

CONFERENCES IN RESEARCH AND PRACTICE IN
INFORMATION TECHNOLOGY

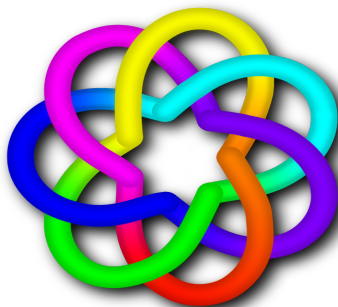
VOLUME 108

HEALTH INFORMATICS AND KNOWLEDGE MANAGEMENT 2010

AUSTRALIAN COMPUTER SCIENCE COMMUNICATIONS, VOLUME 32, NUMBER 7



AUSTRALIAN
COMPUTER
SOCIETY



 **CORE**
Computing Research & Education

HEALTH INFORMATICS AND KNOWLEDGE MANAGEMENT 2010

Proceedings of the Fourth Australasian Workshop on
Health Informatics and Knowledge Management
(HIKM 2010), Brisbane, Australia,
January 2010

Anthony Maeder and David Hansen, Eds.

Volume 108 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 2010. Proceedings of the Fourth Australasian Workshop on Health Informatics and Knowledge Management (HIKM 2010), Brisbane, Australia, January 2010

Conferences in Research and Practice in Information Technology, Volume 108.

Copyright ©2010, Australian Computer Society. Reproduction for academic, not-for-profit purposes permitted provided the copyright text at the foot of the first page of each paper is included.

Editors:

Anthony Maeder

School of Computing and Mathematics
University of Western Sydney
Locked Bag 1797
Penrith South DC, NSW 1797
Australia
Email: A.Maeder@uws.edu.au

David Hansen

CSIRO Australian e-Health Research Centre
Royal Brisbane and Women's Hospital
Herston, Queensland 4029
Australia
Email: david.hansen@csiro.au

Series Editors:

Vladimir Estivill-Castro, Griffith University, Queensland
Simeon J. Simoff, University of Western Sydney, NSW

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 108.
ISSN 1445-1336.
ISBN 978-1-920682-89-7.

Printed, January 2010 by UWS Press, Locked Bag 1797, South Penrith DC, NSW 1797, Australia
Document engineering by Susan Henley, University of Western Sydney
Cover Design by Matthew Brecknell, Queensland University of Technology
CD Production by FATS Digital, 318 Montague Road, West End QLD 4101, <http://www.fats.com.au/>

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

Table of Contents

Proceedings of the Fourth Australasian Workshop on Health Informatics and Knowledge Management (HIKM 2010), Brisbane, Australia, January 2010

Preface	vii
Programme Committee	viii
Organising Committee	ix
Welcome from the Organising Committee	x
CORE - Computing Research & Education	xi
ACSW Conferences and the Australian Computer Science Communications	xii
ACSW and HIKM 2010 Sponsors	xiv

Invited Paper

Health LinQ Using Routine Health Data to Link to Better Patient Care	3
<i>Steve Kisely</i>	

Contributed Papers

A Secure Architecture for Australias Index Based E-health Environment	7
<i>Vicky Liu, William Caelli, Jason Smith, Lauren May, Min Hui Lee, Zi Hao Ng, Jin Hong Foo and Weihao Li</i>	
Design an Automatic Appointment Aystem to Improve Patient Access to Primary Health Care	17
<i>Hongxiang Hu, Ping Yu and Jun Yan</i>	
Access to E-Health Information for the eNomad	23
<i>Anthony D. Stiller</i>	
A Multidimensional Temporal Abstractive Data Mining Framework	29
<i>Heidi Bjering and Carolyn McGregor</i>	
Automatic Sleep Stage Identification: Difficulties and Possible Solutions	39
<i>N. Sukhorukova, A. Stranieri, B. Ofoghi, P. Vamplew, M. Saleem, L. Ma, A. Ugon, J. Ugon, N. Muecke, H. Amiel, C. Philippe, A. BaniMustafa, S. Huda, M. Bertoli, P. Lévy and J. G. Ganascia</i>	
Visualising a State-wide Patient Data Collection: A Case Study to Expand the Audience for Health-care Data	45
<i>Wei Luo, Marcus Gallagher, Di O’Kane, Jason Connor, Mark Dooris, Col Roberts, Lachlan Mortimer and Janet Wiles</i>	
High Accuracy Information Retrieval and Information Extraction System for Electronic Clinical Notes	53
<i>Jon Patrick and Min Li</i>	
A Study on the Use of Search Engines for Answering Clinical Questions	61
<i>Andreea Tutos and Diego Molla</i>	

Customisable Query Resolution in Biology and Medicine	69
<i>Peter Ansell, James Hogan and Paul Roe</i>	
Assessing Text Characteristics of Electronic Discharge Summaries and their Implications for Patient Readability	77
<i>Mehnaz Adnan, Jim Warren and Martin Orr</i>	
Author Index	85

Preface

We are pleased to introduce papers from the 2010 Australasian Workshop on Health Informatics and Knowledge Management (HIKM 2010), delivered in Brisbane on Thursday 21 January 2010 as a special session of the 2010 Australasian Computer Science Week conferences (ACSW 2010).

The papers appearing here continue the HIKM trend in covering a broad scope of research topics in Health Informatics, ranging from theoretical considerations to practical applications. This year a greater number of papers than previous years deal with the design and implementation aspects of software solutions. We take this to indicate a degree of maturing of Computer Science research in Health Informatics in Australia and New Zealand, to a point where deployment and uptake of research results are becoming expected. Three areas of distinctive topics emerged as representative of our current research community interests this year: eHealth Systems, Text Mining / Searching, and Data Mining / Knowledge Discovery.

We have continued to enforce a high standard of acceptance for HIKM, in alignment with the expectations of ACSW and CRPIT. From a total of 21 submissions, all subjected to detailed peer review by 3 independent expert reviewers, the best 10 papers were selected for presentation and publication. Many of these papers report substantial bodies of work, that could equally well be published as journal papers or presented at leading international conferences. We hope those authors will seek such opportunities to report their work more comprehensively in the future. We also hope they will continue to develop new ideas in their areas of interest and concentration to offer to future HIKM workshops.

We congratulate the winners of the two best paper awards for HIKM 2010. The Best Overall Paper was “A Multidimensional Temporal Abstractive Data Mining Framework” by Heidi Bjering (University of Western Sydney) and Carolyn McGregor (University of Western Sydney and University of Ontario Institute of Technology). The Best Student Paper was “Assessing Text Characteristics of Electronic Discharge Summaries and their Implications for Patient Readability” by Mehnaz Adnan, Jim Warren and Martin Orr (University of Auckland).

Anthony Maeder

University of Western Sydney

David Hansen

Australian eHealth Research Centre / CSIRO ICT Centre

HIKM 2010 Programme Chairs

January 2010

Programme Committee

Chairs

Anthony Maeder, University of Western Sydney, Australia

David Hansen, Australian eHealth Research Centre / CSIRO ICT Centre, Australia

Members

Jeremy Barker, Queensland Facility for Advanced Bioinformatics, Australia

Peter Croll, Southern Cross University, Australia

Paul Frosdick, National eHealth Transition Authority, Australia

Marianne Hibbert, BioGrid Australia / University of Melbourne, Australia

Yogesha Kanganasam, Lions Eye Institute / University of Western Australia, Australia

Stephen Kisely, University of Queensland, Australia

John O'Brien, Queensland Health, Australia

Simon McBride, Australian eHealth Research Centre / CSIRO ICT Centre, Australia

Jon Patrick, University of Sydney, Australia

Vitali Sintchenko, New South Wales Health / University of New South Wales, Australia

Jeffrey Soar, University of Southern Queensland, Australia

Andrew Stranieri, University of Ballarat, Australia

Jim Warren, University of Auckland, Australia

Ping Yu, University of Wollongong, Australia

Organising Committee

Co-Chairs

Dr. Wayne Kelly
Prof. Mark Looi

Budget and Facilities

Mr. Malcolm Corney

Catering and Booklet

Dr. Diane Corney

Sponsorship and Web

Dr. Tony Sahama

Senior Advisors

Prof. Colin Fidge
Prof. Kerry Raymond

Finance and Travel

Ms. Therese Currell
Ms. Carol Richter

Registration

Mr. Matt Williams

DVD and Signage

Mr. Matthew Brecknell

Satchels and T-shirts

Ms. Donna Teague

Welcome from the Organising Committee

On behalf of the Australasian Computer Science Week 2010 (ACSW2010) Organising Committee, we welcome you to this year's event hosted by the Queensland University of Technology (QUT). Striving to be a "University for the Real World" our research and teaching has an applied emphasis. QUT is one of the largest producers of IT graduates in Australia with strong linkages with industry. Our courses and research span an extremely wide range of information technology, everything from traditional computer science, software engineering and information systems, to games and interactive entertainment.

We welcome delegates from over 21 countries, including Australia, New Zealand, USA, Finland, Italy, Japan, China, Brazil, Canada, Germany, Pakistan, Sweden, Austria, Bangladesh, Ireland, Norway, South Africa, Taiwan and Thailand. We trust you will enjoy both the experience of the ACSW 2010 event and also get to explore some of our beautiful city of Brisbane. At Brisbane's heart, beautifully restored sandstone buildings provide a delightful backdrop to the city's glass towers. The inner city clusters around the loops of the Brisbane River, connected to leafy, open-skied suburban communities by riverside bikeways. QUT's Garden's Point campus, the venue for ACSW 2010, is on the fringe of the city's botanical gardens and connected by the Goodwill Bridge to the Southbank tourist precinct.

ACSW2009 consists of the following conferences:

- Australasian Computer Science Conference (ACSC) (Chaired by Bernard Mans and Mark Reynolds)
- Australasian Computing Education Conference (ACE) (Chaired by Tony Clear and John Hamer)
- Australasian Database Conference (ADC) (ADC) (Chaired by Heng Tao Shen and Athman Bouguettaya)
- Australasian Information Security Conference (AISC) (Chaired by Colin Boyd and Willy Susilo)
- Australasian User Interface Conference (AUIC) (Chaired by Christof Lutteroth and Paul Calder)
- Australasian Symposium on Parallel and Distributed Computing (AusPDC) (Chaired by Jinjun Chen and Rajiv Ranjan)
- Australasian Workshop on Health Informatics and Knowledge Management (HIKM) (Chaired by Anthony Maeder and David Hansen)
- Computing: The Australasian Theory Symposium (CATS) (Chaired by Taso Viglas and Alex Potanin)
- Asia-Pacific Conference on Conceptual Modelling (APCCM) (Chaired by Sebastian Link and Aditya Ghose)
- Australasian Computing Doctoral Consortium (ACDC) (Chaired by David Pearce and Rachel Cardell-Oliver).

The nature of ACSW requires the co-operation of numerous people. We would like to thank all those who have worked to ensure the success of ACSW2010 including the Organising Committee, the Conference Chairs and Programme Committees, our sponsors, the keynote speakers and the delegates. Special thanks to Justin Zobel from CORE and Alex Potanin (co-chair of ACSW2009) for his extensive advice and assistance. If ACSW2010 is run even half as well as ACSW2009 in Wellington then we will have done well.

Dr Wayne Kelly and Professor Mark Looi

Queensland University of Technology

ACSW2010 Co-Chairs

January, 2010

CORE - Computing Research & Education

CORE welcomes all delegates to ACSW2010 in Brisbane. 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 ACSWin the mid 1990s now share the week with seven other events, which build on the diversity of the Australasian computing community.

In 2010, we have again chosen to feature a small number of plenary speakers from across the discipline: Andy Cockburn, Alon Halevy, and Stephen Kisely. I thank them for their contributions to ACSW2010. I also thank the keynote speakers invited to some of the individual conferences. The efforts of the conference chairs and their program committees have led to strong programs in all the conferences again, thanks. And thanks are particularly due to Wayne Kelly and his colleagues for organising what promises to be a strong event.

In Australia, 2009 saw, for the first time in some years, an increase in the number of students choosing to study IT, and a welcome if small number of new academic appointments. Also welcome is the news that university and research funding is set to rise from 2011-12. However, it continues to be the case that per-place funding for computer science students has fallen relative to that of other physical and mathematical sciences, and, while bodies such as the Australian Council of Deans of ICT seek ways to increase student interest in the area, more is needed to ensure the growth of our discipline.

During 2009, CORE continued to work on journal and conference rankings. A key aim is now to maintain the rankings, which are widely used overseas as well as in Australia. Management of the rankings is a challenging process that needs to balance competing special interests as well as addressing the interests of the community as a whole. ACSW2010 includes a forum on rankings to discuss this process. Also in 2009 CORE proposed a standard for the undergraduate Computer Science curriculum, with the intention that it be used for accreditation of degrees in computer science.

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 2009; in particular, I thank Gill Dobbie, Jenny Edwards, Alan Fekete, Tom Gedeon, Leon Sterling, and the members of the executive and of the curriculum and ranking committees.

Justin Zobel

President, CORE
January, 2010

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.

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 AUC.

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
AUIC	Australasian User Interface Conference
APCCM	Asia-Pacific Conference on Conceptual Modelling
AusPDC	Australasian Symposium on Parallel and Distributed Computing (replaces AusGrid)
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 2010 Sponsors

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



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



CEED,
www.corptech.com.au



Queensland University of Technology,
www.qut.edu.au



CSIRO ICT Centre,
www.csiro.au/org/ict.html



AUSTRALIAN
COMPUTER
SOCIETY

Australian Computer Society,
www.acs.org.au



SAP Research,
www.sap.com/about/company/research

INVITED PAPER

Health LinQ – Using routine health data to link to better patient care

Steve Kisely

Queensland Centre for Health Data Services

The University of Queensland

Brisbane, Qld, Australia

Steve Kisely [s.kisely@uq.edu.au]

Abstract

Australia is one of few countries that have comprehensive, high-quality, population data on many aspects of health and health care. Administrative health data have several advantages over community surveys, or data derived from individual clinical settings. They provide accessible and timely longitudinal data for an entire jurisdiction without the intrusion and cost of additional data collection, and so can be useful for both research and chronic disease surveillance. The Australian Government has provided \$30 million to establish the Population Health Research Network (PHRN), with representation from all States and Territories to facilitate population health research through data linkage. The Queensland Centre for Health Data Services (QCHDS) is the Queensland node and trades under Health LinQ. The centre involves 4 Queensland universities, Queensland Health and the Australian e-Health Research Centre. Functions of the QCHDS include: 1) Facilitating access to linked datasets involving Queensland Health and national data; 2) Developing methodologies for data linkage and analysis; 3) Capacity building around data linkage; 4) Participation in the

national coordination of data linkage and research. This talk describes the procedures for, and applications of, health data linkage. Researchers can either define cohorts for study within the administrative data or link them to their own data. Linkages are by probabilistic and deterministic matching using iterations of Link King and Febrl software packages, with robust protocols to preserve patient confidentiality. Access to the following data has been achieved: hospital morbidity, mortality, peri-natal, mental health data, as well as vital statistics. Privacy of linked data is protected by using a broken chain of information.

Data custodians provide demographic information without any health service data for linkage. A unique key is attached to the health service data before being handed through Health LinQ to the researcher. The researcher now has access to linked data without receiving any of the identifying demographic information. Current projects include preventable deaths from physical illness in psychiatric patients and an evaluation of whether the increased tax on alcopops has reduced alcohol-related health service use such as admissions to hospital or visits to emergency departments.

CONTRIBUTED PAPERS

A Secure Architecture for Australia's Index Based E-health Environment

Vicky Liu, William Caelli, Jason Smith, Lauren May,
Min Hui Lee, Zi Hao Ng, Jin Hong Foo and Weihao Li

Information Security Institute and Faculty of Science and Technology
Queensland University of Technology Australia
PO Box 2434 Brisbane, Queensland 4001, Australia

v.liu@qut.edu.au

Abstract

This paper proposes a security architecture for the basic cross indexing systems emerging as foundational structures in current health information systems. In these systems unique identifiers are issued to healthcare providers and consumers. In most cases, such numbering schemes are national in scope and must therefore necessarily be used via an indexing system to identify records contained in pre-existing local, regional or national health information systems. Most large scale electronic health record systems envisage that such correlation between national healthcare identifiers and pre-existing identifiers will be performed by some centrally administered cross referencing, or index system. This paper is concerned with the security architecture for such indexing servers and the manner in which they interface with pre-existing health systems (including both workstations and servers). The paper proposes two required structures to achieve the goal of a national scale, and secure exchange of electronic health information, including: (a) the employment of high trust computer systems to perform an indexing function, and (b) the development and deployment of an appropriate high trust interface module, a Healthcare Interface Processor (HIP), to be integrated into the connected workstations or servers of healthcare service providers. This proposed architecture is specifically oriented toward requirements identified in the Connectivity Architecture for Australia's e-health scheme as outlined by NEHTA and the national e-health strategy released by the Australian Health Ministers.

Keywords: architecture of health information systems, security for health information systems, health informatics, network security for health systems, trusted system, indexing based system for e-health regime, HL7.

1 Introduction

Undoubtedly, the adoption of e-health has much potential to improve healthcare delivery and performance (Goldschmidt 2005; AHM 2008). Anticipated improvements relate to better management and coordination of healthcare information and increased quality and safety of healthcare delivery. On the other

hand, a security violation in healthcare records, such as an unauthorised disclosure or unauthorised alteration of individual health information, can significantly undermine both healthcare providers' and consumers' confidence and trust in the e-health system. A crisis in confidence in national e-health systems would seriously degrade the realisation of potential benefits.

Evidence from the NEHTA's Report on Feedback Individual Electronic Health Record (NEHTA 2008c) suggests that numerous healthcare consumers and providers embrace the adoption of national individual electronic health records because of the potential benefits. There are a number of consumers, however, who are reluctant to embrace e-health because of privacy concerns. Obviously, the security and privacy protection of information is critical to the successful implementation of any e-health initiative. NEHTA, therefore, rightly places security and privacy protection at the centre of its e-health approach.

In order to address the requirements for enabling a secure national e-health environment, we propose a security architecture based around the current strategic directions from the Australian Government's National E-Health Strategy (AHM 2008) and Connectivity Architecture (NEHTA 2008b) proposed by NEHTA, both recently released in December 2008.

This proposed architecture defines a model to support secure communications between healthcare providers and the Index System in the national e-health environment, which some other approaches fail to address. We draw on important lessons from the Internet's Domain Name System (DNS) for the development and deployment of the national healthcare Index System. Our approach embraces the hierarchical and distributed nature of DNS and defines the required components for a secure architecture for Australia's national e-health scheme. This proposed architecture employs a high trust computer platform to perform indexing functions and a high trust interface module as the application proxy to connect to the healthcare Index System and other healthcare service providers.

