to derive the 2 clusters was *'readmission within 7 days'* with relative importance value of 0.4. Once again none of the patients in cluster 1 were re-admitted within 7 days; however 5.4% of patients in cluster 2 were re-admitted within 7 days.

The next quality of care attribute was *'total in-hospital LOS'* with relative importance value of 0.04. Patients in cluster 1 had a longer mean LOS (6.14 days) compared to patients in Cluster 2 with mean LOS of (5.54 days).

The final quality of care attribute was *'time spent in the ED'* with relative importance value of 0.03. Patients in cluster 1 spent slightly longer time in the ED (5.7 hours) compared to patients in cluster 2 with mean time of (5.18 hours).



Quality of Care Variables	Cluster 1 (n=9968)		Cluster 2 (n=7837)	
	Mean / Mode *	Predictor Importance **	Mean / Mode *	Predictor Importance **
Discharge Summary sent within 2 days of discharge	Yes (100%)	1	No (94.8%)	1
In-hospital mortality	No (100%)	0.61	No (91.7%)	0.61
Re-admitted within 7 days	No (100%)	0.4	No (94.6%)	0.4
Total in-hospital LOS, days	6.14	0.04	5.54	0.04
Time spent in the ED, hours	5.7	0.03	5.18	0.03

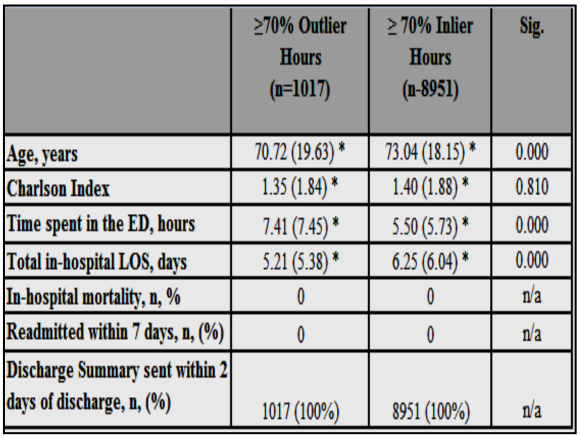
* Mean for continuous variables; Mode for binary variables (indicating the most frequent category); ** Relative importance of each quality of care variable/attributes in estimating the model

Table 5: Summary of quality of care variables/attributes

The next important step in this study was to investigate if there were any significant differences between the quality of care attributes and patient characteristics in both clusters for those patients in the outlier and the inlier groups defined earlier. Table 6 and Table 7 below summarises the quality of care attributes and patient characteristics for the outlier and inlier group in cluster 1 (n=9968) and cluster 2 (n=7837) respectively. The *'Sig.'* column shows the p-value where significance level α < 0.05 is considered significant. Mann-Whitney U test was used for significance level test for continuous variables. Chi-square test was used for significance level test for proportions.

In cluster 1, 10.20% of patient journeys were in the outlier category with the rest of the patient journeys under the inlier category.

The age difference between outliers and inliers were statistically significant (p=0.000; Mann-Whitney U test). According to domain experts this age difference is not of clinical importance. The difference in Charlson Index (CI) between the inliers and the outliers was not statistically significant (p=0.810; Mann-Whitney U test) suggesting that disease complexities was not an important characteristic in differentiating patients. Outliers in cluster 1 spent much longer time in the ED waiting for an in-patient bed to become available after the decision to admit compared with the inliers. The difference in ED time between the outlier and the inlier group was statistically significant (p=0.000; Mann-Whitney U test). Despite spending longer time in the ED waiting for in-patient bed, outliers had overall shorter in-hospital LOS compared with the inliers. The differences in the LOS is statistically significant (p=0.000; Mann-Whitney U test). As noted before, there was no in-hospital mortality for patients in cluster 1. All cluster 1 patients were not re-admitted within 7 days and their discharge summaries were sent within 2 days of discharge regardless of their outlier or inlier status.