2 Paper Structure

This paper begins with a summary of the benefits associated with increased adoption of e-health; however, risks to privacy in such e-health systems must be addressed. Addressing the security appropriately is considered as key to success of the e-health implementation. Section 3 defines the paper's scope and details our assumptions in the context of the Australian

national e-health environment. Section 4 investigates three representative e-health initiatives resembling the approach being adopted in Australia. Section 5 reasons the lesson we can learn from Internet's DNS to design the national e-health Index System. The authors' proposal for a secure connectivity architecture with the required structures is described in Section 6. Section 7 illustrates a request for a specific patient's health records via the Index System with a set of information flows. The analysis of this work is incorporated in Section 8. Finally, the conclusion is drawn and future direction for work is outlined in Section 9.

3 Scope and Assumptions

The Australian National E-health Strategy (AHM 2008) defines the basic building blocks for a national e-health system including: (1) the implementation of the healthcare identifier (HI) scheme for healthcare consumers and providers, (2) the establishment of standards for the consistent collection and exchange of health information, (3) the establishment of rules and protocols for secure healthcare information exchange, and (4) the implementation of underlying physical computing and network infrastructure. We propose a secure architecture to address the protection of clinical information exchange in a reliable and secure manner. This proposed architecture is specifically concerned with the secure architecture design and development to facilitate interactions between healthcare providers, healthcare organisations and the national Index System rather than focusing on healthcare consumers accessing healthcare information.

It is anticipated that the national HI scheme will be established by mid 2010 (AHM 2009). This paper assumes that an adequate national legislative framework will be established to support the management and operation of the healthcare identifier scheme (NHHRC 2009) to enable a national e-health implementation by July 2010. Presumably, the National Authentication Service for Health (NASH) becomes available for Public Key Infrastructure (PKI) services to support digital signing and data encryption in the national e-health environment. It is also assumed that the National Broadband Network (NBN) infrastructure will be constructed for electronically enabling access and transfer of health information nationally.

In the context of this paper, a service requester refers to the entity that uses a service provided by another entity. A service provider is an entity that offers a service used by another entity. A service provider can be a healthcare provider, healthcare organisation or organisation commissioned to provide services for healthcare providers or healthcare organisations.

4 Related Work

While most nations would appear to have some e-health initiatives at some stage of investigation or implementation, this section focuses on three national e-health architectures resembling the approach being adopted in Australia.

4.1 Dutch National E-health Strategy

The Dutch e-health infrastructure is constructed by the National IT Institute for Healthcare in the Netherlands

(NICTIZ)¹. The Dutch national e-health approach uses the National Healthcare Information Hub, National Switch Point (Landelijk Schakelpunt or LSP) to enable the exchange of healthcare information. There is no clinical information stored at the LSP. The clinical data details reside at local health information systems. The Dutch national index system, LSP, includes services such as identification and authentication, authorisation, addressing, logging and standardization of messages services (The Dutch Ministry of Health 2007)

The LSP links healthcare providers' information systems together to enable the electronic exchange of health information nationally. The Dutch national e-health network connectivity architecture requires the healthcare partitioners' health information system to comply with the security requirements for a "Qualified Health Information System to be allowed to connect to the LSP via a qualified commercial service provider. Such IT service providers are commissioned to provide secure communications between healthcare information systems and the LSP" (Spronk 2008).

While the healthcare provider requests specific patient information which is located in other healthcare information systems, all queries are relayed via the LSP. The healthcare service provider responds to the LSP. Namely, the LSP aggregates the requested health data from the health service providers and then routes the health data to the requester. There is no direct communication between the healthcare service providing system and requesting system. The LSP also logs which healthcare practitioners have accessed patient data for accountability (The Dutch Ministry of Health 2007).

The Dutch national index system, LSP, is the central coordination point for exchange health information, including authentication, authorisation, routing and logging. Such an implementation model may appear suitable for a small scale of national e-health structure. Implementation of this model in a geographically large country will produce more network traffic, possibly creating performance bottlenecks; it is particularly prone to a single point of failure weakness.

4.2 National Health Service (NHS) in England

The National Health Service (NHS) in England implements the National Programme for IT (NpIT) to deliver the central electronic healthcare record system. This central system is known as Spine. Spine provides national e-health services in England including:

- The Personal Demographics Service (PDS), which stores patients' demographic information including unique patient identifiers - NHS Numbers;
- Spine Directory Services (SDS), which provides directory services for registered healthcare providers and organisations;
- National Care Record (NCR), which contains clinical information summaries as well as the location of the detailed healthcare information;

¹ NICTIZ is Dutch national e-health coordination point and knowledge centre. The related information is available at <http://www.nictiz.nl/>, accessed 28/08/2009.

- Legitimate Relationship Service (LRS), which is an authorisation logic containing details of relationships between healthcare professionals and patients and patient preferences on information accessing; and
- Transaction and Messaging Spine (TMS), which provides routing for querying and responding to clinical messages via the NCR (Spronk 2007).

The English national e-health services include identification and authentication, authorisation logic, clinical summary information, directory services and routing. This programme is implemented in England, while Wales is running another national programme. The separate provisions of national e-health systems need to be made interoperable for information traversing across national borders.

4.3 USA Health Information Exchange (HIE)

USA National Institute for Standards and Technology (NIST) recently released a document entitled Draft Security Architecture Design Process for Health Information Exchanges (HIEs) (Scholl et al. 2009) to provide guidance for the development of a security architecture particularly for the exchange of healthcare information. The HIE security architecture design process includes five layers to construct a security architecture for healthcare information exchange. The five layers include: (a) policies for overall legal requirements to protect healthcare information access, (b) services and mechanisms to meet policy requirements, (c) operational specifications for the business processes, (d) definitions of technical constructs and relationships to implement enabling processes, and (e) provisions for technical solutions and data standards for implementing the architecture.

USA health information exchange architecture is based upon a hierarchical structure. Namely, it consists of a National Federation Health Information Exchange (HIE), Multi-Regional Federation HIEs, and Regional HIEs. The National Federation HIE, national federated technical architecture, connects a number of Multi-Regional Federation HIEs, involving multiple states jurisdictions. Multi-Regional Federation HIEs connect multiple regional HIEs. Regional HIEs can consist of two or more independent healthcare providers to share healthcare information. The participating healthcare providers set up their own trust agreement to define security and privacy requirements for the exchange of healthcare information (Scholl, Stine, Lin and Steinberg 2009).

The Identity Federation Service provides identification and authentication services. The entity can be authenticated via the Identity Federation Service or its home organisation's authentication service to support single sign on for accessing the HIE services. The privilege management is performed by service providers locally (Scholl, Stine, Lin and Steinberg 2009).

The USA approach is different from the Dutch and English national e-health architectures. In a large nation like the USA, the distributed national e-health scheme seems suitable for scalability. USA e-health architecture is similar to the context of the DNS hierarchical model. This

type of approach can mitigate the network traffic and performance bottleneck on the centralised e-health system.

5 Lesson Learnt from the Internet's Domain Name System (DNS)

The Internet's "*Domain Name System (DNS)*" has become a critical part of the Internet and of the "*World Wide Web (WWW)*" in particular. Without its services many current information systems and services provided over the Internet would not function. Indeed, as Web-based applications rapidly become the "norm", particularly in the public sector but also in the private sector, the resilience and high speed performance of the DNS have become mandatory requirements. The use of Web-based structures has been nominated as the basic functional structure of the Australia Federal e-health, NEHTA scheme. The DNS structure, determined some 25 years ago, is based around a globally distributed, hierarchical database architecture that relies upon replication for resilience and caching for performance. However, it has been realised that the basic DNS scheme is insecure, in the sense that both confidentiality and integrity, including authenticity and authorisation, were not part of the overall design during the original design and development time of the early to mid 1980s.

"Robustness and adequate performance are achieved through replication and caching" (Liu and Albitz 2006). Essentially, the client-server model chosen, via use of client "resolvers" and then "name-servers", has been proven over time and is the model suggested in this architecture. The hierarchical nature of the DNS structure again appears suitable given that the Australian system must cater for a federated national structure with roles for the various State level participants. The "*ccTLD*" or "*country top level domain*" coupled with a "*2nd level*" structure appears to offer suitable benefits in organisation and management as well as the necessary backup resilience that is required in the overall scheme.

The appropriate security arrangements, the "*Transaction Signatures (TSIG)*" structure based on a single-key cryptographic system again helps in this regard in relation to the secure synchronisation of actual DNS nameserver systems themselves. As Liu and Albitz (2006) state, "*TSIG uses shared secrets and a one-way hash function to authenticate DNS messages, particularly responses and updates.*" Similar schemes exist for confidentiality, integrity and authenticity services in data networks in the banking and finance sector.

As mentioned above, the original DNS structure did not consider matters of confidentiality and integrity. At the same time, the TSIG scheme is not scalable to any real dimension as nameservers correspond with an arbitrary set of other nameservers. The "*DNS Security Extensions (DNSSEC)*" (Arends et al. 2005a; Arends et al. 2005b; Arends et al. 2005c), through use of "Public Key Cryptography", enable DNS "zones" to "digitally sign" the necessary nameserver tables so that, on distribution, such tables can be checked for authenticity and integrity by the receiver. The addition of appropriate DNSSEC records to the overall database structure provides a useful model that may be incorporated into the proposed architecture that is the subject of this paper.

In summary, the overall DNS experience, and the structure of the DNSSEC security extensions provide a most suitable model for incorporation, in modified form, into the healthcare index architecture proposed. The DNSSEC structure assists in combating known attacks on the Internet system through such techniques as “cache poisoning”, “traffic diversion”, “man-in-the-middle attacks” and so on. At the same time, however, the basic

index systems, like the Internet’s DNS nameserver systems, must be installed and managed on basic computer systems, including the necessary operating systems (OS) that are sufficiently secure for the purpose. The immediate use of DNSSEC style structures is seen as essential given that many aspects of the proposed e-health record infrastructure will reside on the general purpose Internet.

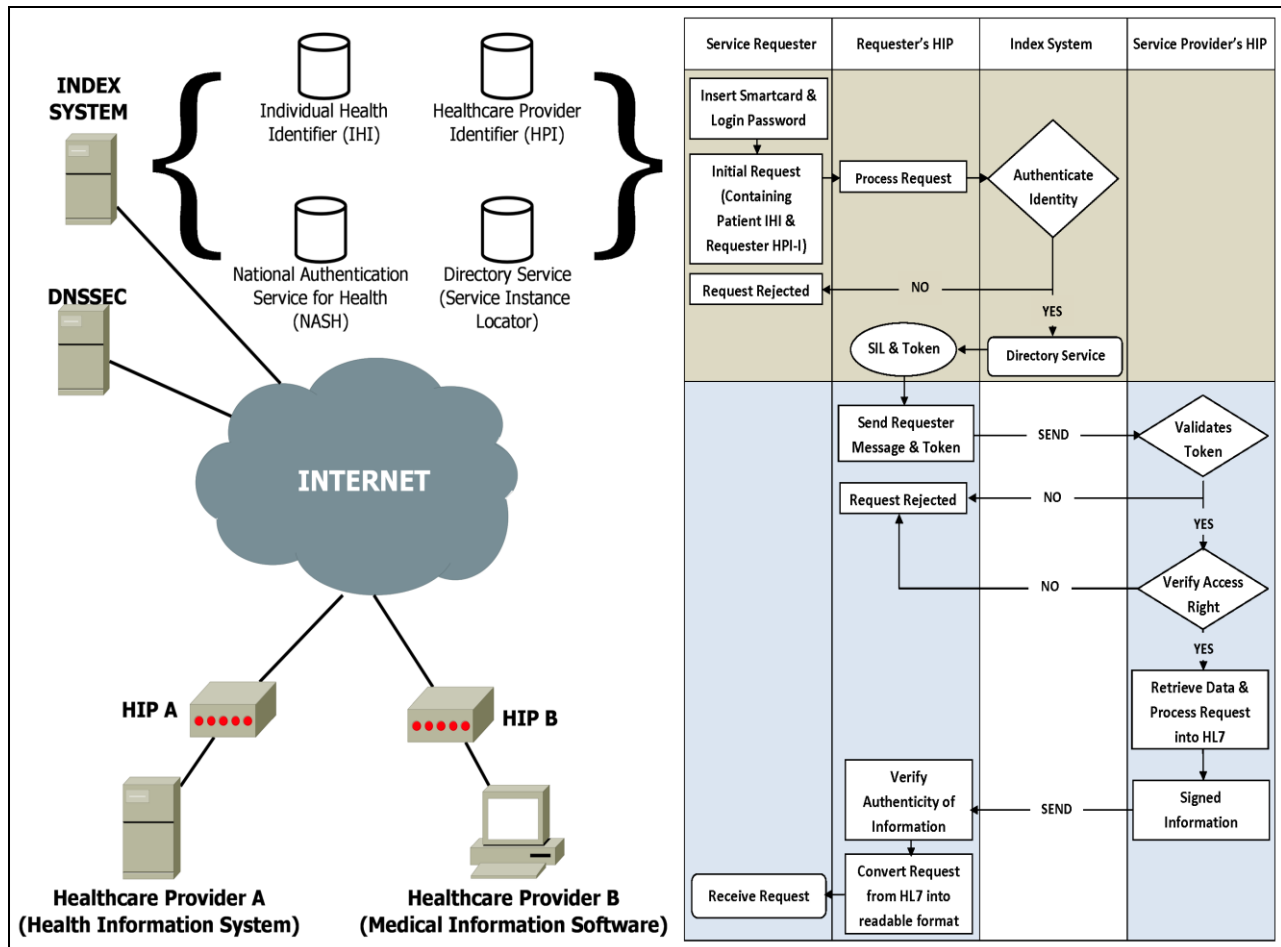


Figure 1 Proposed Architecture Overview and Key Information Flows

6 Our Approach

Generally, health information is stored over a number of different health information systems. A national index system must be available for the provision of directory services to determine the distributed locations of the source systems holding the related health records. Our proposal addresses this need by defining a model to support secure communications between healthcare providers and the Index System in the national e-health environment as shown in Figure 1. This proposed architecture is based on the broad architecture of the Australian Government’s National E-health Strategy (AHM 2008) and NEHTA’s Connectivity Architecture (NEHTA 2008b), both released in December 2008.

Our proposed architecture defines the required constructs to share and transfer healthcare information securely between healthcare providers and the authorised national Index System. This architecture proposes that the Index System should be built on a high trust computer

platform as well as mandating that the participating healthcare provider’s need to adopt a high trust interface module - HIP as the application proxy to link to the Index System and other health information systems. Additionally, the authors argue that a fundamental security issue, that of name resolution, must be addressed prior to the interactions between the healthcare providers and national Index System. This paper, therefore, proposes a trusted architecture not only providing the indexing service but also incorporating a trusted name resolution scheme for the enforcement of communicating to the authorised Index System.

Since the Index System is itself a critical application under any operating system, that Index System must be protected from even internal threats through the use of modern “flexible mandatory access control (FMAC)” structures. Under such an operating system, and as distinct from the less secure “discretionary access control (DAC)” systems, even a systems manager may not have permission to access the health record data. In simple terms, in these

systems there is no “super-user” capable of obtaining access to all system resources at any time. If an individual nameserver system is “captured”, propagation of exposure will not extend beyond the compromised application itself, a vital concern in any e-health record indexing structure. Such systems exist and are commercially available, e.g. the “Secure LINUX (SELinux)” systems, “Solaris/SE” system, etc. The proposed “HIP” structure would make use of such security enforcement to provide the necessary protection levels.

6.1 Index System (IS)

The authors argue that the load of the national Index System should be relatively lightweight to perform e-health indexing services efficiently. This can mitigate the Index System explosion and traffic bottleneck risks. Such an approach is favourable in a geographically large country such as Australia. To maximise the efficiency of the indexing services, the proposed Index System does not provide network connectivity services, messaging translation, addressing and routing functions and extensive logging of all message access. These services can be performed at the level of the local health information systems via the HIP, which is detailed in Section 7.2. The access control and authorisation process is best performed close to where the source system is, as each healthcare service provider might implement the service differently based on its own health information system access requirements. NEHTA (NEHTA 2008b) also states that there are no centralized network provisions to handle peer-to-peer communications; each service must manage its own interface to the network.

The Index System will be a centralised facility run at a national level. It is envisioned that the directory service is devised in the context of a DNS, which uses hierarchical distributed database architecture.

Our proposed national Index System performs common and fundamental functionalities including:

- Identification and authentication, and
- Directory services.

6.1.1 Identification and Authentication Services

The national Healthcare Identifiers Service (HI Service) is indeed one of the building blocks for the national e-health infrastructure. The national HI scheme for identification services must be deployed prior to the implementation of the national e-health system. The HI Service will provide accurate identification of individuals and healthcare providers in the national e-health environment.

Individuals receiving healthcare services will be assigned an Individual Healthcare Identifier (IHI). All authorised Healthcare providers will receive a Healthcare Provider Identifier – Individual (HPI-I). Healthcare centres and organisations in Australia will be provided with a Healthcare Provider Identifier – Organisation (HPI-O). To be eligible to query the HI Service, a requesting entity must be nominated by a healthcare organisation and have an HPI-I associated with an HPI-O. The IHI Service will allow authenticated healthcare providers to lookup a specific IHI. The HPI Service of the Index System will provide lookup services to navigate the

locations of healthcare providers to facilitate communication and the exchange of healthcare information (AHM 2009).

National Authentication Service for Health (NASH) is designed by NEHTA to provide PKI authentication services. NASH will issue digital certificates and tokens for registered and certified healthcare providers and organisations (AHM 2009).

6.1.2 Directory Services

The Directory Service is one of the fundamental services in national e-health infrastructure. Since healthcare data are located at various places, directory services are used to identify and locate the available information. The Directory Service in the Index System provides a mechanism for obtaining the necessary information for invoking a service. This information contains the network location of the service, the digital certificate required to use it and other information required to invoke the service. It is envisaged this will be specified in Web Services Description Language² (WSDL) format, which equates to Service Instance Locator (SIL) (NEHTA 2008d) functionalities outlined by NETHA.

6.1.3 Operation of the Directory Services

Based upon NEHTA’s definitions (NEHTA 2008a) on concepts and patterns for implementing services, the service patterns can be divided into two broad categories: synchronous and asynchronous services. A synchronous service occurs in direct response to a request. An asynchronous service has no relationship between the events. For example, to request a specific individual’s health records is a synchronous service. To send out a discharge summary report to a healthcare provider is an asynchronous service.

With a synchronous service, when interacting with the directory service the requesting entity will provide proof of their identity (HPI-O) and the IHI associated with the records they are requesting. Once the requester has been authenticated by the Index Server, it will respond with the following: (a) a signed token attesting to the identity of the requester ($\{\text{token}\}\text{Sign}_{\text{IS_PrivKey}}$) and (b) a list of service instances containing health records for the person identified by the IHI ($\text{Service_Instance_1}, \dots, \text{Service_Instance_N}$).

The entire response is signed so that the requester can be assured that it is a legitimate response from an authorised Index System and that any alterations to the response will be detectable. The response is also encrypted under a key known by the requester ($\{\dots\}\text{Encrypt}_{\text{HPI-O_PubKey}}$), in order that the confidentiality of both the requester and the individual identified by the IHI is maintained.

The token is signed independently of the entire response in order that it can be reused with requests to each service instance. The full response is depicted in Figure 2.

² WSDL is used for describing how to access the network services in XML format. More detail is available at http://www.w3.org/TR/wsdl#_introduction accessed 30/08/2009.

```
{ {token}SignIS_PrivKey,Service_Instance_1,...,  
Service_Instance_N}EncryptHPI-O_PubKey
```

Figure 2: Service Instance Response Message Format

The service instance information contained in the response identifies the target system location and information necessary for securely invoking that service. This may include, but will not be limited to the credentials / certificates required to access the service. The signed token provided in the Index System response may be the only credential required, in which case the effort expended by the Index System in authenticating the requester is reused. It is, however, conceivable that additional authentication may be required by a given service instance. For example, the requester may need to prove that they are a member of a given practice or college of medical practitioners.

With an asynchronous service, such as when a discharge summary message needs to be sent to the patient's primary healthcare provider, the healthcare provider issuing the summary queries the Index System for the primary healthcare provider's HPI, location and the digital certificate and then signs and encrypts the discharge message prior to transmission.

6.2 Healthcare Interface Processor (HIP) – Proxy Service

Our design philosophy of HIP draws on principles used in the Interface Message Processor (IMP) of the Advanced Research Projects Agency Network (ARPANET). Each site uses an IMP to connect to the ARPANET network in order to isolate the potential hostile system connecting the ARPANET network. Our design rationale underlying HIP is to provide a secured communication channel for an untrusted health information system connected to the Index System as well as for health information exchange between healthcare providers. Wherever a connection to the national indexing system is required, a HIP facility has to exist. The design goal for HIP is to make it as a “plug and operate” facility, which is easy and simple to use for healthcare providers as well as with characteristics of high security, reliability, efficiency and resilience. Such a design would be very beneficial and useful particularly for healthcare providers.

HIP contains its own on-board crypto-processor based on a trusted computing based module to store cryptographic keys. Any information system depends, therefore, upon a trusted base for safe and reliable operation, commonly referred to as a “trusted computing-base”. Without a trusted computing base any system is subject to compromise. For this reason HIP aims to run on top of trusted hardware, firmware and operating system. HIP, a self-contained unit configured with an IP address, is capable of running Web services. HIP carries out its works from layer 1 to 7 of the seven-layer OSI model.

It is envisaged that HIP achieves provisions of security services and mechanisms based upon the security and management concepts of the OSI IS7498-2, including:

- To establish a **trusted path** to connect to the authorised Index System,

- To provide **peer-entity authentication** between healthcare providers and national Index System,
- To facilitate **secure healthcare information exchange** in transit,
- To provide **data protection** with appropriate **access control** mechanisms,
- To provide **interoperability** to enable healthcare information exchange between disparate healthcare systems with varying security mechanisms,
- To support **accountability** when healthcare information has been accessed, and
- To provide **operation flexibility** with “emergency override” and **capacity flexibility** for various scales of healthcare organizations.

6.2.1 Trusted Path Establishment

In response to the recent increase in DNS cache poisoning and traffic diversion attacks, we propose that the first step is to perform the enforcement of communicating to the authorised Index System prior to the interactions between the service requesting entity and the Index System. To achieve this, from a technical underlying process, HIP should be pre-configured to contact a DNSSEC capable server to perform a trusted name resolution in order to defend against false DNS data and assure that connections are only established with the legitimate Index System.

6.2.2 Peer-Entity Authentication

Many proposals are only concerned with the authenticity of the requesting entity (i.e. one-way authentication) but fail to address the importance of two-way authentication. Our proposed architecture provides a mutual peer-entity authentication service complying with the ISO 7489-2. To authenticate the authenticity of the Index System, the service requesting entity must validate the certificate of the Index System. Once the authenticity of the national Index System is assured, the Index System authenticates the identity of the healthcare service requesting entity. In this sense, the authentication service of the Index System acts as a notarization mechanism in line with the philosophy of peer-entity authentication stated in ISO IS7498-2.

6.2.3 Secured Communication Channel for Health Information Exchange

The healthcare provider's computer may have its security compromised. HIP, a hardened and qualified facility, acts as a proxy server establishing a secured communication channel connecting to the Index System and bringing isolation from the untrusted computer.

HIP will be assigned a standard unique identifier (i.e. HPI-O) and be issued an asymmetric key pair for digitally signing and encrypting to achieve integrity and confidentiality goals. HIP contains its own on-board crypto-processor, thus it can facilitate the secure exchange of health information. In addition, HIP is built on the Trusted Platform Module (TPM) that is used to store cryptographic keys.

6.2.4 Provision of Data Protection

As various healthcare organisations may have their own specific access authorisation requirements and processes, access authorisation is best performed where the resource system is located. Once the requesting entity's identity is authenticated, the request of particular healthcare information is presented to the target service provider. The HIP of the target service provider will provide the verified identity and the profile of the requester to the authorisation logic unit to perform access decision making. The authorisation decision depends upon the requesting entity's profile and defined privilege management policy. The implementation of the authorisation logic unit is based on the "Sensitivity Label" function outlined by NEHTA (NEHTA 2008c).

6.2.5 Interoperability Platform

NEHTA³ is responsible for selecting electronic messaging standards in Australia's health sector. It has endorsed Health Level 7 (HL7)⁴ as the national standard for the electronic exchange of health information. HIP provides an interoperability platform by incorporating an HL7 Interface Engine and Message Mapping Sets conforming to the HL7 v3 Message Standards for healthcare information exchange. HIP also incorporates an HL7 Interface Engine and Message Mapping Sets for messaging Interoperability.

HL7 Interface Engine

Any non-HL7-compliant data contents are translated into the HL7 standard format (XML-based data structure) by the HL7 Interface Engine prior to information transmission. The HL7 Interface Engine contains a set of mapping algorithms to map data contents with an appropriate HL7 Message Template to generate an HL7 message.

Message Mapping Sets

The Message Mapping Sets contain a repository of HL7 Message Templates for various clinical and administrative messages. Each set provides one HL7 Message Template to serve for one clinical or administrative message. Message Mapping Sets will be designed and developed to meet the current healthcare service needs and will be imported into HIP. The HL7 Message Template guides and directs data contents to form an HL7 message.

HL7 Clinical Document Architecture (CDA)

HL7 Clinical Document Architecture (CDA) provides a framework for clinical document exchange. HIP imports the HL7 message into a CDA document. This CDA document is also associated with an appropriate stylesheet. The CDA document and the stylesheet will be sent to the requesting entity through Web services. The requesting

entity renders the received document with the stylesheet in a human-readable form with a Web browser.

6.2.6 Privacy Accountability

Audit trail mechanisms can be used to deter unauthorised access to data to improve privacy accountability. To enforce privacy accountability, HIP could be configured to automatically trigger an audit trail event particularly when data is being accessed.

6.2.7 Operation and Capacity Flexibility

HIP aims to accommodate emergency override whereby any delays that may potentially occur through authentication and authorisation may be overridden. This is particularly relevant in the case of defined emergency including pandemic circumstances. HIP is designed to provide an emergency override provision called "Hit-the-HIP" for ease of operation.

The HIP architecture is flexible enough to cater for interfacing at various levels. Examples of healthcare organisational structures include a one-person general practice clinic, and small or medium clinics to large hospitals. It is proposed that a number of design variations for the HIP facilities, depending on the healthcare structure, may include:

- One-person healthcare practitioner,
- Smaller healthcare practitioners,
- Hospital administration, and
- Regional hospital administration

7 Envisioned Key Information Flows

This section uses a scenario to illustrate the key information flows (see Figure 1) based on the proposed architecture described in Section 6. While a requester needs to inquire about a specific patient's health information, the key information flows of the interactions between the requester, Index System and service provider are illustrated in the following steps. Note that all request and response messages prior to transmission are signed and encrypted for confidentiality, authentication and message integrity reasons.

1. Peer-Entity Authentication Process

- 1.1. Prior to peer-entity authentication, to ensure the secure resolution, the service requester's HIP obtains the address of the Index System from the DNSSEC system which is pre-configured in the HIP.
- 1.2. The service requester initiates a connection with the Index System via the service requester's HIP. To ensure the authenticity of the Index System, the service requester's HIP validates the certificate of the Index System.
- 1.3. To ensure the identity of the service requester, the service requester logs into the Index System with his/her smart card containing their credentials.

2. **Health Record Enquiry Process** The service request, containing the patient's IHI and requester's HPI-I, is sent to the Directory Services of the Index System

³NEHTA Sets Direction for Electronic Messaging in Health is available at <http://www.nehta.gov.au/nehta-news/423-nehta-sets-direction-for-electronic-messaging-in-health>, accessed 19/08/2009

⁴ Health Level 7, an American National Standards Institute accredited standard, has been developed to enable disparate healthcare applications to exchange key sets of clinical and administrative data.

to inquire which health providers hold the health records of the specific patient.

- 2.2. The Directory Services of the Index System responds with a token and a list of the service instance information for service invocation to the requesting entity. This token indicates the requester identity assertion to enable single sign on for service invocation.
- 2.3. The requester verifies the received information and then contacts each target service provider for service invocation. The requester sends the request including the token with other necessary information to invoke the service.

3. Verification and Authorization Evaluation Process

- 3.1. Each target service provider validates the request message containing the token and other necessary information for service invocation.
- 3.2. In turn, the request is passed to the authorization logic to make an access authorisation decision based on the service requester's profile indicated in the ticket and any additional authorisation attributes which are mutually agreed by the policy.

4. Provision of Requested Health Record Process

- 4.1. If the access is granted, the service provider extracts the health record from the data source.
- 4.2. The service provider processes the requested health record into the HL7 message format.
- 4.3. The target service provider sends the signed and encrypted information to the requester.
- 4.4. The service provider records the information access for auditing purposes.

5. Reception of Requested Health Record Process

- 5.1. The requested information arrives at the service requester's HIP.
- 5.2. The service requester's HIP verifies the information arrived and then extracts the requested information which is in HL7 message format.
- 5.3. The message must be presented in a human readable format. The representation of HL7 message is rendered and displayed to the requester.

8 Analysis

A first point of contact in any Index System must be itself verified for authenticity and integrity. In Internet terms the client system must be sure that it is connected to the correct Index System and not to some fraudulent system or via some intermediate node point capable of monitoring all traffic. The suggestion for use of a DNSSEC style structure in the overall architecture is seen as a minimum requirement for overall trust in the system.

In turn, this implies that all systems used in the creation and operation of a "centralised" Index System must be security verified in line with accepted international standards. The main such standard is the "Common

Criteria (CC)" set,⁵ under international standard IS-15408, accepted by many nations⁶ as the base for evaluation of the security stance of any system. Isolation of critical security functions into verifiable hardware and software structures capable of CC "protection profile" definition is envisaged along with the acceptance of a requirement for an associated evaluation at a minimum of an evaluation level of "EAL5". This would apply to the HIP. It should also be a requirement that the USA's "FIPS 140-2", the Federal Information Processing Standard, be used for the security verification of cryptographic functions, in line with accepted industry practice.

Unlike previous structures, the HIP may operate at all seven layers of the OSI model and, indeed, be seen as a "proxy" for Internet interaction. For example, the functionalities of HIP include:

- Routing control functions operating at layer 3, the "network layer" of the OSI model;
- HL7 interpreter functions working at the "presentation layer", layer 6;
- Web service operations carried out at layer 7 of the OSI model, the "application layer"; and
- The encryption/decryption mechanisms at layers 2, 3 and 4 of the OSI model.

The proposed structure is cognisant of NEHTA's architectural designs for the overall national health record index scheme proposed for Australia. Moreover, the main aim of the HIP concept is to simplify overall security control and management of the e-health environment from the point of view of those health professionals and practitioners who will be using the system in the future. The whole HIP architecture is seen as being able to be explained and understood by health professionals and related people who are not ICT experts. Moreover, the HIP and its security should be transparent to them in normal operation. The goal of the proposed system is to make the HIP understandable and essentially transparent to users so that health practitioners can focus on their primary functions to deliver quality healthcare service. In this regard, control and management of the overall system is vested in appropriate information and network systems professionals, not the end users or health partitioners themselves.

9 Conclusion and Future Work

This paper proposes three distinct suggestions on the architecture set:

(1) Trusted domain name services are a critical element in the overall trusted architecture of any indexing based healthcare systems to combat name resolution cache poisoning and traffic diversion attacks;

(2) A trusted architecture for the Index System which provides the critical solution to determine the locations of

⁵ The Common Criteria Portal is available at – <http://www.commoncriteriaportal.org>, accessed 7/09/2009.

⁶ More information about the Mutual Recognition and the Common Criteria Recognition Arrangement is available at http://www.dsd.gov.au/infosec/evaluation_services/aisep_pages/aisep_partners.html accessed 07/09/2009.

distributed health records. This Index System plays a vital role in the national e-health scheme for identification and authentication and directory services. The Index System, therefore, must be a high trust system running on a trusted platform; and

(3) HIP plays a vital role as a proxy server connecting to the national Index System as well as linking to untrusted health information systems. The proposed “HIP” structure will be built on top of a trusted platform. This makes use of available security enforcement to provide the necessary protection levels.

We envisage that the HIP would be subject to security functionalities and evaluation at the minimum requirements of EAL5 under the Common Criteria/ISO15408⁷, in which Australia participates under the Common Criteria Recognition Agreement (CCRA)⁸.

There are a number of proposals to maintain summarised healthcare records within the overall index system/switching system (Spronk 2007; Spronk 2008). A summary of healthcare records in Australia is called an Individual Electronic Health Record (IEHR) (AHM 2008). Our architecture can accommodate IEHR: for example an IEHR database added in Figure 1. This proposal needs to be further examined in light of prior experience in other sectors, such as banking and finance industries. While it would appear possible to maintain IEHRs within the national Index System, practicality may indicate that, in line with the DNS system discussed in this paper and in the banking sector, IEHRs may be best implemented at the point where such aggregation is most feasible. In Australia, this would indicate, in light with the DNS system, a second level Index System at the state level which would also contain IEHRs. Under investigation in the overall project is the feasibility of aggregating IEHRs on demand for the use of point access.

Point of Sale (EFTPOS) is a model that can be used to develop HIPs. Part of our future work is to design a prototype to demonstrate this. This paper forms a foundation for the creation of such a prototype/demonstrator high trusted Index System coupled with a prototype HIP. This will form a base of future requests for research funding. HIP will be developed as proof-of-concept which may be used when tendering for supply and installation. It is suggested that the government will issue the development and testing of HIP which involves the production of 5-6 laboratory prototypes and 50-100 production prototypes. Upon the successful bidder testing, this proposal suggests that the government would issue tenders for the production and installation of HIP. This is based upon the successful experience in the financial sector, in particular, the successful structure and deployment of Australian Electronic Funds Transfer at EFTPOS systems over the last 25 years.

Although this paper concentrates on the Australian national e-health environment from a security perspective, our conclusions could be equally applied to any distributed, indexed based healthcare information systems involving cross referencing of disparate health data collections or repositories.

10 References

- AHM (2008): National E-Health Strategy Summary. [http://www.health.gov.au/internet/main/publishing.nsf/Content/604CF066BE48789DCA25751D000C15C7/\\$File/Summary%20of%20National%20E-Health%20Strategy-final051208.pdf](http://www.health.gov.au/internet/main/publishing.nsf/Content/604CF066BE48789DCA25751D000C15C7/$File/Summary%20of%20National%20E-Health%20Strategy-final051208.pdf). Accessed 21/08/2009.
- AHM (2009): Healthcare Identifiers and Privacy: Discussion paper on Proposals for Legislative Support. www.health.gov.au/.../Typeset%20discussion%20paper%20-%20public%20release%20version%20070709.pdf. Accessed 13/08/2009.
- Arends, R., Austein, R., Larson, M., Massey, D. and Rose, S. (2005a): RFC4033 DNS Security Introduction and Requirements. <http://www.ietf.org/rfc/rfc4033.txt>. Accessed 07/09/2009.
- Arends, R., Austein, R., Larson, M., Massey, D. and Rose, S. (2005b): RFC4034 Resource Records for the DNS Security Extensions. <http://www.ietf.org/rfc/rfc4034.txt>. Accessed 07/09/2009.
- Arends, R., Austein, R., Larson, M., Massey, D. and Rose, S. (2005c): RFC4035 Protocol Modifications for the DNS Security Extensions. <http://www.ietf.org/rfc/rfc4035.txt>. Accessed 07/09/2009.
- Goldschmidt, P. G. (2005): HIT and MIS: Implications of Health Information Technology and Medical Information Systems. <http://delivery.acm.org/10.1145/1090000/1089141/p68-goldschmidt.pdf?key1=1089141&key2=5606972511&coll=portal&dl=ACM&CFID=15151515&CFTOKEN=6184618>. Accessed
- Liu, C. and Albitz, P. (2006). DNS and BIND, O'Reilly Media Inc.,.
- NEHTA (2008a): Concepts and Patterns for Implementing Services Version 2.0 draft - 1 September 2008 Draft for Comment. http://www.nehta.gov.au/component/docman/doc_download/547-service-instance-locator-requirements-v10-draft-archived. Accessed 09/09/2009.
- NEHTA (2008b): Connectivity Architecture Version 1.0 - 1 December 2008 Release. www.nehta.gov.au/component/.../624-connectivity-architecture-v10-. Accessed 18/08/2008.
- NEHTA (2008c): Report on Feedback Individual Electronic Health Record, issued by the National Health and Hospitals Reform Commission
- NEHTA (2008d): Service Instance Locator: Requirements. www.nehta.gov.au/.../606-service-instance-locator-requirements-v11. Accessed 01/09/2009.
- NHHRC (2009): A Healthier Future for All Australians – Final Report <http://www.nhhrc.org.au/internet/nhhrc/publishing.nsf/Content/nhhrc-report>. Accessed 13/08/2009.

⁷ The international standard ISO15408 sets a strict guideline for evaluating security policy, program design documents, source code, manuals and other factors.

⁸ The Common Criteria Recognition Agreement (CCRA) Web site is available at <http://www.commoncriteriaportal.org/theccra.html>, accessed 03/09/2009.