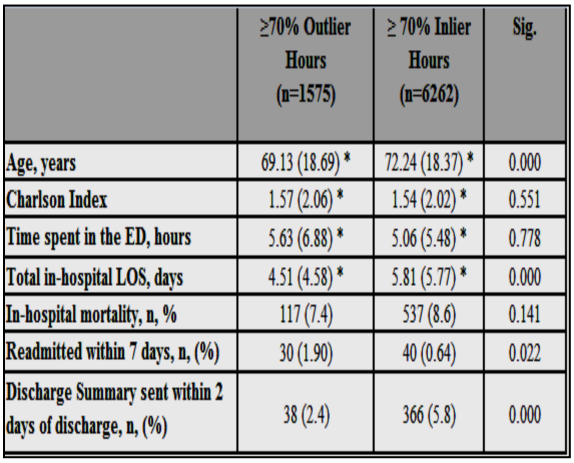


	≥70% Outlier Hours (n=1017)	≥ 70% Inlier Hours (n-8951)	Sig.
Age, years	70.72 (19.63) *	73.04 (18.15) *	0.000
Charlson Index	1.35 (1.84) *	1.40 (1.88) *	0.810
Time spent in the ED, hours	7.41 (7.45) *	5.50 (5.73) *	0.000
Total in-hospital LOS, days	5.21 (5.38) *	6.25 (6.04) *	0.000
In-hospital mortality, n, %	0	0	n/a
Readmitted within 7 days, n, (%)	0	0	n/a
Discharge Summary sent within 2 days of discharge, n, (%)	1017 (100%)	8951 (100%)	n/a

* Mean (SD) for continuous variables

Table 6: Quality of care attributes comparison for inliers and outliers in cluster 1

In cluster 2, there were 20.1% of outlier patient journeys and the rest were inliers. (see Table 7). Charlson Index (CI) was not statistically significant between the outliers and the inliers. This was similar to patients in cluster 1. Age differences between the outliers and the inliers were statistically significant (p=0.000; Mann-Whitney U test), however as noted before this is not of clinical significance. Contrary to patient journeys in cluster 1, although outliers spent slightly longer time in the ED compared to the inliers this was not statistically significant (p=0.778; Mann-Whitney U test) for patients in cluster 2. Similar to outliers in cluster 1, outliers in cluster 2 had shorter overall in-hospital LOS compared with the inliers and this was statistically significant (p=0.000; Mann-Whitney U test). The main differences between patients in cluster 1 and cluster 2 is in relation to the 3 quality of care attributes; *'in-hospital mortality'*, *'readmitted within 7 days'* and *'discharge summary sent within 2 days of discharge'*. All patients with inferior quality of care in relation to these 3 attributes were in cluster 2. In-hospital mortality between outliers and inliers in this cluster was not statistically significant. Outliers were re-admitted more than the inliers and this was statistically significant (p=0.022; chi-square test) suggesting that quality of care for outliers were inferior to those who were inliers. Less outliers had their discharge summaries sent within 2 days of discharge compared to the inliers and this was statistically significant (p=0.000; X^2 test). This again suggests an inferior quality of care for the outliers.



	≥70% Outlier Hours (n=1575)	≥ 70% Inlier Hours (n=6262)	Sig.
Age, years	69.13 (18.69) *	72.24 (18.37) *	0.000
Charlson Index	1.57 (2.06) *	1.54 (2.02) *	0.551
Time spent in the ED, hours	5.63 (6.88) *	5.06 (5.48) *	0.778
Total in-hospital LOS, days	4.51 (4.58) *	5.81 (5.77) *	0.000
In-hospital mortality, n, %	117 (7.4)	537 (8.6)	0.141
Readmitted within 7 days, n, (%)	30 (1.90)	40 (0.64)	0.022
Discharge Summary sent within 2 days of discharge, n, (%)	38 (2.4)	366 (5.8)	0.000

* Mean (SD) for continuous variables

Table 7: Quality of care attributes comparison for inliers and outliers in cluster 2

Another set of analysis was carried out to compare the differences between characteristics of patients in cluster 1 and cluster 2 (table not shown). Apart from Age with (p=0.066; Mann-Whitney U test), CI and Sex were significantly different between patients in cluster 1 and cluster 2 with (p=0.000; Mann-Whitney U test) respectively. All quality of care variables were significantly different between patients in cluster 1 and patients in cluster 2 with (p=0.000; Mann-Whitney U test).