- Scholl, M., Stine, K., Lin, K. and Steinberg, D. (2009):
Draft Security Architecture Design Process for Health
Information Exchanges (HIEs).
<http://csrc.nist.gov/publications/drafts/nistir-7497/Draft-NISTIR-7497.pdf>. Accessed 5/09/2009.
- Spronk, R. (2007): The Spine, an English National
Programme.
http://www.ringholm.de/docs/00970_en.htm. Accessed
30/08/2009.
- Spronk, R. (2008): AORTA, the Dutch National
Infrastructure.
http://www.ringholm.de/docs/00980_en.htm. Accessed
30/08/2009.
- The Dutch Ministry of Health (2007): Overview of the
Architecture on Dutch National E-health.
http://www.uziregister.nl/Images/emd_wdh_uk_tcm38-17362.wmv. Accessed 25/08/2009.

Design an automatic appointment system to improve patient access to primary health care

Hongxiang Hu

School of Information Systems
and Technology
University of Wollongong,
Wollongong 2522 Australia
hh959@uow.edu.au

Ping Yu

School of Information Systems
and Technology
University of Wollongong,
Wollongong 2522 Australia
ping@uow.edu.au

Jun Yan

School of Information Systems
and Technology
University of Wollongong,
Wollongong 2522 Australia
jyan@uow.edu.au

ABSTRACT

Advanced Access model has been introduced in general practice in the United States to improve patient access to primary health care services for more than ten years. It has brought in the benefits of eliminating service provider's waiting lists, improving patients' timely access to services and reducing no-show rate. However, to implement this model, practices need to collect relevant information, develop contingency plans and set up practice strategies to balance the provision of care and patient's demand. These tasks are not always easy to achieve. Understanding the requirements and constraints for effective management of patient booking is essential for developing an automatic appointment system that effectively supports this model in practice. This paper discussed these requirements and constraints, and then proposed a new model for automatic information collection, real time service monitoring and rule-based appointment decision making to balance demand and supply.

Keywords

Advanced Access, appointment system, patient access, primary health care

1. INTRODUCTION

In Australia, medical practitioners are distributed unevenly across the country that the practitioners in remote areas are 40% shortage compared with the average level in the country [1]. However, in some remote areas, such as Great Western and Southern of New South Wales, this ratio could only reach 1/8 of the average level [2]. Thus, patients always find it difficult with accessing to health care services in these areas. For example, a study in Wagga Wagga finds that a patient could wait up to 55 days for a routine appointment to see a General Practitioner (GP) [3].

Despite the shortage of workforce, one of the main reasons for this difficulty is that the way practice takes appointments (appointment model) mainly accounts providers' static schedule. For example, once a GP's schedule on a specific day has been fixed, patient appointment is arranged in the carved out slots. If

patients' demands on that day exceed the provider's supply within the scheduled time, then the exceeded demands will be postponed into the future schedule, which cause service delay. As time goes on, the postponed appointments form a long list of backlog, which seriously impedes patients' access to health care services. Previous studies indicate that service delay occurred more frequently when patients experience long time waiting [3-5], which wastes the precious GP's time slot allocated to this patient. It also blocks another patient's access to the services. In the long run, the recurrent occurrence of this situation will eventually cause the deterioration of the supply of health care services [6, 7].

A new appointment model entitled Advanced Access (AA) was proposed by Murray and Tantau [8] to balance supply and demand, and diminish backlog of appointments and delay of primary health care services. According to the AA model, practices are required to provide same-day service when a patient requests an appointment. The Advanced Access model proposes to achieve this goal through the implementation of six strategies: "balancing supply and demand, reducing backlog, reducing the variety of appointment types, developing contingency plans for unusual circumstances, working to adjust demand profiles, and increasing the availability of bottleneck resources". Direct benefits of Advanced Access include significant improvement in patient accessibility [3, 9-11] and the reduction of patients' no-show rates [3, 4, 9].

To date, the AA model is the best primary health care management model in terms of providing timely services to patients [7]. It is promoted in England by the National Primary Care Development Team as a way of improving access and achieving National Health Service (NHS) planned access targets. Studies indicate that 67% of practices in England claimed to operate Advanced Access [10]. Australian Primary Care Collaborative (APCC) has established its phase 2 program in December 2007, and one of the topics is to improve patient access to primary care [12]. Two empirical studies carried out by Dr Knight suggest that this model could work effectively in Australia as well [3]. Therefore we propose to develop an automated appointment system that is underpinned by the theory of the AA model.

As indicated above, the initiative of AA was well accepted in USA, UK and other countries. However implementing this model poses many new challenges to primary health care services. It requires shifting the criteria for appointment decision making from provider's schedule to patient's demand [13, 14]. The huge effort to manage the demand and supply leads to complex changes not only reflecting on the accessibility and no-show rates, but also continuity of care, providers' workload and practice working

culture [15]. For example, it requires receptionist and practice manager to take extra work to record patient's request on paper, and evaluate the daily change of every service provider's work load, accessibility and continuity of care [16]. Some practices find it difficult to implement Advanced Access because of the intrinsic dynamic nature of medical practices and inadequate guidelines for customising the advanced access approach to fit different styles of practices and demand patterns [17]. Some practices are unable to sustain the Advanced Access because they lack the capacity to dynamically manage the fluctuation of patient demand and provider supply.

The gap between the advantages of Advanced Access and the inability of a practice to implement this model calls for an innovative appointment system to support the implementation of Advanced Access. In this study, we will discuss the requirements and design of such a new, automatic appointment system. We would name this system Advanced Appointment System (AAS).

The rest of this paper is organised as follows. Section 2 describes important aspects need to be considered to design AAS; Section 3 presents the detailed design of AAS; Sections 4 discusses the appointment process handled by AAS. Finally, Section 5 concludes this paper and outlines the authors' future work.

2. Requirements Analysis

Tantau suggests that the three foundational elements for the success and sustainability of the Advanced Access model are capacity, continuity, and demand and supply equilibrium [9]. We will implement these elements in the AAS. Capacity is measured by the providers' working hours. Continuity is an important attribute to decide the quality of care, and can be traced by recording whether a patient is assigned to the appointed provider. There are many methods to balance demand and supply. The primary objective of balancing demand and supply is to guarantee the patient's accessibility to health care services. Accessibility can be measured by the ratio of patients' demands that have been fulfilled and the waiting time from initiation of a patient's request to the fulfillment of this request.

The problem of balancing demand and request is that although patient demand for services is predictable to a certain extent, it is not always accurate. This requires providers to accurately record patients' requests and plans the provision of care services accordingly. Imbalanced demand and supply may be caused by fluctuation of demand or shortage of medical staff members [18]. To balance demand and request, AAS can use 'straight method' to solve the temporary burst of demand by increasing service providers' workload; or adapt an 'alternative method' to solve the problem of shortage of medical staff, such as using the telephone or e-mail instead of visits to respond to patients' questions and to do follow-up care, developing group medicinal visits and extending the intervals between return visits for patients with chronic disease, providing patients and families with home-care education and reference materials [7].

Therefore, in order to improve patient's access to primary health care services, the AAS needs to have the capability to: (1) trace patient's demand; (2) manage provider's supply, and (3) maintain the balance between demand and supply.

3. System Design

Our proposed Advanced Appointment System is composed of six elements (Figure 1): a System Interface, a Data Repository, a Request module, a Performance module, a Strategy module and a Scheduling module. The Interface provides functionalities to exchange information with end users. The Data Repository is used to store all the relevant data that are used. The four modules are the core component in this design. The Request module is used to calculate patients' demand by tracing patients' everyday requests and sorting these requests in different categories. Basically, patient requests are entered by users from the System Interface; the Request module classifies these requests and stores them in the Data Repository. If a request already exists in the Data Repository, the module can retrieve this request and display it onto the System Interface. The Performance module checks the practice performance based on the extent of satisfaction of the demand. If the satisfaction level is below a defined threshold, the Performance module will trigger the Strategy module to adjust the strategy for patient appointment. Activated by the Performance module, the Strategy module will provide recommendations to assist a user to manually make appointments in light of the need of equalising demand and supply. The Scheduling module is used to dynamically manage provider's workload, arrange patient requests with appointments and display provider's schedule.

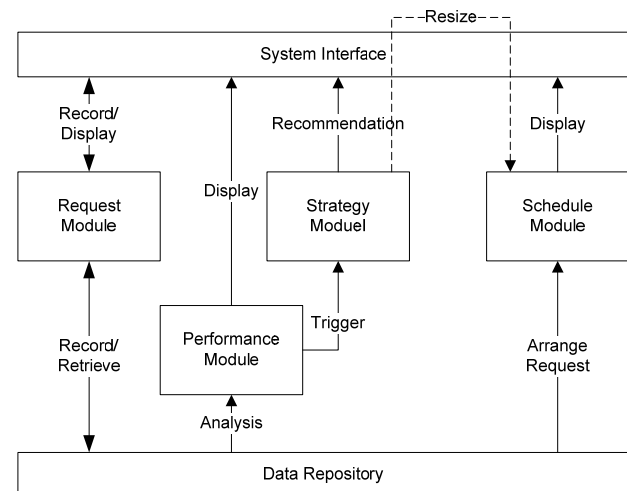


Figure 1: Component model of the AAS

The following sections detail the functions of each module.

3.1 Request module

The Request module is used to calculate the number of patient demands by tracing the processing of a patient's request in three states: booked appointment, pending request and discarded request (see Figure 2). There are three reasons for tracing these requests: First, to satisfy Advanced Access model, a practice aims to satisfy each patient's request for an appointment on the day they want it. Tracing each patient's request can help the practice to find the real demand for each service provider on a daily basis. Second, recoding the patient's demand into different patient categories can help the practice to estimate the types and number of services needed by patients. Third, finding out the ratio of

demands that have been fulfilled can enable the practice to understand the gap between demand and supply.

For example, Dr. Lightman can serve 30 patients per day, but he does not take pre-scheduled appointments. There is a burst of influenza and 45 patients need to see Dr. Lightman on one day. Obviously there would be 15 patients who could not see this GP. The demand for the next day's service from Dr. Lightman is 45 patients. Does it mean that the demand for Dr Lightman is now 45 patients per day? The answer is 'No. The patients come to see Dr. Lightman may include the patients who did not get the opportunity to see him the previous day, besides the patients that he sees regularly and some new patients. In order to accurately assess the patient demand for Dr. Lightman, we need to classify his patient request into three states: (1) booked appointment if the patient is offered an appointment; (2) pending request (or unsatisfied request), a middle state, if the patient is not offered an appointment, but wish to call back to fulfil this request; and (3) discarded request if the patient gives up this request. Figure 2 describes these three patient request states and the relationships amongst them for seeing Dr. Lightman on a day.

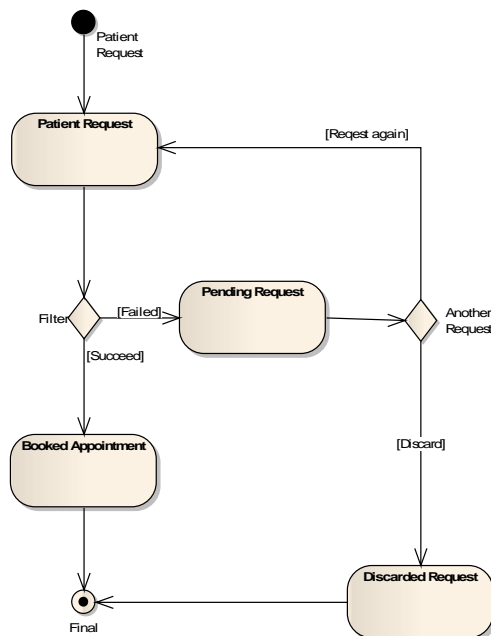


Figure 2: The three states of a patient's request for seeing Dr. Lightman and the transaction processes amongst the three states.

The following formula describes the relationship between the demand and request for Doctor Lightman:

Demand (one day) = All Requests – Pending Request.

This formula suggests that demand on a specific day equals to all the patients' requests on that day minus the number of patients' requests that was not fulfilled before and left in the pending list.

Fulfilled Demand (one day) = Booked Appointment – Booked Appointment from Pending Request.

This formula notes that fulfilled demand on a specific day equals to all the appointments booked on that day minus the number of appointments given to the pending requests.

3.2 Schedule module

Recoding patients' daily demand helps to estimate future supply; however, the estimated supply may not match the true demand on a particular day. If the estimated demand is lower than the actual demand, then extra capacity of supply needs to be established to match the demand [19]. In the example above, it is desirable to put in extra capacity to be in place to match the 15 extra demands for Doctor Lightman on that particular day. Although it is possible to put in extra capacity to handle the increasing demand, there is a constant worry that demand is infinite [17]; therefore, we propose a Schedule model to address this challenge. The Schedule module should sort provider's capacity into two categories: standard capacity and potential capacity. Standard capacity refers to the consultation that the providers can supply within their standard working session. Potential capacity is the quantity of consultation supplied on providers' extra time. By default the potential capacity is not displayed on a provider's schedule, but they are available when there is a shortage of supply. The reservation of the potential capacity is important for a practice to maintain the balance between demand and supply on a daily basis. This can reduce the backlog in the short term but may increase the workload of service suppliers. The size of a provider's capacity is managed by Schedule module; however, how this capacity is decided is supported by the rules from the Strategy module once been triggered by the Performance module (see Figure 1) and approved by an end user.

3.3 Performance module

The Performance module monitors the service's performance in a practice. The performance is measured according to the three targets of Advanced Access model: increasing accessibility and for patients, guarantee continuity of care for patient and balancing workload for service providers. as mentioned in the introduction section. Different practices have different requirements on service accessibility (e.g. same-day access or 48-hours access) [20, 21] and continuity (individual continuity or group continuity) [22] and flexibility on workload [14]. This requires the practice to set up the boundaries and thresholds for each of these attributes. Once these practice "rules" are determined the Performance module will be able to effectively execute its function of monitoring the performance of a practice and sending alarms to the system when the rules and standards of performance are violated.

3.4 Strategy module

To help with appointment decision making, the AAS incorporates a Strategy module to store all of the relevant rules for managing patient appointments. Once triggered by the signal sent from the Performance module, a relevant rule-based recommendation will be presented to an end user to facilitate the person to make the relevant appointment decisions. Currently 13 rules to be used in practice [3, 4, 18, 23-25] have been gathered and would be placed in our Strategy module, as listed below:

1. **Increase Provider Workload (use optional hours) (IPW(H))**: Providers provide small size extra capacity by using optional hours.
2. **Increase Provider Workload (use extra sessions) (IPW(S))**: Providers provide large size extra capacity by using extra sessions.
3. **Restrict New Patient (RNP)**: Providers refuse new patients added to their panel to reduce patients' demands.
4. **Restrict Prescheduled Appointment (RPA)**: Providers restrict prescheduled appointments on certain days for certain people to provide sufficient capacity on the specific day. This method is commonly used when practice try to shift the prescheduled appointment from certain heavy duty day, such as the day after holiday.
5. **Deny Prescheduled Appointment (DPA)**: practice restricts the number of days that patients can make pre-booked appointments. This may reduce the number of missed appointments and improve capacity available for patient, however may sacrifice some convenience for same patients such as aged people.
6. **Increase provider standard session (IPSS)**: increase a provider's standard work session, so that this provider's routine capacity will be increased, such as some of part-time providers changed to full-time during implementation of AA.
7. **Recruit a Temporary Physician (RTP)**: organisation recruits a provider to temporarily increase the health care supply. It temporarily improves the capacity to work down backlogs or to fill the capacity gap when a provider is on holidays.
8. **Recruit a Physician (RP)**: practice recruits a provider to increase the health care supply for long term to improve long term capacity.
9. **Assign Roles to Practice Nurse (ARPN)**: practice assign practice nurses to deal with certain cases. In this way, the provision can be increased for long term. Nurses can be potential alternatives to improve access to diabetes care in settings where physicians are not available [26].
10. **Group Consultation (GC)**: Provider provides consultation to a group of people at the same time. In this way, this provider could improve healthcare supply.
11. **Telephone Consultation (TC)**: provider provides telephone consultation to patients. In this way, the appointment interval can be reduced , It is not applicable for Medicare claim.
12. **Shift Demand to Other Providers (SDOP)**: practice shifts patients from high workload providers to low workload providers.
13. **Appointment Redesign (AR)**: practice redesigns the appointment types and intervals to increase the supply, such as evidence based practice to decide patient follow-up interval.

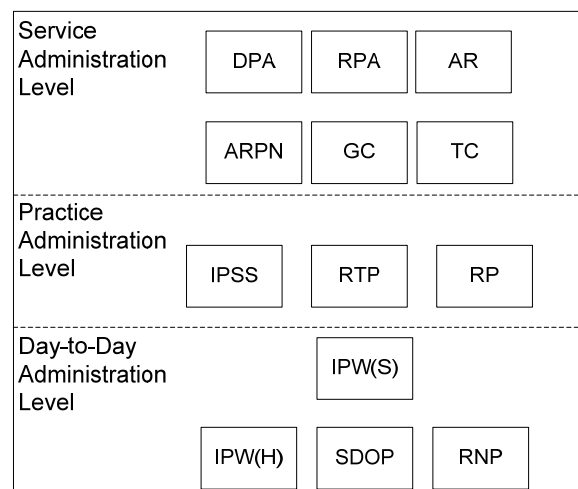


Figure 3: practice rules are organised in three levels: Day-to-Day Administration Level, Practice Administration Level, and Service Administration Level

These strategies have been organised at three levels based on the length of effects and complexity of implementation: Day-to-Day Administration level, Practice Administration level, and Service Administration level (In Figure 3), so that they can be invoked in different states. IPW(H), IPW(S), SDOP and RNP work at the Day-to-Day Administration level, because these rules are always used when a patient calls in to balance daily fluctuation of demands, and have short-term impact on balancing demand and supply. IPSS, RTP and RP work at the Practice Administration level that relays on the leadership to introduce new staff member into the practice, which result in the growth of supply in the long term. Strategies at this level will be suggested to work out the backlogs that providers could deal with, such as when a provider has left. The rest of the rules work at the Service Administration level, because GC and TC change the way that primary health care is provided to a patient; RPA, DPA and AR decide the way to take appointment; and ARPN changes the providers' structure. All of these rules at Service Administration level will have profound effects on the provision of care.

4. System Process

After explaining the functionalities provided by each model of the AAS, the next step is to design the work flow of the AAS to support Advanced Access. The proposed process of using the AAS to make an appointment for a patient includes three phases as shown in Figure 4.

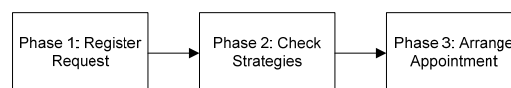


Figure 4: The process of arranging an appointment using advanced appointment system

4.1 Phase 1: Registering Request

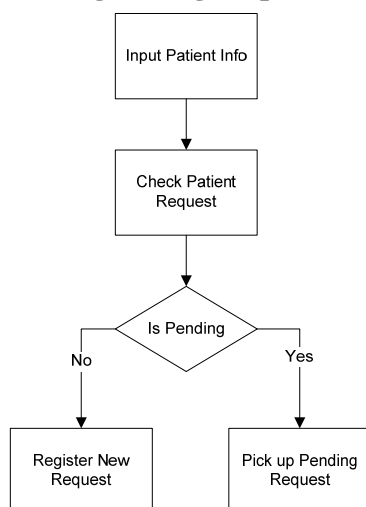


Figure 5: Phase 1, Record patient request

In phase 1 (See Figure 5), a receptionist receives a request from a patient and inputs this request to the appointment system. The appointment system checks the type of this request. If this request comes from a pending request in the system, the system will pick up this pending request; otherwise, it will register a new request into the system.

4.2 Phase 2: Checking Strategies

After registering patient request, the appointment system checks the performance of this type of service, which consists of three attributes: accessibility, continuity and workload. If the system performance remains at the acceptable level, the system will execute according to the default strategy. If the Performance module identifies abnormal performance, it will trigger a corresponding alarm in accordance with the relevant rules in its rules database, which is managed by Strategy module. This will alert an end user to implement the relevant strategy to improve the service performance (Figure 6).

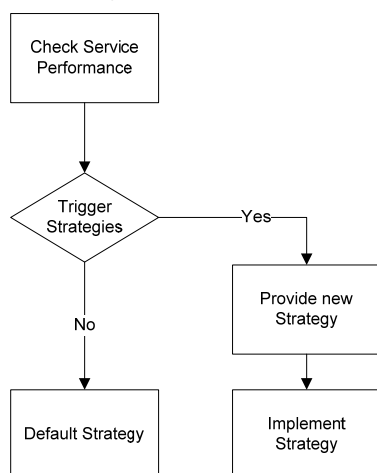


Figure 6: Phase 2, Check Strategies

As mentioned in Section 2.4, we have identified 13 rules that may affect the provision of care for a patient at three levels and from short term to long term. If the decline of service performance is caused by the fluctuation of request or demand, the system will provide Appointment Administration level rules. If the decline of service performance is caused by the shortage of supply, the system will not only provide Appointment Administration level rules to temporarily balance demand and supply, but also higher level rules to radically improve the service supply. The Appointment Administration level rules IPW(H) and IPW(S) can both provide extra capacity of supply but may increase a provider's workload. SDOP could increase the provider's capacity but affects the continuity of care as the patient is allocated to another provider (not group continuity); RNP can balance the request and demand with no effect on the existing patients and providers, but it decreases patient's accessibility to service. Therefore the selection of these strategies will be based on the priorities of the practice whether they prefer to privilege accessibility, continuity or workload.

4.3 Phase 3: Arrange Appointment

In phase 3, a receptionist follows the selected rules to arrange an appointment for a patient. If the patient is satisfied with the appointment, then the patient's demand has been fulfilled and the request turns to booked appointment. For any other reasons that the appointment is not booked, the system will postpone this request, and this request becomes a pending request.

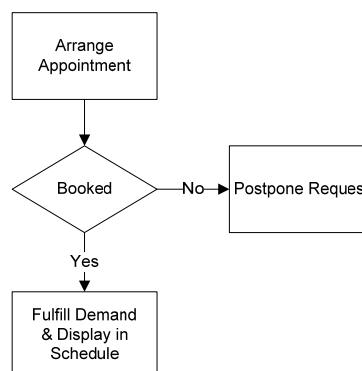


Figure 7: Phase 3, Arrange Appointment

5. Conclusions and Future Work

The Advanced Appointment System proposed in this paper presents an innovative solution to resolving the challenges for patients to access primary health care services. The mechanism for AAS to improve patient appointment process in primary health care includes: (1) revealing patient demand by steadily tracing patients' requests during the whole appointment process; (2) providing a performance triggered process to sustain the provision of care; and (3) structuring the practice rules to balance demand and supply under different circumstances. The fundamental work in this study, such as workflow analysis, has been taken in Centre Health Complex (CHC) in Shellharbour. In the future, a rigorous validation of the data model of AAS is required to validate the design of the AAS. An algorithm for the Performance module should be established to accurately trigger the implementation of

the right practice strategy based on the culture of Australian General Practice. Meanwhile, the pilot software of AAS is planned to be implemented in CHC for GP sectors and Allied Health Sectors, as a core component of integrated appointment system project.

ACKNOWLEDGMENTS

This research is supported by UOW Internal Industry Linkage Grant, 2008. Centre Health Complex has partly funded the project. We acknowledge Dr Rashid, Dr Niraj, and all of the healthcare workers and administrative staff in Centre Health Complex for their support for this project.

REFERENCES

- [1] Australian Institute of Health and Welfare. 2006. Medical labour force 2004.
- [2] Workforce Development & Leadership. 2009. Profile of the Medical Practitioners Workforce in NSW. City, 2009.
- [3] Knight, A. W., Padgett, J. and George, B. 2005. Reduced waiting times for the GP: two examples of "advanced access" in Australia. *Medical Journal of Australia*, 183, 2 (Jul 18 2005), 101-103.
- [4] Mehrotra, A., Keehl-Markowitz, L. and Ayanian, J. Z. 2008. Implementing open-access scheduling of visits in primary care practices: a cautionary tale. *Annals of Internal Medicine*, 148, 12 (Jun 17 2008), 915-922.
- [5] Belardi, F. G., Weir, S. and Craig, F. W. 2004. A controlled trial of an advanced access appointment system in a residency family medicine center. *Family Medicine*, 36, 5 (May 2004), 341-345.
- [6] Martin, C., Perfect, T. and Mantle, G. 2005. Non-attendance in primary care: the views of patients and practices on its causes, impact and solutions. *Family Practice*, 22, 6 (Dec 2005), 638-643.
- [7] Murray, M. and Berwick, D. M. 2003. Advanced access: reducing waiting and delays in primary care. *JAMA*, 289, 8 (Feb 26 2003), 1035-1040.
- [8] Murray, M. and Tantau, C. 1998. Must patients wait? The Joint Commission journal on quality improvement, 24, 423.
- [9] Tantau, C. Accessing patient-centered care using the advanced access model. 2009. *Journal of Ambulatory Care Management*, 32, 1 (Jan-Mar 2009), 32-43.
- [10] Salisbury, C., Montgomery, A. A., Simons, L., Sampson, F., Edwards, S., Baxter, H., Goodall, S., Smith, H., Lattimer, V. and Pickin, D. M. 2007. Impact of Advanced Access on access, workload, and continuity: controlled before-and-after and simulated-patient study. *British Journal of General Practice*, 57, 541 (Aug 2007), 608-614.
- [11] Salisbury, C., Goodall, S., Montgomery, A., Pickin, D., Edwards, S., Sampson, F., Simons, L. and Lattimer, V. 2007. Does Advanced Access improve access to primary health care? Questionnaire survey of patients. *The British Journal of General Practice*, 57, 615.
- [12] Knight, A. W. 2009. Learning from four years of collaborative access work in Australia. *Quality in Primary Care*, 17, 71-74.
- [13] Hroschickski, M. C., Solberg, L. I., Sperl-Hillen, J. M., Harper, P. G., McGrail, M. P. and Crabtree, B. F. 2006. Challenges of change: a qualitative study of chronic care model implementation. *Annals of Family Medicine*, 4, 4 (Jul-Aug 2006), 317-326.
- [14] M., Ahluwalia, S. and Offredy, M. 2005. A qualitative study of the impact of the implementation of advanced access in primary healthcare on the working lives of general practice staff. *BMC Family Practice*, 6(Sep 27 2005), 39.
- [15] Dixon, S., Sampson, F. C., O'Cathain, A. and Pickin, 2006. M. Advanced access: more than just GP waiting times? *Family Practice*, 23, 2 (Apr 2006), 233-239.
- [16] Baxley, E. G., Weir, S., Baxley, E. G. and Weir, S. 2009. Advanced access in academic settings: definitional challenges. *Annals of Family Medicine*, 7, 1 (Jan-Feb 2009), 90-91.
- [17] Gill, J. S. 2004. A nonfinancial approach to financial improvement of medical groups through advanced access. *Journal of Healthcare Management*, 49, 4 (Jul-Aug 2004), 271-277.
- [18] Baugh, R. F., Alpard, C. R. and Colon, E. Advanced access to otolaryngology: lessons learned. *Otolaryngology - Head & Neck Surgery*, 138, 2 (Feb 2008), 140-142.
- [19] Gupta, D., Potthoff, S., Blowers, D. and Corlett, J. Performance metrics for advanced access. *Journal of Healthcare Management*, 51, 4 (Jul-Aug 2006), 246-258; discussion 258-249.
- [20] Campbell, J. L., Ramsay, J., Green, J. and Harvey, K. Forty-eight hour access to primary care: practice factors predicting patients' perceptions. *Family Practice*, 22, 3 (Jun 2005), 266-268.
- [21] Mitchell, V. 2008. Same-day booking: success in a Canadian family practice. *Canadian Family Physician*, 54, 3 (Mar 2008), 379-383.
- [22] Salisbury, C., Sampson, F., Ridd, M. and Montgomery, A. A. 2009. How should continuity of care in primary health care be assessed? *British Journal of General Practice*, 59, 561 (Apr 2009), e134-141.
- [23] Murray, M., Bodenheimer, T., Rittenhouse, D. and Grumbach, K. 2003. Improving timely access to primary care: case studies of the advanced access model.[see comment]. *JAMA*, 289, 8 (Feb 26 2003), 1042-1046.
- [24] Pomerantz, A., Cole, B. H., Watts, B. V. and Weeks, W. B. 2008. Improving efficiency and access to mental health care: combining integrated care and advanced access. *General Hospital Psychiatry*, 30, 6 (Nov-Dec 2008), 546-551.
- [25] Meade, J. G. and Brown, J. S. 2006 Improving access for patients - a practice manager questionnaire. *BMC Family Practice*, 7, 37.
- [26] Fall, C. 2001. Non-industrialised countries and affluence: Relationship with Type 2 diabetes. *British medical bulletin*, 60, 33.

Access to E-Health information for the eNomad

Anthony D. Stiller

C Management Service Pty Ltd

Central Queensland University Brisbane International Campus

108 Margaret Street, Brisbane 4000, Queensland

t.stiller@bris.cqu.edu.au

Abstract

The concept and implementation of an E-Health scheme is not new. There is an array of literature supporting the benefits derived by individuals, health industries and government agencies who gain access to digital health records. Access improves the level of health literacy and assist in making informed decisions. Senior citizens are becoming more techno savvy and purchasing affordable mobile wireless enabled Information and Communication Technologies (ICTs) that easily connect to the expanding broadband Internet footprint across cities and remote regions in Australia. This provides opportunities for the electronic nomad (eNomad) and health professionals to gain access to personal digitally stored medical records regardless of the geographical location. This paper will focus on the adoption, use and impact of ICTs to gain access to digital health records by the eNomad and the health industry and its use as a tool to improve information literacy and informed decision making.

Keywords: electronic nomad, individual electronic health record, individual healthcare identifier.

1 Introduction

Research on the elderly and the adoption of ICTs indicate that seniors are more willing to adopt technologies and go online empowering them to take control of their social interaction, communication and health needs. Further studies show that seniors who use the Internet to seek information report greater satisfaction levels and experience less anxiety (Juznic et al. 2006).

The Australian population is more mobile now than ever before, relocating geographically through work, education, lifestyle choices and this includes the grey nomads. This mobile and aging population is placing greater pressure on health services across Australia as they call upon health centres for consultations and treatment. Health professionals rely on accurate data and information on health records and the status of the person attending that health centre.

While ICT systems have the ability and capacity to store large volumes of data on individual patients who access these health services in their immediate location, gaining access to health records by the individual and/or health professional in distant or remote locations is not a simple process. This is due to ICT systems interconnectivity, interoperability, records being stored using different database platforms and formats. This also raises issues of completeness and accuracy of records, security, privacy, ethics and individual sensitivity.

For an E-Health initiative to be successful a holistic approach on the adoption and diffusion of ICTs, design, development and rollout using a national E-Health framework and platform required. It requires user input and collaboration from all entities that have a vested interest and access to this service.

2 eNomad

Obst et al. (2008) applies the term grey nomad to 'semi-retired, or retired people who travel for all, or part of the year, throughout Australia [and] generally seen as being aged 55 and over. These long term travellers usually use a caravan or a motorhome as their base'. Grey nomads generally travel with no particular schedule or date to return to their normal place of residence. They are known to join or establish large social networks, carry limited documentation on medical and medication history and members of the baby boomer generation seeking lifestyle choices. As more grey nomads purchase ICTs and use them in their everyday life, a new group of nomads have emerged, the electronic or eNomad. With no universal definition for the eNomad, the definition of the grey nomad has been extended to incorporate '*using ICTs to communicate and interact with other members of family and society as they travel to places of choice*'.

In the study conducted by Obst et al. (2008) on the grey nomads, 21% of females and 32% of males experienced a health incident during the 2 year period of their study. Their major concern was reaching medical assistance in a timely manner. Examples quoted in the study included severe chest pains, back injuries, deep vein thrombosis, severe arthritis, cervical cancer diagnosis, pneumonia, and fainting fits. These medical incidents caused added concern and stress as they travelled in unfamiliar parts of the country without the knowledge of available medical facilities and services in that region. Stress levels increased due to time delays in reaching a medical facility

and having to explain in great detail their full medical history (if they were fit to do so) to a health professional who then has to make an informed decisions as to what treatment is best for the patient.

For the eNomad, making a doctors or dentist appointment on the move is made much easier using a wireless enabled PDA/GPS/phone or laptop device. These devices enable the person to search for an available doctor, dentist and chemist or health professional as they travel from town to town or State to State across Australia. These ICTs enable them to use the Internet to make appointments, set the date and time, pay online and use the inbuilt GPS to guide them to the health facility. After the consultation and treatment, the updated information could be entered into the E-Health database and the eNomad could download the latest information into their mobile device for reference at a later date.

The eNomad could be a driver of innovation change guiding the development and rollout of eHealth interconnected services due to their individual health needs and being part of their social network. National E-Health Transition Authority (NEHTA) would be well advised to include the grey nomad organisations and eNomads in the eHealth project by including these groups during the information gathering, development and implementation lifecycle of the E-Health project. This will ensure the services provided on an E-Health portal and database meets the needs of a group of users whose numbers are increasing year by year.

3 E-Health

The concept and implementation of an E-Health individual electronic health record (IEHR) system in various countries around the world including Australia is not new. Countries such as Canada, USA, United Kingdom, Germany, France, Ireland, New Zealand, Denmark, Sweden, South Africa, Norway, Singapore and Australia (Cornwall 2002; HIMSS 2008) have either sponsored research or implemented E-Health projects in one form or another to digitize medical records so they can be accessed by various health related organisations to assist in managing medical records (MMR), transaction processing (TP) and in decision support systems (DSS).

The 2004-2005 National Health Survey (NHS) collected information to describe various aspects of the health status of the Australian population, with a particular focus on the National Health Priority Areas (NHPA) of asthma, cancer, heart and circulatory conditions, diabetes, injuries, mental wellbeing and musculoskeletal conditions, particularly arthritis and osteoporosis (ABS 2004-2005). The same survey identified additional medical conditions the eNomad faces as they travel around Australia that may require medical assistance and includes attacks by another person, bites or stings, bruising, burns or scalds, choking, cuts, dislocations, sprains, strains, electric shocks, falling over, fractures and broken bones, hit by something, hitting something,

inhaling fumes, internal injuries, loud sounds, near drowning, swallowing poisons and vehicle accidents.

At the Australian Health Ministers Conference (AHMC) in December 2008, the definition for E-Health proposed by the World Health Organisation as 'the combined use of electronic communication and information technology in the health sector' was adopted. When implemented in Australian, the objectives of the scheme is to ensure that 'the right health information is provided to the right person at the right place and time in a secure, electronic form for the purpose of optimizing the quality and efficiency of health care delivery' would be realized (AHMC 2008).

Ad hoc E-Health initiatives are being developed by private and public organisations include NEHTA, the federal Department of Health and Ageing (HealthConnect), the Queensland State Government via the delivery of electronic medical records and the Australian Capital Government commitment investing in E-Health technology. According to the Australian Health Information Council (AHIC), there is no coordination or cooperation between federal and state authorities on a national framework to E-Health and the sharing of electronic records (AHIC 2007). This was made evident in a consultant report prepared Deloitte Touche Tohmatsu for AHMC alerting Ministers to the fact that:

"the health information landscape is characterised by discrete islands of information with significant barriers to the effective sharing of information between health care participants. It also poses challenges when trying to understand and report on what is really happening to support population health surveillance and guide policy, service planning, innovation and clinical and operational decision-making (AHMC 2008).

Table 1 shows the results of a study conducted by The Royal Australasian College of Physicians (2007, p.20) indicating the highest percentage of computer usage by physicians in all three categories is searching for information on the Internet, while using a computer for patient records is low.

Computer use of electronic applications	Public hospital		Private hospital		Consulting rooms	
	n	%	n	%	n	%
Electronic prescribing	92	6	34	2	173	11
Electronic request for investigations	246	16	42	3	130	8
Receiving results electronically	854	54	124	8	310	20
Electronic patient notes	302	19	50	3	230	15
Electronic referrals	140	9	26	2	65	4
Electronic letters to patients	234	15	33	2	125	8
Patient held records/USB/Data keys	185	12	42	3	133	8
Administration practice finance	310	20	67	4	377	24
Educational activities	837	53	63	4	348	22
Searching internet for information	1123	71	120	8	486	31

Table 1: Computer use of electronic applications

The most encouraging outcome of the report was that 94% indicated they would use electronic applications for online prescribing and 95% would use online computers systems for evidence based information.

Pearce (2009) in his study on 'Electronic medical records – where to from here?' says that while 90% of general practitioners have a computer on the desk where they work, only 65% use them for clinical records and processing notes to update patient information at the point of care.

While it is technically feasible to rollout a national E-Health system across Australia and create a database management system to capture, interrogate and display information on a range of ICTs both fixed and wireless, it does not mean that patients, physicians, GPs and health professionals would necessarily use the technologies for decision making. Accuracy of medical records recorded by GPs and health professionals resulting from diagnosis, procedures and blood tests relies on the competency of the staff (Wan et al. 2009). Adoption of a national coding scheme for all routines would ensure consistency.

Access to E-Health records regardless of geographical location is important for the grey nomads and eNomads. However, health literacy has a marked impact on the ability of the individual to assess E-Health records and make an informed decision on how to manage medical conditions. Access to E-Health records has the potential to 'reduce costs in the health system, prevent illnesses and chronic diseases, and reduce the rates of accidental death' (ABS 2009). Data from 2007-2008 national health survey shows that 85% of people over the age of 65 years had 'three or more long-term health conditions' to manage (ABS 2009) at any one time.