4 Discussion

The main differences between patients in cluster 1 and cluster 2 relates to the quality of care attributes. Patients in cluster 2 had inferior quality of care compared to those in cluster 1 regardless of whether they were outliers or inliers. In cluster 1, there were no in-hospital mortality, none were re-admitted within 7 days and discharge summaries were sent within 2 days of discharge for all the patients. Analysing patients in cluster 2 (those who had inferior quality of care) in regards to the outlier and inlier status revealed meaningful in-sight as the comparison was done on a cluster of patients with similar characteristics and quality of care attributes. One of the major challenges of the study was the considerable effort that went into exploring the data to discover the best way to derive the outlier and the inlier population. Over the period of 6 years, investigating the spread of time spent in an outlier ward and the spread of time spent in an inlier ward lead to the dichotomisation of this variable into "≥ 70%" of outlier or inlier time. The method used and the dichotomisation of this variable was believed to be the best approach for this data set to discover the effect of being a ward outlier or ward inlier on the quality of care received by these 2 groups of patients.

It was also necessary to investigate the effect of the quality of care attributes on the outliers and inliers status based on a homogenous group of patients. The relationships discovered based on analysing the quality of care attributes on homogenous clusters were different when the patients were not clustered. This study demonstrates the complexity of analysing hospital data and the need to identify group of patients with more similar characteristics from the raw data. Although outliers in both clusters were younger and the association was statistically significant, it was not a clinically significant association. This emphasised the importance of involving domain experts to make meaningful conclusion.

Patient co-morbidity, (CI) did not have a significant association on whether the patient was admitted in a "home-ward" or outside of a "home-ward". This result was also obtained when the analysis was carried out without clustering the patients into 2 homogenous clusters.

There was a linear relationship between being an outlier or inlier and the amount of time spent in the ED. This association is only significant for outliers in cluster 1. The point-biserial correlation was used to capture the relationship between a dichotomous variable and a continuous variable (DeCoster and Claypool 2004). Point-biserial correlation showed a high correlation between the time spent in ED and being an outlier (r = 0.097, p = 0.000).

Contrary to the hypothesis, outliers in both clusters had shorter in-hospital LOS, similar association obtained when analysing the patient population without clustering. According to domain experts, this is a promising indicator as being an outlier did not compromise the efficiency of care in relation to the overall in-hospital LOS but outliers had inferior quality of care in relation to the extended time spent in the ED.

Point-biserial was used to further analyse the correlation between total in-hospital LOS and in-hospital mortality for patients in cluster 2. The correlation showed lower in-hospital LOS was associated with patients who did not die whilst in-hospital (r = - 0.139, p = 0.000). This finding calls for further research into the nature of this relationship. The finding reveals that outliers' short LOS is not associated with in-hospital mortality.

Using point-biserial for patients in cluster 2, lower in-hospital LOS was associated with patients who were not readmitted within 7 days of discharge (r = -0.042, p = 0.000) suggesting that re-admission might not necessarily be linked with shorter LOS or the outliers. Again, applying advanced modelling and analysis would reveal further in-sight to this association.

5 Conclusion & Future Work

In conclusion, patients in cluster 2 had significant association with inferior quality of care attributes. Outliers had shorter in-hospital LOS contrary to the hypothesis. Also, contrary to the results of un-clustered patient journeys, there were no significant association between being outlier and in-hospital mortality. Discharge summaries were not sent as promptly for the outliers compared to the inliers compromising continuation of care after discharge for the outlier group. Higher percentage of outliers was re-admitted within 7 days again suggesting inferior quality of care and conforming to the hypothesis.

Future work in regards to the inlier and outlier group of patients will include undertaking process mining to discover the patterns of patient movement and their correlation with LOS. Preliminary work on ward movement for the 2 groups has been initiated. Adapting process mining techniques to discover the control flow and ward movement for the two groups of patients will reveal further in-sight into the process of ward allocation in relation to quality of care.