With the objective of an E-Health system to provide ease of access to authorised users, there is a possibility that data being entered into the system could be subject to error leading to misdiagnosis and compound the medical condition. During the design, development and testing of the E-Health system, information extracted from the system would need to be checked for accuracy, completeness, reliability and readability from the perspective of the health professional (Rawson & D'Arcy 1998) and the grey and eNomad.

The National E-Health Transition Authority (NEHTA) is one of a number of organisations working with federal, state and territory governments to develop a national approach for electronically collecting, securing and exchanging health information across borders' using multi-channel communications means. This will provide the eNomad with the surety that regardless of where they are, so long as they have access to Internet, should a medical situation arise the most current medical information can be shared between GPs, specialists, hospitals and other medical related professions.

4 ePrescriptions

The use of electronic prescription (ePrescription) or e-prescribing replaces the paper based system used by entering prescription details directly into a computer based system that can be accessed by any authorised pharmacy or chemist connected to the network (Lapen

2007). A media release from The Pharmacy Guild of Australia (2009) reported that of the '250 million prescriptions dispensed each year 60 per cent are repeat prescriptions'. Carrying around repeat prescriptions could be a risk to the grey nomad if they are misplaced, lost or out of date.

In the Northern Territory, a small number of medical centres and pharmacies are providing an ePrescription service under their ehealthNT program that requires a digitally encrypted signature and barcode before the system is accessed (Katehar 2008). While this initiative provides a limited service in Darwin, experiences of countries such as England, United States who have standalone or integrated systems shows that patients safety is improved by preventing errors through medication management, delivers cost-benefits and creates linkages between laboratories and pharmacies (Vatanara 2008).

The benefits of incorporating an ePrescriptions into the E-Health database would enable the grey nomad and eNomad to have their medications dispensed at any pharmacy or chemist in Australia. When changing medical conditions occur, the outstanding ePrescription can be cancelled, removing the possibility of increasing medical complications since the medication will no longer be required or appropriate. Using mobile ICTs, the eNomad can use their device to check on the status of prescriptions, dosage rates, expiry dates and locate a pharmacy or chemist to have them dispensed.

A study conducted by McGuire (2005: in Reisenwitz et al. 2007), showed that 37% of online seniors used the Internet for the purpose of seeking information on drugs and other health related issues. For the eNomad, being on the road allows them to continue online searches, post blogs, chat online and use other forms of social networking to improve their level of medical literacy on which to make informed decisions.

At present the NHETA E-Health model does not provide for the integration of the MediCare numbering identification number as it is not unique (Dearne 2009a). The proposed E-Health card will contain a 16 digit individual healthcare identifier (IHI) number to access the system. This will require the eNomad to carry two cards, their family MediCare card to access the payments system, and E-Health card with their IHI number to access the E-Health system. This will lead to confusion for the grey and eNomad.

5 eCommunication

To gain the most benefit from a national E-Health system requires an effective and secure communication pathway between the medical professionals, the E-Health database and the grey and eNomads. Accuracy and reliability will increase where data entered into the records uses a national standardized code set (Pearce 2009). It also requires seamless connectivity across disjointed islands of health record systems across Australia. The present

independent development of E-Health decision support systems by various state and territory governments is not based on a national framework or technical specifications. This is preventing the independent systems providing interconnectivity and the sharing of data (AHIC 2008). The Australian federal health minister has called for full cooperation between state and territory governments (Dearne 2009a) and the health industry to ensure there is a balance between access, privacy and security of personal information (Dearne 2009b). Drawing upon the study by Parnaby and Towill (2007), their 'patient-centered seamless supply chain' model could be replicated into the Australian national E-Health initiative by adopting a systems engineering approach to replace the uncoordinated independent systems by the federal, state and territory governments.

The Australian federal health minister and the Council of Australian Governments are committed to introduce the 16 digit individual healthcare identifier (IHI) as a way to accurately identify healthcare providers, organizations and individuals (NEHTA 2007a). In a survey conducted by NEHTA (2008), over 90% of respondents indicated that would want their health records, and those of their children included in the IEHR system.

One of the most effective communications tools is an Internet based portals as the window to gain access and distribute E-Health information to health professionals and the eNomad. Secure encrypted text messaging over the telephone network (fixed or wireless) and by completing predefined web based forms on the portal can be used to streamline the process. The success however is dependent on a well designed and functional portal containing rich and relevant information based on the computer literacy level of the user (Liederman et al. 2005). The convergence of Internet protocols (IP), availability of faster broadband networks, together with the emergence of new ICT devices incorporating voice, video and data across the telecommunication network could easily access the E-Health system. This will provide communication between the eNomad, their general practitioner, specialist, hospital system, health services and other health professionals including ambulance and other emergency services.

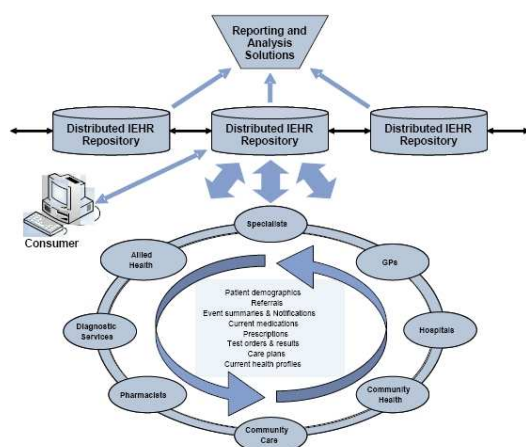


Figure 1: Distributed Database Model

The distributed database model (Figure 1) proposed for the IEHR (AHMC 2008 p. 15) links compliant separate subsystems of healthcare providers and organisations into the national repository. This will ensure the data and information remains accurate, reliable, secure, private and traceable (audited) regardless of the access point and ICTs used.

6 mScanner

The application of mScanner (mobile scanner) hardware and software is available on the 3G, 4G and NextG mobile device equipped with a digital camera. An image of a barcode could be transmitted to the E-Health database or the Food Standards Australian and New Zealand (FSANZ) database storing data from manufacturers', food processors, packaging companies and distributors using food labels containing information on nutrition value, ingredients and major allergens foods. Cross referencing would provide instant feedback on the product and its suitability for consumption so as to prevent a health reaction based on the individuals' medical condition. The information in electronic form would provide the eNomad with a rich data source on which to make informed decisions on the purchase, preparation, consumption and storage of foods.

This technology also has the potential when connected to a dietary database to assist in the selection of healthy foods, their preparation, quantities of consumption and information on their safe handling. Once connected to the Internet and the E-Health portal, a search of the Food Standards Australia New Zealand (FSANZ) database for information on any additives, food safety, labeling, GM foods and recommend alternatives. The same technology could also be used by the eNomad to scan pharmaceutical products and natural health products sold in supermarkets and health stores and cross referenced to the database so as to reduce medical related incidents requiring medical assistance.

A further extension to the E-Health system could include the service provided by dietitians and nutritionists to provide advice on food types, consumption and suggested menus and exercise routines for the eNomad on the road. Advice and recommendations would be based on their medical condition, the geographic location, climate, season of the year and food types in that region.

7 Barriers to e-Health Adoption

While the benefits of introducing a seamless national E-Health have been well documented, there are a number of barriers to its widespread acceptance. GPs and other health professionals may not want to store sensitive information about the patient in a national database and suggest they seek consent from the patient before storing the data where it is to be shared (Tap et al. 2009). Many medical professionals respond to the sensitive needs of their patients so as not hinder treatment (Saiid, Tonsi & Baig 2008). The need to engage health professionals and the public in identifying issues that may hinder its successful design, development and implementation and

adoption and diffusion have been recognized by the bodies charged with overseeing the E-Health initiative (AHIC 2008; NEHTA 2007b).

A summary report provided by NEHTA (2007c) from attendees at a workshop on E-Health initiatives identified the cost of developing new systems, redeveloping existing ICT interfaces, change management and resources as barriers. In addition, the report identified the need for a consistent use of terminology and agreeing on a common code set for recording data and generating reports. The roadmap spread over 3 to 10 years could also act as a barrier to its successful implementation as the medical needs of professional bodies, state and territory governments and the public may change over that period of time. In addition, the political will of successive governments to continue the project could also be a barrier as it spans across a 3 to 4 year political lifecycle (NEHTA 2007c).

It would appear from the same summary report the key issues seen as barriers to the successful implementation of an E-Health initiative centre around the governance principles of 'clarity of accountability, transparency, stakeholder representation, sustainability, support for activity at multiple levels, effective leadership and coordination, balance local innovation and national outcomes' (NEHTA 2007c p. 18).

8 Conclusion

The numbers of grey nomads and eNomads travelling around Australia seeking a lifestyle change are increasing. As their health needs increase, access to secure, high quality, equitable and sustainable health information systems in health centers and facilities located away from their normal geographical region is essential if their travels are to be stress free. One way to meet these needs is through the implementation of the integrated E-Health initiative being developed by NEHTA. The aims and objectives for developing such a scheme will provide access to medical records anywhere in Australia through the IEHR system. Inclusion of a national ePrescription service would also benefit the eNomad and general public alike. Access to the distributed IEHR database model using the 16 digit individual healthcare identifier will improve health literacy so that more informed decisions can be made on individual's diagnosis and treatment options. It will enable health professionals to use the most up to date ICTs to access information stored in the IEHR database.

The key benefits derived through the E-Health initiative include cost reductions on health delivery by governments, hospitals and patients. This can only be realized when the barriers to its rollout have been overcome. While these are admirable objectives, cost reduction should not be the driving force for its implementation. Rather, the focus should be on delivering a national E-Health service that promotes a healthier society which includes access to digital medical records by the eNomad.

9 References

- Australian Bureau Statistics (2009): Australian Social Trends, **4102.0**. Canberra Australia.
- Australian Bureau of Statistics (2004-2005): National Health Survey: Summary of Results, **4364.0**. Australia.
- Australian Health Information Council (2008): Electronic Decision Support Systems Report. Australia.
- Australian Health Information Council (2007): E-Health Future Directions Briefing Paper. Australia.
- Australian Health Ministers Conference (2008): National E-Health Strategy Summary, *Victorian Department of Human Services*. Australia.
- Australian Health Information Council (2008): Communiqué. Australia.
- Cornwall A. (2002): Electronic Health Records: An International Perspective, *Health Issues*, **73**:19-23.
- Dearne, K. (2009a): Small steps better in e-health, *Australian IT*, August 18, 2009. Australia.
- Dearne, K. (2009b): Health rebate cuts could fund e-health: Roxon, *Australian IT*, August 19, 2009. Australia.
- HIMSS Enterprise Systems Steering Committee (2008): Electronic Health Records: A Global Perspective: *Healthcare Information and Management Systems Society*.
- Juznic, P., Blazic, M., Mercun, T., Plestenjak, B. & Majcenovic, D. (2006): Who says that old dogs cannot learn new tricks? A survey of internet/web usage among seniors: *New Library World*: **107** (1226/1227):332-345.
- Katehar, L. (2008): NT delivers national first for electronic prescriptions: *Department of Health and Families*. Australia.
- Lapane, K.L., Dube', C., Schneider, K.L. & Quilliam, B.J. (2007): Patient Perceptions Regarding Electronic Prescriptions: Is the Geriatric Patient Ready?: *Journal American Geriatrics Society*: **55**:1254-1259.
- Liederman, E.M., Lee, J.C., Baquero, V.H. & Seites, P.G. (2005): Patient-Physician Web Messaging The Impact on Message Volume and Satisfaction: *Journal of General Intern Medicine*: **20**:52-57.
- National E-Health Transition Authority (2008): National E-Health Transition Authority. Australia.
- National E-Health Transition Authority (2007a): Action Plan for Adoption Success The response to the independent review of NEHTA. Australia.
- National E-Health Transition Authority (2007b): Individual Healthcare Identifier Fact Sheet. Australia.
- National E-Health Transition Authority (2007c): Question 1: 'What are the main barriers to implementing the proposed path forward? Australia.
- Obst, P., Brayley, N., & King, M. (2008): Grey Nomads: Road Safety Impacts and Risk Management: In: *2008 Australasian Road Safety Research, Policing and Education Conference*. Adelaide, South Australia.
- Parnaby, J. & Towill, D.R. (2008): Seamless healthcare delivery systems: *International Journal of Health Care Quality Assurance*: **21**(3):249-273.
- Rawson, N. & D'arcy, C (1998): Assessing the Validity of Diagnostic Information in Administrative Health Care Utilization Data: Experience in Saskatchewan:

- Pharmacoepidemiology and Drug Research Safety*: 7:389-398. John Wiley & Sons, Ltd.
- Pearce, C. (2009): Electronic medical records – where to from here?: *Australian Family Physician*: **38**(7).
- Queensland Health (2008) ‘*Nutritionist Fact Sheet*’ <http://www.health.qld.gov.au/phcareers/workers/nutritionist.asp>. Accessed 9 September 2009.
- Reisenwitz, T., Iyer, I., Kuhlmeier, D.B. & Eastman, F.K. (2007): The elderly’s internet usage: an updated look: *Journal of Consumer Marketing*, **24**(7):406–418.
- Sajid, M.S., Tonsi, A. & Baig, M.K. (2008): Health-related quality of life measurement, *International Journal of Health Care Quality Assurance*: **21**(4):365-373.
- Tapp, L., Elwyn, G., Edwards, A., Holm, S. & Eriksson, T. (2009): Quality improvement in primary care: ethical issues explored, *International Journal of Health Care Quality Assurance*, **22**(1):8-29.
- The Royal Australasian College of Physicians (2007): The computer will see you now. A study to determine access and use of computers and electronic applications by physicians. Australia.
- Vatanara, A., . Vahdat, D., Rouholamini Najafabadi, A. & Noori, L.K. (2008): E-prescribing: A preliminary paradigm for Iran health system, *The Second Conference on Electronic City*, **36**:2592–2602.
- Wan, E., Wan, R. & Kamaruzaman, J (2009): Service quality in health care setting, *International Journal of Health Care Quality Assurance*, **22**(5):471-482.

A Multidimensional Temporal Abstractive Data Mining Framework

Heidi Bjering¹ and Carolyn McGregor^{2,1}

¹School of Computing and Mathematics
University of Western Sydney
Locked Bag 1797, Penrith South DC NSW 1797, Australia
h.stratti@uws.edu.au

²Faculty of Business and IT/Faculty of Health Sciences
University of Ontario Institute of Technology (UOIT)
2000 Simcoe St North, Oshawa ON L1H 7K4, Canada
c.mcgregor@ieee.org

Abstract

This paper presents a framework to support analysis and trend detection in historical data from Neonatal Intensive Care Unit (NICU) patients. The clinical research extensions contribute to fundamental data mining framework research through the integration of temporal abstraction and support of null hypothesis testing within the data mining processes. The application of this new data mining approach is the analysis of level shifts and trends in historical temporal data and to cross correlate data mining findings across multiple data streams for multiple neonatal intensive care patients in an attempt to discover new hypotheses indicative of the onset of some condition. These hypotheses can then be evaluated and defined as rules to be applied in the monitoring of neonates in real-time to enable early detection of possible onset of conditions. This can assist in faster decision making which in turn may avoid conditions developing into serious problems where treatment may be futile.

Keywords: Clinical research, data mining, temporal abstraction.

1 Introduction

In the industrialized world, premature birth has been recognized as one of the most significant perinatal health issues (Kramer et al., 1998). In Australia 8.1% of babies are born before 37 weeks gestation (Laws et al., 2007). Premature babies often have prolonged stays in Neonatal Intensive Care Units (NICUs) and can suffer from a number of different conditions during their stay. Some of these conditions have been shown to exhibit certain variations in their physiological parameters that can indicate the onset of such conditions, before it can be detected by other means. Sepsis, a common illness in neonates, has been shown to exhibit changes in physiological data before the condition can be diagnosed through blood cultures (Griffin et al., 2007). This

indicates that subtle changes that may not be apparent through the current practice of manual recording of the physiological data at regular intervals can be important in detecting the onset of conditions in neonates. There is a need for systems aimed at clinical management to help analyse complex multidimensional data produced by the monitoring and life support devices connected to the babies, such as the one described in (Stacey et al., 2007). There is also a need for clinical research systems and frameworks to facilitate clinical research on stored historical physiological patient monitoring data to enable the discovery of previously unknown trends and patterns that may be indicative of the onset of some condition.

Medical monitoring equipment produces large amounts of data, which makes analysing this data manually impossible. Adding to the complexity of the large datasets is the nature of physiological monitoring data – the data is multidimensional, where it is not only changes in individual dimensions that are significant, but sometimes simultaneous changes in several dimensions. As the time-series produced by the monitoring equipment is temporal, there is a need for clinical research frameworks that enables both the dimensionality and temporal behaviour to be preserved during data mining. When analysing time series data, the individual data values by themselves often provide little meaning; however when considered over time and context the values can become meaningful. Temporal abstraction is a technique used to summarize raw time series data to a higher level while preserving context and time (Shahar, 1997), usually adding qualitative information such as states and trends to a particular abstraction. In a NICU setting it is usually trends or changes in the physiological data over time, sometimes across multiple parameters, which are significant when analysing and predicting patient conditions, therefore making temporal abstractions of this data prior to mining very relevant.

The Temporal Abstractive MultiDimensional Data Mining (TAMDDM) framework is a flexible framework designed not just for one particular study or research project, but to enable multiple varied studies on the collected historical data. It can be used for the detection of trends and patterns to recognize possible indicators of some condition in the neonate, to enable the creation of hypotheses that can be transformed into rules suitable for use in intelligent monitoring systems.

Copyright © 2010, Australian Computer Society, Inc. This paper appeared at the Australasian Workshop on Health Informatics and Knowledge Management (HIKM 2010), Brisbane, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 108. Anthony Maeder and David Hansen, Eds. Reproduction for academic, not-for-profit purposes permitted provided this text is included.

This paper is presenting a multidimensional data mining framework that is suitable for use in clinical research, including provisions for application of temporal abstraction on the time series data, and support for null hypothesis testing. The framework produces hypotheses that can be translated into rules to be used by a real-time event stream processor used for intelligent monitoring and alerting in a NICU environment.

In the rest of this paper we first present related research in the area of temporal abstraction and data mining in a medical setting. The TAMDDM framework is then introduced and a demonstration of the temporal abstraction and realignment tasks are provided, followed by the conclusion including future research.

2 Related work

In medicine temporal abstraction can be used to convert low level raw numeric time series data into a higher level qualitative description which better matches the language used by medical professionals (Stacey and McGregor, 2007). Some healthcare systems use temporal abstraction to abstract to the level of descriptions or guidelines. Abstracting to this level enables the matching of these abstractions for guideline execution in clinical management. An example of which is the system developed by Seyfang et al (2001) for optimising oxygen supply for newborn infants. Their system, which is part of the Asgaard framework (Shahar et al., 1998, Seyfang and Miksch, 2004), uses the Asbru language (Seyfang and Miksch, 2004, Fuchsberger et al., 2005) and abstracts raw monitoring data collected by NICU monitoring devices to the abstract concepts that are used in therapeutic plans. The data enters the system as a stream and the high-level abstractions derived from the raw data are compared to predefined conditions described in the therapeutic plans. RÉSUMÉ is a system which provides a “framework for deep knowledge representation to perform temporal abstraction of patient data” (Stacey and McGregor, 2007). It uses CAPSUL, a temporal pattern representation language (Antunes and Oliveira, 2001, Chakravarty and Shahar, 2000). RÉSUMÉ is used on stored database data of low frequency of the abstracted parameters. The Tzolkin architecture uses RÉSUMÉ to create abstractions (Boaz and Shahar, 2003). RASTA (A System for Temporal Abstraction) (O'Connor et al., 2001) adds distributed capabilities to RÉSUMÉ to allow the system to be used for more complex settings (Augusto, 2005).

PROTEMPA (Post and Harrison, 2007) has a system for implementing temporal abstractions in stored time series data, both lower level (simple) and higher level (complex) abstractions. These abstractions are used for identifying pre-defined patterns in the time-series data. The system has the potential to be used in patient monitoring and decision support where the patterns being looked for are predefined, however it is not used for discovering *new* patterns and relationships in the abstracted data.

Shahar's pivotal work (Shahar, 1997) presents a framework for Knowledge Based Temporal Abstraction (KBTA) which infers abstractions based on domain-specific knowledge stored in a formal knowledge base. Boaz and Shahar (2003) discusses the need for a

temporal-abstraction database mediator to provide a useful method for “*querying* not only *raw data*, but also its *abstractions*”. In the research presented in this paper we need to data mine the abstractions, rather than just query them. IDAN (Boaz and Shahar, 2003) is a temporal abstraction mediator which uses the generic temporal abstraction system ALMA (Boaz and Shahar, 2005) for its temporal reasoning task, and ALMA uses KBTA/Temporal Abstraction Rule (TAR) language (Balaban et al., 2003, Boaz et al., 2003) and CAPSUL. IDAN is used by multiple applications; KNAVE-II (Boaz and Shahar, 2005) and DeGeL are examples.

Recent research abstracts multidimensional time series data to produce alerts when certain trends are detected (Stacey et al., 2007). Currently, the rules for detecting these trends are human-defined; however, there may be as yet undiscovered trends and patterns that could indicate the onset of some condition, found by analysing historical data. Opportunities exist to apply data mining to temporally-abstracted cross correlated historical time series data of previous NICU patients, to identify new patterns and trends that may be of significance in the early identification of the onset of medical conditions in new NICU patients. These trends and patterns can be used to create rules for clinical alert systems within NICU monitoring equipment. When dealing with time series data or temporal data, data mining is rarely straightforward. Some pre-processing of the data usually needs to take place. Temporal data mining is an important extension to data mining and is discussed in the paper “A survey of temporal knowledge discovery paradigms and methods” (Roddick and Spiliopoulou, 2002). Many approaches to temporal data mining are covered; however the problem of providing a flexible environment to support various temporal data mining studies on multidimensional data streams is not discussed. The paper discusses some interesting systems that appear to partially address the areas of interest to the research presented in this paper. For example, the RX project uses temporal data to discover causal relationships. Also of interest for our research is the discussion on sequence mining and SDL (Shape Definition Language). Duchene et al (2007) has developed a prototype system to be used in the area of home health telemedicine. Although this is a different environment from the intensive care unit setting in terms of data rates and types of data, the system they developed is of interest to this research due to the way the data is pre-processed using temporal abstraction before being mined. The prototype system is mining heterogeneous multivariate time-series data for a patient to discover and learn usual patterns for that particular patient. The purpose of the system is to be able to detect changes in the pattern profile, which can indicate a problem for the patient at home. In the system the focus is data for only on *one* patient at a time. The research presented in this paper will extend this concept to mining across multiple parameters for multiple patients to discover trends that can be indicative of the onset on some condition.

A review of papers related to temporal abstraction in intelligent data analysis (IDA) with particular emphasis on abstraction of multivariate/multidimensional data,

and particularly those papers combining research in both temporal abstraction and data mining as applied to clinical data has been done by Stratti (2008). The review found that there is an absence of flexible applications and frameworks for data mining of multidimensional time series data. The applications using temporal abstractions as a pre-processor to such data was aimed at a specific study, rather than as a flexible framework for many varied studies on the collected data. Clinical environments such as ICU deal with high frequency, high volume clinical and physiological data from monitoring equipment (Verduijn et al., 2007, Tusch G., 2007, Azulay et al., 2007, Moskovitch et al., 2007, Silvent et al., 2004), whereas others may deal with low frequency data such as test results over time (Ho et al., 2004, Abe and Yamaguchi, 2005, Post and Harrison, 2007). Two papers reviewed considered both high and low frequency data in a multi-stream environment (Verduijn et al., 2007, Azulay et al., 2007), and one of these (Verduijn et al., 2007) also considered real-time data. Three of the remaining papers dealt only with high frequency multi-stream data (Moskovitch et al., 2007, Silvent et al., 2004, Charbonnier and Gentil, 2007), and the remaining papers utilised low frequency data (Tusch G., 2007, Abe and Yamaguchi, 2005, Sacchi et al., 2007, Post and Harrison, 2007). Only Bellazzi et al. (2005) is working with distributed data. With all the papers reviewed, each of the papers had a particular study as the motivation for the data collection, and hence, none of the papers created an environment for flexible exploration to support different clinical research problems. The creation of a flexible environment to support different clinical research problems was identified as an open research area.

A variety of techniques were used for the temporal abstractions. A data driven approach to temporal abstraction was used by Azulay (2007) and Moskovitch (2007). Verduijn et al (2007) utilised qualitative temporal abstractions to create state and trend abstractions. Sacchi et al (2007) uses Shahar's (1997) knowledge based approach; KBTA. Four of the papers discussed creation of complex abstractions (Silvent et al., 2004, Post and Harrison, 2007, Bellazzi et al., 2005, Charbonnier and Gentil, 2007).

Heath (2006) argues that for knowledge discovery in data (KDD) and data mining results to be accepted by clinicians and the medical community, adaption must be made to introduce more rigor in the form of scientific-method approach into the process, and to include provisions for hypothesis creation and null hypothesis testing within the framework. According to Heath (2006), clinicians are sceptical of data mining (DM) results, largely because current frameworks do not support the scientific method of Null Hypothesis testing. The null hypothesis is usually created to be demonstrated as incorrect, in order to support an alternative hypothesis. When used in medical experiments the null hypothesis is typically stated as there being no significant difference between compared groups. Null Hypothesis testing is used when conducting clinical trials, and Heath states that "the null hypothesis driven medical research paradigm must inform DM

investigative methods in the medical domain" (Heath, 2006).

Heath (2006) proposes the *extended* CRISP-DM model as a solution to this issue. This model uses exploratory data mining as a tool to find unknown patterns or relationship and creating hypotheses. Confirmatory data mining is subsequently used for null hypothesis testing.

Of the papers reviewed there was none that explicitly discuss null hypothesis testing, however one paper does discuss creating hypotheses to be evaluated (Abe and Yamaguchi, 2005). As a result the incorporation of null hypothesis testing within the data mining framework was recognised as an open research area. To enable the use of the *extended* CRISP-DM when discovering new trends and patterns in neonatal physiological data, the challenge existed to further extend this model to include provisions for temporal abstractions and for use on multidimensional parameters. Alignment for condition onset prediction is also not discussed in any of the systems and hence these represent open research areas. There exists a challenge in developing systems that will enable exploratory data mining on high frequency, heterogeneous, multi-parameter, historical time series data, when incorporating temporal abstraction as a preprocessing method so as not to lose the information of time and context during the mining process. To enable the discovery of new trends and patterns that may be indicative of the onset of a condition in intensive care patients where the timing of certain events in a patient's condition can be of high importance, there is a need for integrated temporal abstraction data mining systems to include methods to enable realignment of historical data in relation to the onset of the condition being investigated. As a result of the investigations from the literature the TAMDDM framework was designed.

3 Architecture

The proposed TAMDDM framework architecture is aiming to bridge the gap between clinical management and clinical research, enabling the secondary use of some of the vast amount of data collected by monitoring equipment attached to a baby in the NICU.

The framework has three layers (figure 1); the multi-agent system layer which drives the framework, the data mining model layer defining the data mining tasks, and the specific TAMDDM framework task layer.

The framework utilizes components from Foster's (Foster and McGregor, 2005) multi-agent system which has been extended to facilitate the tasks needed in the TAMDDM framework. Heath (2006) has previously extended the CRISP-DM data mining model to facilitate null hypothesis testing. The *extended* CRISP-DM model is integrated into the extended multi-agent framework to provide the data mining model to complete the tasks of the TAMDDM framework, as shown in figure 1.

Foster's multi agent system consists of an Agent Server, Processing Agent, Functional Agent, Sub Agents, Human Agent, and a Database Access Server. For descriptions of each of these please see Foster and McGregor (2005).

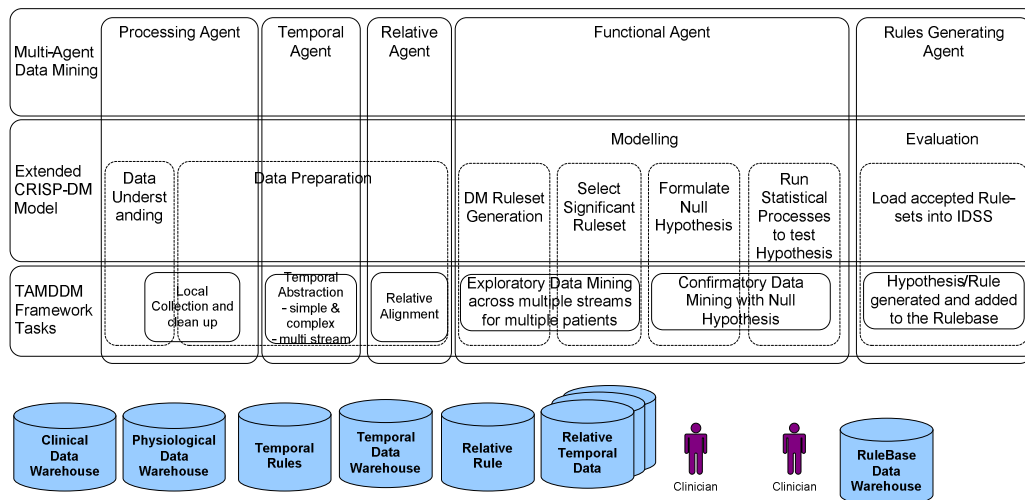


Figure 1: The TAMDDM Framework

The agents utilized by the TAMDDM framework are the processing agent, functional agent and rules generating agent. The system also utilizes the database access server. To enable its use in the TAMDDM framework the multi-agent system is extended to include two new agents, namely a temporal agent and a relative agent. In addition, the functionality of the functional agent has changed to incorporate mining of temporal data and has been expanded to support null hypothesis testing to support research utilising the scientific method as described by Heath (2006).

The following sections describe the TAMDDM framework by describing the tasks completed by the agents, with the main focus on the temporal, relative and functional agent.

3.1 Processing Agent

In the TAMDDM framework the processing agent acts as a pre-processor for the functional agent, performing the tasks of getting and preparing the data from external databases and storing it within the physiological data store and/or the clinical data store ready for further processing by the temporal agent. This agent is used to support and partially support the phases of Data Understanding and Data Preparation in the *extended* CRISP-DM model.

3.2 Temporal Agent

The temporal agent is a new agent added to the multi-agent system. The temporal agent processes new physiological data entering the framework, creating temporal abstractions as defined by temporal rules in the system. The temporal abstraction process is a preprocessing method before data mining which allows the temporal aspects and the context of the data to be preserved.

For every patient each of the physiological streams is temporally abstracted into appropriate abstractions such as trends (increasing, decreasing) and level shifts (high, low). Each raw piece of data may belong to several abstractions. For example, a particular measurement may

be part of an 'increasing' abstraction, and at the same time be within 'normal' limits. Complex abstractions can also be done across multiple abstracted parameters. Each abstraction, including actual start and end times for the particular abstraction instance, is stored in the TAMDDM's Temporal Data Warehouse.

The temporal agent has six main functions:

- 1) Retrieve the physiological data from the physiological data store for each parameter for each patient
- 2) Retrieve the relevant abstraction rules from the temporal rules table
- 3) Apply the rules to the physiological data, creating simple abstractions for individual data streams for individual patients
- 4) Store the created abstractions in the TAMDDM's temporal data store
- 5) Create complex abstractions from the simple abstractions created in step 3, according to any rules found in the temporal rules table.
- 6) Store any complex abstractions created in the TAMDDM's temporal data store.

The temporal abstraction tasks are part of the data preparation phase of the *extended* CRISP-DM model.

3.3 Relative Agent

Each of the abstractions created by the temporal agent from the physiological data can be part of many clinical research studies. Once created, the abstractions are stored in the TAMDDM's data stores until needed for a particular study. The next agent in the TAMDDM framework, the relative agent, is not invoked until a particular clinical research study is to be completed. The relative agent is a new agent added to the multi-agent system. The relative agent uses the abstractions created by the temporal agent, together with clinical information of individual patients.

When a particular study is prepared, it will often be necessary to realign the time of abstractions relative to a particular point in time of interest, such as diagnosis. Using absolute times for the start and end time of

abstractions give no indication of what time this abstraction takes place in relation to the diagnosis. A mechanism to enable relative timing is needed, and this need led to the design of the relative agent for this task. The relative agent calculates start and finish times for each abstraction relative to a particular event, such as time of diagnosis. This calculation will occur for every abstraction for each data stream for each patient taking part in the particular study.

The relative agent has three main functions:

- 1) Retrieve the relevant data and temporal abstractions from the TAMDDM framework's data store, based on the selection specifications given by the clinician/researcher.
- 2) Applying the transformations specified for the study to be undertaken to the absolute timed temporal abstractions to create the set of aligned temporal abstractions, called relative abstractions, as time (start and end times) is relative to the alignment point.
- 3) Store the relatively aligned abstractions in the relative temporal data store to allow for further processing by the functional agent.

When researching a particular condition, the abstractions will be matched with the diagnosis table holding the patient's diagnosis time and date. This information will be fed through a transform algorithm to enable a measurement in time for the abstractions relative to the point in time of the diagnosis. T_0 will be the point of diagnosis, and $T_{-1}, T_{-2}, T_{-3}, \dots, T_{-n}$ will indicate the distance in time between an abstraction before the time of diagnosis, and the diagnosis.

Many studies can be conducted on the same temporal abstractions; therefore the same temporal abstractions may require realignment in several different ways. Each re-aligned temporal abstraction stored in the relative temporal data table will belong to a particular study. The realigned temporal abstractions will form the basis for the exploratory and confirmatory data mining performed in later stages of the process.

3.4 Functional Agent

The realigned temporal abstractions created by the relative agent are further processed by the functional agent. For the TAMDDM framework the functional agent is extended from Foster's design (Foster and McGregor, 2005) to include exploratory and confirmatory data mining. The functional agent is used to facilitate the modeling tasks of the *extended* CRISP-DM model, including rule set generation through exploratory data mining, selecting significant rule sets, null hypothesis formulation and running statistical processes to test the null hypothesis during confirmatory data mining.

Exploratory data mining is used to analyse the realigned temporal abstractions, across multiple data streams for multiple patients, to detect new trends and patterns. Significant rule sets are selected from the results of the exploratory mining for further analysis. This is part of the data mining rule set generation phase of the *extended* CRISP-DM model.

After exploratory data mining has been employed and significant rule sets have been selected, the 'confirmatory data mining with null hypothesis' task begins with

formulating the null hypothesis for any results that indicate interestingness and further investigation. Once the null hypothesis has been defined, confirmatory data mining is employed on the data abstractions with the aim of either proving or disproving the null hypothesis. The 'run statistical processes to test null hypothesis' phase of the *extended* CRISP-DM model performs the confirmatory data mining with null hypothesis task of the TAMDDM framework, aiming to prove or disprove the null hypothesis.

If the null hypothesis is disproven for a particular rule set, a hypothesis is put forward about the rule set and further testing can be initiated if desired before committing the new rule set to the rule base.

3.5 Rules Generating Agent

The purpose of investigating the historical data is to possibly find new hypotheses that can be defined as rules to be used for intelligent patient monitoring in the NICU. The rules generating agent is used for converting findings made by the functional agent into rules that can be inserted into the rules database (Foster and McGregor, 2005). The TAMDDM framework will use the rules generating agent to convert hypotheses created by the exploratory data mining into rules that can be stored and utilized by an event stream processor in neonatal monitoring.

Once the confirmatory data mining is concluded, if the null hypothesis has been disproven a clinician/researcher will evaluate the hypotheses produced and decides if these are to be incorporated into the rule base used by an intelligent patient monitoring system in the NICU. This task is part of the 'load accepted rule sets into IDSS' task of the Evaluation phase in the *extended* CRISP-DM model, and is performed by the Rules Generating Agent.

4 Demonstration

This section presents a demonstration of TAMDDM, within the context of its application to support clinical research within neonatal intensive care.

A description is provided of how historical neonatal physiological and clinical data moves through the system, starting as raw unprocessed data, the transformation of this data into qualitative temporal abstractions, realignment of these abstractions in preparation for analysis and finally hypothesis creation through exploratory and confirmatory data mining.

The TAMDDM framework is a flexible framework not designed for any particular study or research project, but to enable multiple varied studies on the historical data. The first two agents, processing and temporal, are independent of any particular case study and are invoked on collection of new data, whereas the remaining agents are run for each case study/research project.

The course of action through the first two agents, the processing agent and the temporal agent, is the same for all data entering the system. These two agents perform part of the pre-processing necessary for using the physiological data in clinical research to discover new trends and patterns in the data indicative of some condition. The course of action through the processing

and temporal agent happens as raw time series and clinical data is entering the system. Once this procedure is finished those agents are not re-invoked unless there is new data entering the system.

Once the data has been processed by the processing and temporal agent, the temporal abstractions created are stored in the TAMDDM's data store until a particular clinical study is taking place. The particular study determines the further course of action for the temporal abstractions. First the abstractions are re-aligned relatively to a particular time or event of interest such as date/time of birth, date/time of diagnosis or other event. Once the abstractions are realigned the functional agent performs the exploratory and confirmatory data mining on the realigned abstractions. If the result of this process has created accepted hypotheses, the rule generating agent can be invoked to transform the accepted hypotheses into rules. These rules will be stored in the rule base external to the TAMDDM framework. This rule base can be accessed by an intelligent monitoring and alerting system for real-time patient alerting.

Due to space limitations, the demonstration section is mainly focusing on the new and innovative additions to data mining frameworks that this research presents; the abstraction and re-alignment tasks.