Further analysis is needed to discover the reasons behind the longer ED time for outliers in Cluster 1 and why this is not a significant association for patients in cluster 2. Further study is also needed to discover the cause of shorter LOS for the outlying patients. According to domain experts, although LOS is a measure of efficiency, further analyses are needed to conclude the association with quality of care attributes. Additional in-sight is needed to understand the association between shorter in-hospital LOS and lower in-hospital mortality.

6 References

Armstrong, J. J., M. Zhu, et al. (2011). "K-Means Cluster Analysis of Rehabilitation Service Users in the Home Health Care System of Ontario: Examining the Heterogeneity of a Complex Geriatric Population." Archives of Physical Medicine and Rehabilitation.

Braitberg, G. (2007). "Emergency department overcrowding: dying to get in?" Med J Aust **187**(11-12): 624-625.

DeCoster, J. and H. Claypool. (2004). "Data Analysis in SPSS." Retrieved 25/08/2012, 2012, from http://www.stat-help.com/spss.pdf.

FitzGerald, G., S. Toloo, et al. (2012). "Demand for public hospital emergency department services in Australia: 2000-2001 to 2009-2010." Emerg Med Australas **24**(1): 72-78.

IBM. (2011). "Number of Clusters (auto-clustering) (TwoStep clustering algorithms)." Retrieved 27th October, 2012, from http://publib.boulder.ibm.com/infocenter/spssstat/v20r0m0/index.jsp?topic=%2Fcom.ibm.spss.statistics.cs%2Ftwostepcluster_table.htm.

IBM. (2012). "Goodness Measures (cluster evaluation algorithms)." Retrieved 20th August, 2012, from http://publib.boulder.ibm.com/infocenter/spssstat/v20r0m0/index.jsp.

Kaufman, L. and P. J. Rousseeuw (2005). Finding Groups in Data: Introduction to Cluster Analysis.

Kolker, A. (2008). "Process Modeling of Emergency Department Patient Flow: Effect of Patient Length of Stay on ED Diversion." Journal of Medical Systems **32**(5): 389-401.

Li, J. Y. Z., T. Y. Yong, et al. (2011). "Timeliness in discharge summary dissemination is associated with patients' clinical outcomes." Journal of Evaluation in Clinical Practice: no-no.

Lobach, D. F. and W. E. Hammond (1997). "Computerized decision support based on a clinical practice guideline improves compliance with care standards." Am J Med **102**(1): 89-98.

Luke, D. A. (2005). "Getting the big picture in community science: methods that capture context." Am J Community Psychol **35**(3-4): 185-200.

Mans, R., H. Schonenberg, et al. (2008). "Process mining techniques: an application to stroke care." Stud Health Technol Inform **136**: 573-578.

Mooi, E. and M. Sarstedt (2011). Cluster Analysis. A Concise Guide to Market Research, Springer-Verlag**:** 237 - 284.

Perimal-Lewis, L., S. Qin, et al. (2012). "Gaining Insight from Patient Journey Data using a Process-Oriented Analysis Approach " Australasian Workshop on Health Informatics and Knowledge Management (HIKM 2012) Melbourne, Australia **129**(ACS): 59-66

Rebuge, Á. and D. R. Ferreira (2012). "Business process analysis in healthcare environments: A methodology based on process mining." Information Systems **37**(2): 99-116.

Richardson, D. B. (2006). "Increase in patient mortality at 10 days associated with emergency department overcrowding." Med J Aust **184**(5): 213-216.

Shetty, A., N. Gunja, et al. (2012). "Senior Streaming Assessment Further Evaluation after Triage zone: A novel model of care encompassing various emergency department throughput measures." Emerg Med Australas **24**(4): 374-382.

Simon, T. G., M. D. Beland, et al. "Charlson Comorbidity Index predicts patient outcome, in cases of inoperable non-small cell lung cancer treated with radiofrequency ablation." European Journal of Radiology(0).

SPSS. (2001). "The SPSS TwoStep Cluster Component." SPSS - White Paper - Technical Report Retrieved 27th October, 2012, from http://www.spss.ch/upload/1122644952_The%20SPSS%20TwoStep%20Cluster%20Component.pdf.

Tan, P.-N., M. Steinbach, et al. (2005). Cluster Analysis: Basic Concepts and Algorithms. Introduction to Data Mining, Addison-Wesley**:** 487-568.

Thomas, J. W., K. E. Guire, et al. (1997). "Is patient length of stay related to quality of care?" Hosp Health Serv Adm **42**(4): 489-507.

van der Aalst, W. M. P., H. A. Reijers, et al. (2007). "Business process mining: An industrial application." Information Systems **32**(5): 713-732.

Van Walraven, C., R. Seth, et al. (2002). "Effect of Discharge Summary Availability During Post-discharge Visits on Hospital Readmission." Journal of General Internal Medicine **17**(3): 186-192.

Vezyridis, P., S. Timmons, et al. (2011). "Going paperless at the emergency department: a socio-technical study of an information system for patient tracking." Int J Med Inform **80**(7): 455-465.

Author Index

Recent Volumes in the CRPIT Series

ISSN 1445-1336

Listed below are some of the latest volumes published in the ACS Series *Conferences in Research and Practice in Information Technology*. The full text of most papers (in either PDF or Postscript format) is available at the series website `http://crpit.com`.

Volume 113 - Computer Science 2011
Edited by Mark Reynolds, The University of Western Australia, Australia. January 2011. 978-1-920682-93-4.
Contains the proceedings of the Thirty-Fourth Australasian Computer Science Conference (ACSC 2011), Perth, Australia, 1720 January 2011.

Volume 114 - Computing Education 2011
Edited by John Hamer, University of Auckland, New Zealand and Michael de Raadt, University of Southern Queensland, Australia. January 2011. 978-1-920682-94-1.
Contains the proceedings of the Thirteenth Australasian Computing Education Conference (ACE 2011), Perth, Australia, 17-20 January 2011.

Volume 115 - Database Technologies 2011
Edited by Heng Tao Shen, The University of Queensland, Australia and Yanchun Zhang, Victoria University, Australia. January 2011. 978-1-920682-95-8.
Contains the proceedings of the Twenty-Second Australasian Database Conference (ADC 2011), Perth, Australia, 17-20 January 2011.

Volume 116 - Information Security 2011
Edited by Colin Boyd, Queensland University of Technology, Australia and Josef Pieprzyk, Macquarie University, Australia. January 2011. 978-1-920682-96-5.
Contains the proceedings of the Ninth Australasian Information Security Conference (AISC 2011), Perth, Australia, 17-20 January 2011.

Volume 117 - User Interfaces 2011
Edited by Christof Lutteroth, University of Auckland, New Zealand and Haifeng Shen, Flinders University, Australia. January 2011. 978-1-920682-97-2.
Contains the proceedings of the Twelfth Australasian User Interface Conference (AUIC2011), Perth, Australia, 17-20 January 2011.

Volume 118 - Parallel and Distributed Computing 2011
Edited by Jinjun Chen, Swinburne University of Technology, Australia and Rajiv Ranjan, University of New South Wales, Australia. January 2011. 978-1-920682-98-9.
Contains the proceedings of the Ninth Australasian Symposium on Parallel and Distributed Computing (AusPDC 2011), Perth, Australia, 17-20 January 2011.

Volume 119 - Theory of Computing 2011
Edited by Alex Potanin, Victoria University of Wellington, New Zealand and Taso Viglas, University of Sydney, Australia. January 2011. 978-1-920682-99-6.
Contains the proceedings of the Seventeenth Computing: The Australasian Theory Symposium (CATS 2011), Perth, Australia, 17-20 January 2011.

Volume 120 - Health Informatics and Knowledge Management 2011
Edited by Kerryn Butler-Henderson, Curtin University, Australia and Tony Sahama, Qeensland University of Technology, Australia. January 2011. 978-1-921770-00-5.
Contains the proceedings of the Fifth Australasian Workshop on Health Informatics and Knowledge Management (HIKM 2011), Perth, Australia, 17-20 January 2011.

Volume 121 - Data Mining and Analytics 2011
Edited by Peter Vamplew, University of Ballarat, Australia, Andrew Stranieri, University of Ballarat, Australia, Kok–Leong Ong, Deakin University, Australia, Peter Christen, Australian National University, , Australia and Paul J. Kennedy, University of Technology, Sydney, Australia. December 2011. 978-1-921770-02-9.
Contains the proceedings of the Ninth Australasian Data Mining Conference (AusDM'11), Ballarat, Australia, 1–2 December 2011.

Volume 122 - Computer Science 2012
Edited by Mark Reynolds, The University of Western Australia, Australia and Bruce Thomas, University of South Australia. January 2012. 978-1-921770-03-6.
Contains the proceedings of the Thirty-Fifth Australasian Computer Science Conference (ACSC 2012), Melbourne, Australia, 30 January – 3 February 2012.

Volume 123 - Computing Education 2012
Edited by Michael de Raadt, Moodle Pty Ltd and Angela Carbone, Monash University, Australia. January 2012. 978-1-921770-04-3.
Contains the proceedings of the Fourteenth Australasian Computing Education Conference (ACE 2012), Melbourne, Australia, 30 January – 3 February 2012.

Volume 124 - Database Technologies 2012
Edited by Rui Zhang, The University of Melbourne, Australia and Yanchun Zhang, Victoria University, Australia. January 2012. 978-1-920682-95-8.
Contains the proceedings of the Twenty-Third Australasian Database Conference (ADC 2012), Melbourne, Australia, 30 January – 3 February 2012.

Volume 125 - Information Security 2012
Edited by Josef Pieprzyk, Macquarie University, Australia and Clark Thomborson, The University of Auckland, New Zealand. January 2012. 978-1-921770-06-7.
Contains the proceedings of the Tenth Australasian Information Security Conference (AISC 2012), Melbourne, Australia, 30 January – 3 February 2012.

Volume 126 - User Interfaces 2012
Edited by Haifeng Shen, Flinders University, Australia and Ross T. Smith, University of South Australia, Australia. January 2012. 978-1-921770-07-4.
Contains the proceedings of the Thirteenth Australasian User Interface Conference (AUIC2012), Melbourne, Australia, 30 January – 3 February 2012.

Volume 127 - Parallel and Distributed Computing 2012
Edited by Jinjun Chen, University of Technology, Sydney, Australia and Rajiv Ranjan, CSIRO ICT Centre, Australia. January 2012. 978-1-921770-08-1.
Contains the proceedings of the Tenth Australasian Symposium on Parallel and Distributed Computing (AusPDC 2012), Melbourne, Australia, 30 January – 3 February 2012.

Volume 128 - Theory of Computing 2012
Edited by Julián Mestre, University of Sydney, Australia. January 2012. 978-1-921770-09-8.
Contains the proceedings of the Eighteenth Computing: The Australasian Theory Symposium (CATS 2012), Melbourne, Australia, 30 January – 3 February 2012.

Volume 129 - Health Informatics and Knowledge Management 2012
Edited by Kerryn Butler-Henderson, Curtin University, Australia and Kathleen Gray, University of Melbourne, Australia. January 2012. 978-1-921770-10-4.
Contains the proceedings of the Fifth Australasian Workshop on Health Informatics and Knowledge Management (HIKM 2012), Melbourne, Australia, 30 January – 3 February 2012.

Volume 130 - Conceptual Modelling 2012
Edited by Aditya Ghose, University of Wollongong, Australia and Flavio Ferrarotti, Victoria University of Wellington, New Zealand. January 2012. 978-1-921770-11-1.
Contains the proceedings of the Eighth Asia-Pacific Conference on Conceptual Modelling (APCCM 2012), Melbourne, Australia, 31 January – 3 February 2012.

Volume 134 - Data Mining and Analytics 2012
Edited by Yanchang Zhao, Department of Immigration and Citizenship, Australia, Jiuyong Li, University of South Australia, Paul J. Kennedy, University of Technology, Sydney, Australia and Peter Christen, Australian National University, Australia. December 2012. 978-1-921770-14-2.
Contains the proceedings of the Tenth Australasian Data Mining Conference (AusDM'12), Sydney, Australia, 5–7 December 2012.