4.1 Temporal Abstraction

After the processing agent has placed the data in the local data stores within the TAMDDM framework, the stored data is processed by the Temporal Agent. The temporal agent uses the rules defined in the temporal rules table in the database within the TAMDDM framework to create temporal abstractions from the physiological data that has been collected from monitoring equipment.

The data for each patient consist of multiple data streams. Each of these streams of time-stamped physiological readings is abstracted separately into simple temporal abstractions. These simple abstractions can then be used to create complex temporal abstractions; abstractions that could be a combination of two or more simple abstraction (see fig. 4). Abstractions can consist of level shifts (e.g., high/low/normal) or trends (e.g., increase/decrease/stable). A particular time-stamped physiological reading for a particular patient can be part of several simple abstractions. It may be that it is appropriate to abstract for both level shifts and trends.

Let's consider an example for arterial oxygen saturation (SaO₂). Normal values for SaO₂ is considered as 90% or above (House et al., 1987), so it would be natural to have a level shift abstraction rule for the SaO₂ data stream, where continuous intervals of SaO₂ values at or above 90% are made into a 'normal' abstraction, and continuous intervals of SaO₂ values below 90% are made into a 'low' abstraction. Table 1 contains raw SaO₂ readings for a patient. Creating a graph of the SaO₂ values in Table 1 against the 90% threshold (Figure 2), makes it apparent that the values in the table above can be reduced into several normal and low abstractions (level shift abstractions).

Patient_Id	DateTime	Phys_Id	Value
102	20061116_11:29:02.001	4	90.1
102	20061116_11:29:02.002	4	90
102	20061116_11:29:02.003	4	89.9
102	20061116_11:29:02.004	4	89.8
102	20061116_11:29:02.005	4	89.9
102	20061116_11:29:02.006	4	89.9
102	20061116_11:29:02.007	4	89.9
102	20061116_11:29:02.008	4	90
102	20061116_11:29:02.009	4	90
102	20061116_11:29:02.010	4	90
102	20061116_11:29:02.011	4	89.8
102	20061116_11:29:02.012	4	89.9
102	20061116_11:29:02.013	4	89.9
102	20061116_11:29:02.014	4	89.9
102	20061116_11:29:02.015	4	89.8
102	20061116_11:29:02.016	4	90
102	20061116_11:29:02.017	4	90
102	20061116_11:29:02.018	4	89.9
102	20061116_11:29:02.019	4	89.9
102	20061116_11:29:02.020	4	90
102	20061116_11:29:02.021	4	90.1
102	20061116_11:29:02.022	4	90.5
102	20061116_11:29:02.023	4	90.9
102	20061116_11:29:02.024	4	91.1
102	20061116_11:29:02.025	4	91
102	20061116_11:29:02.026	4	90.8

Table 1: Raw SaO₂ readings

For SaO₂, readings of 90 and above are seen as normal, and readings below 90 can be problematic. This threshold is indicated by the stippled red line at 90. The rule for this particular abstraction, using 90 as a threshold as discussed above, will be:

Low = SaO₂ < 90

Normal = SaO₂ ≥ 90

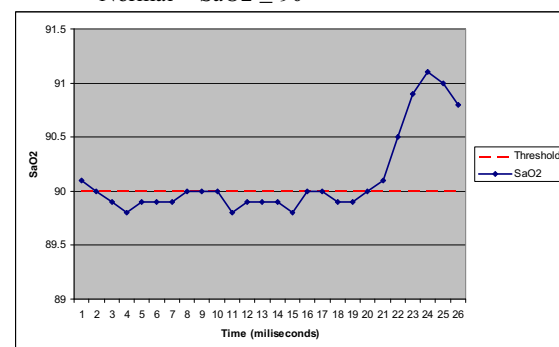


Figure 2: Graph SaO₂ against a threshold of 90%

Here, we can see that the two first readings in figure 2 are within the normal range with a start at 1 and end at 2 and would create a 'normal' abstraction. The next five readings are below the 90% threshold and therefore would create a 'low' abstraction, starting at 3 and finishing at 7.

The abstractions created are stored in the TemporalAbstraction table in the TAMDDM framework (Table 2). As can be seen from the table below, abstracting the readings also condenses the data:

Abstraction Value	ActualStartTime	ActualEndTime
Normal	20061116_11:29:02.001	20061116_11:29:02.002
Low	20061116_11:29:02.003	20061116_11:29:02.007
Normal	20061116_11:29:02.008	20061116_11:29:02.010
Low	20061116_11:29:02.011	20061116_11:29:02.015
Normal	20061116_11:29:02.016	20061116_11:29:02.017
Low	20061116_11:29:02.018	20061116_11:29:02.019
Normal	20061116_11:29:02.020	20061116_11:29:02.026

Table 2: Abstractions created (Partial table) from all SaO2 readings in Table 1: Raw SaO2 readings

Another abstraction is created for blood pressure. Table 3 shows the blood pressure values for the same patient and the same timeframe as the SaO2 example above:

Patient_Id	DateTime	Phys_Id	Value
102	20061116_11:29:02.001	6	23.8
102	20061116_11:29:02.002	6	23.1
102	20061116_11:29:02.003	6	22.9
102	20061116_11:29:02.004	6	23.5
102	20061116_11:29:02.005	6	23
102	20061116_11:29:02.006	6	24
102	20061116_11:29:02.007	6	24.2
102	20061116_11:29:02.008	6	23.5
102	20061116_11:29:02.009	6	24.5
102	20061116_11:29:02.010	6	24.4
102	20061116_11:29:02.011	6	24
102	20061116_11:29:02.012	6	24.8
102	20061116_11:29:02.013	6	24.1
102	20061116_11:29:02.014	6	24.7
102	20061116_11:29:02.015	6	24.3
102	20061116_11:29:02.016	6	23.5
102	20061116_11:29:02.017	6	23.9
102	20061116_11:29:02.018	6	23.3
102	20061116_11:29:02.019	6	23.7
102	20061116_11:29:02.020	6	23.9
102	20061116_11:29:02.021	6	23.3
102	20061116_11:29:02.022	6	23.8
102	20061116_11:29:02.023	6	23.2
102	20061116_11:29:02.024	6	23.3
102	20061116_11:29:02.025	6	23.9
102	20061116_11:29:02.026	6	23.7

Table 3: Blood pressure values

The rule used to abstract the blood pressure parameter is:

Low = BP < 24

Normal = BP ≥ 24

Graphing the blood pressure values against a threshold of 24 is illustrated in figure 3. The abstractions are recorded in table 4.

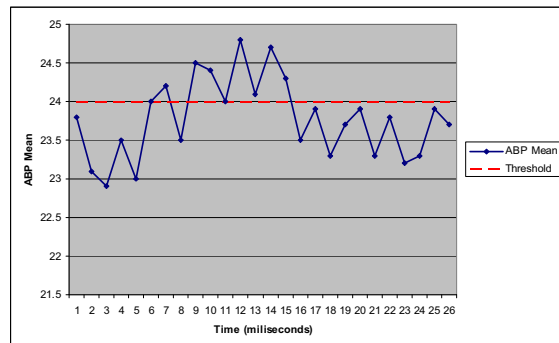


Figure 3: Graphing of blood pressure values against a threshold of 24mm/Hg

Abstraction Value	ActualStartTime	ActualEndTime
Low	20061116_11:29:02.001	20061116_11:29:02.005
Normal	20061116_11:29:02.006	20061116_11:29:02.007
Low	20061116_11:29:02.008	20061116_11:29:02.008
Normal	20061116_11:29:02.009	20061116_11:29:02.015
Low	20061116_11:29:02.016	20061116_11:29:02.026

Table 4: Level shift abstractions (Partial table) created from all blood pressure readings in Table 3: Blood pressure values

Complex abstractions can be created from simple abstractions such as those created above for blood pressure and blood oxygen saturation. For example a complex abstraction can be specified where the two abstraction rules above are combined, and create complex abstractions only when both blood pressure and blood oxygen saturation is below their respective thresholds. The rule that must hold true for this example is $SaO_2 < 90$ AND $BP < 24$, meaning only intervals where both these conditions are true are of interest for this particular complex abstraction. The diagram below (figure 4) illustrates the two physiological parameters as the graphs against time in milliseconds. The top graph shows arterial oxygen saturation, and the bottom graph shows blood pressure. The green lines indicate the time spans of the low abstractions for each parameter. The low abstractions show what is of interest to this example. As the complex abstractions to be created consist of time periods where there is an overlap of the low level shift abstractions for both parameters, the full time intervals from the simple abstractions are not always included.

The time points which satisfy the complex abstraction rule are marked with red circles in Figure 4. There are only two time intervals in this section of monitoring data that can be used for the complex abstractions where both parameters have a low abstraction at the same time.

These points are from t_3 - t_5 inclusive, and t_{18} - t_{19} inclusive. These abstractions are summarized in Table 5.

As can be seen there are only two intervals in the sample data that satisfy the criteria for this complex abstraction.

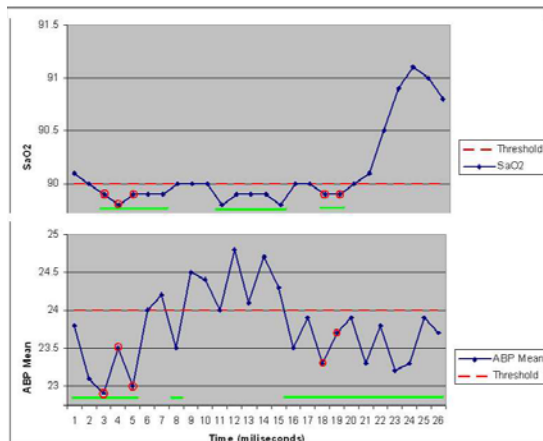


Figure 4: Complex Abstraction

Once this abstraction process is completed, the created abstractions are stored for use in future clinical studies.

Abstraction Value	ActualStartTime	ActualEndTime
LowSaO2BP	20061116_11:29:02.003	20061116_11:29:02.005
LowSaO2BP	20061116_11:29:02.018	20061116_11:29:02.019

Table 5: Complex Abstractions

4.2 Relative Alignment

When a particular research study is taking place, the relative agent is invoked. Consider a hypothetical study of babies who experience cardiac arrest. A clinician wishes to conduct a study to learn if there exist simultaneous variations in the arterial oxygen saturation (SaO2) parameter and blood pressure parameter in the time leading up to the incidents of cardiac arrest. The variations considered are consistent with the complex abstraction created by the temporal agent in the previous section, namely:

$$\text{SaO2} < 90 \text{ AND } \text{BP} < 24$$

To enable the detection of particular patterns of this abstraction at a particular time before the cardiac arrest, re-alignment of the abstractions relative to the time of cardiac arrest is necessary.

In the first diagram in figure 5, the lines going out from each baby indicates various abstracted data streams; multi-dimensional data. The green boxes on the group of lines for each baby indicate groups of some unusual behaviour in the physiological data. These intervals would be abstracted by the temporal agent, and are now stored in the TAMDDM's data stores. As these abstractions are using absolute time for the start and finish time for each abstraction, it will usually be necessary to give these abstractions start and finish times relative to a particular event that is of interest, such as the time of diagnosis. This will enable the comparison and mining of the abstractions, allowing the distance from diagnosis (or other event) to be taken into account. It may be, as in Figure 5, the physiological data is exhibiting a

particular kind of behavior a certain time before a particular diagnosis is made.

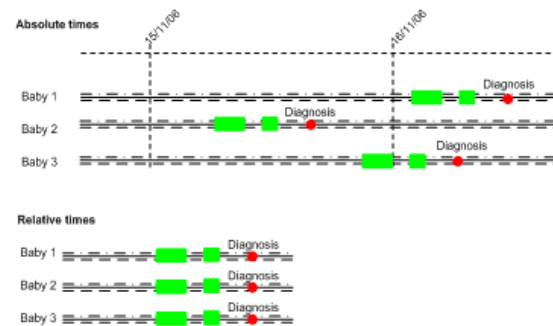


Figure 5: Realignment of abstracted parameters relative to diagnosis.

This particular behavior will be apparent in the temporal abstractions created, however if we mine these abstractions each of these indicators above will appear with different timing and will therefore not indicate that there is a common behavior in the physiological data at a particular time interval *before* diagnosis for the patients. To overcome this issue the TAMDDM framework uses the relative agent to realign the abstractions created from the physiological data for the babies who are taking part in a particular study. This realignment will use as basis a particular event of interest, and this can be different depending on the study. In this example the event is when a baby is given a diagnosis for a particular condition.

As can be seen from figure 5, the unusual patterns in the data and the diagnosis for baby 2 occurs after each other on 15/11/06. For baby 1, this occurs the following day, and for baby 3, some of the unusual patterns occur on the 15th and some on the 16th. When the abstractions for these unusual patterns are given start and end times relative to the diagnosis, the abstractions from the babies' data streams can be re-aligned relative to the diagnosis event, as shown in the bottom diagram in figure 5.

Consider the abstractions created for SaO2 in the previous section. The first abstraction has a start time of 20061116_11:29:02.001 and a finish time of 20061116_11:29:02.002. This patient was diagnosed with the condition being researched exactly one hour after the start time of the first abstraction. To relate this patient's abstracted data to the time of diagnosis, we are interested in when the abstraction was valid in relation to the time of diagnosis. As the diagnosis was exactly one hour after the start time of the first abstraction recorded, the relative abstraction start time will be 00000000_01:00:00.000, exactly one hour before diagnosis. The relative times are created by calculating the difference between the actual times and the time of diagnosis (or other event of interest). The end time for this abstraction was 20061116_11:29:02.002, which gives a relative end time of 00000000_00:59:59.999. Table 6 contains the relative temporal abstractions for this particular example, using diagnosis time as the event of interest for patient data realignment. When the event of interest is a particular diagnosis, we are interested in data in the time before and up to the diagnosis.

Abstract onValue	RelativeStartTime	RelativeEndTime	Study Id
Normal	00000000_01:00:00.000	00000000_00:59:59.999	119
Low	00000000_00:59:59.998	00000000_00:59:59.994	119
Normal	00000000_00:59:59.993	00000000_00:59:59.991	119
Low	00000000_00:59:59.990	00000000_00:59:59.986	119
Normal	00000000_00:59:59.985	00000000_00:59:59.984	119
Low	00000000_00:59:59.983	00000000_00:59:59.982	119
Normal	00000000_00:59:59.981	00000000_00:59:59.975	119

Table 6: relative temporal abstractions

There is no need to create relative changes to abstractions for data after the time of the particular diagnosis for the individual patients. If we are looking for trends before a certain event, the data after that event is not of interest and should not be realigned.

Once the relative agent has completed the relative alignment, control is passed to the functional agent.

4.3 Data Mining

Considering the example mentioned in the previous section, where a clinician is investigating the relationship between variations in the arterial oxygen saturation (SaO₂) parameter and blood pressure parameter in the time leading up to the incidents of cardiac arrest. First exploratory data mining will be employed to find any new hypotheses. This could be done using rule association mining. Exploratory data mining may produce several significant rule sets that needs null hypothesis testing. A domain expert must select each set, create a hypothesis and confirm the hypothesis through null hypothesis testing using confirmatory data mining performed by the functional agent. Once completed, the hypothesis (if found valid) is passed to the rules generating agent for processing.

If exploratory mining found a link between both arterial oxygen saturation and blood pressure simultaneously being low for 20 seconds or more and the event of cardiac arrest, a hypothesis could be formulated:

$$(\text{SaO}_2 < 90 \text{ AND BP} < 24) \geq 20 \text{ seconds} \Rightarrow \text{Cardiac Arrest}$$

Once a hypothesis is formulated, a null hypothesis can be formulated and tested. A null hypothesis in this case would state that there is no trend of arterial oxygen saturation dropping below 90 and blood pressure dropping below 24 for 20 seconds or more before patients experience cardiac arrest. If confirmatory mining proves the null hypothesis test to be right, there is no need to continue the process. If however the result of the confirmatory data mining is that these changes in SaO₂ and blood pressure can predict the onset of cardiac arrest, the null hypothesis is disproven and further investigations can be made. The clinician may decide that the hypothesis is mature enough to be used as a rule for an intelligent monitoring system, or decide that further investigation is warranted. If the former is decided, the hypothesis is passed to the rules generating agent.

The rules generating agent processes the hypotheses created by the functional agent into appropriate rules that can be stored in the rule base to be used by an intelligent

monitoring and alerting system such as the monitoring system created by Stacey (2007).

5 Conclusion and further research

The TAMDDM framework is an innovative framework for mining multidimensional temporal data, incorporating null hypothesis testing to allow clinical research to be conducted on historical physiological and clinical data. The research presented in this paper has certain limitations that require further research and development. Monitoring data from patient monitoring typically has invalid data known as artifact, for example when the patient is moved, or the transducers attached to the body is moved. Further development has commenced to incorporate artifact identification and appropriate processing of that data within the data mining.

Currently the storage of the data stream data within the TAMDDM framework is not standard based, as such standards are absent. The monitoring devices generate enormous amounts of data and better standards based storage methods are currently being investigated.

Local data storage is assumed within TAMDDM and further research has commenced to support distributed environments utilising the service oriented architecture, where access to agents is exposed through web services. In this way we can support multi-centre retrospective studies. This extended framework providing Service-oriented Multi-Dimensional Temporal Data Mining supporting null hypothesis testing is known as STDMⁿ. STDMⁿ is being utilised as part of clinical research investigating condition onset indicators for nosocomial infection in neonates in project Artemis (McGregor et al., 2009).

6 References

- Abe, H. and Yamaguchi, T. (2005) Implementing an Integrated Time-Series Data Mining Environment - A Case Study of Medical KDD on Chronic Hepatitis -. *Fist International Conference on Complex Medical Engineering (CME2005)*. Takamatsu, Kagawa, Japan.
- Antunes, C. and Oliveira, A. (2001) Temporal Data Mining: an overview. *Workshop on Temporal Data Mining at the 7th International Conference on Knowledge Discovery and Data Mining (KDD'01)*.
- Augusto, J.C. (2005) Temporal reasoning for decision support in medicine. *Artificial Intelligence in Medicine*, 33, 1-24.
- Azulay, R., Moskovitch, R., Stopel, D., Verduijn, M., Jonge, E.d. and Shahar, Y. (2007) Temporal Discretization of medical time series - A comparative study. *IDAMAP 2007 workshop*. Amsterdam.
- Balaban, M., Boaz, D. and Yuval, S. (2003) Applying Temporal Abstraction in Medical Information Systems. *Ann Math Comput Teleinform*, 1, 56-64.
- Bellazzi, R., Larizza, C., Magni, P. and Bellazzi, R. (2005) Temporal data mining for the quality assessment of hemodialysis services. *Artificial Intelligence in Medicine*, 34, 25-39.
- Boaz, D., Balaban, M. and Shahar, Y. (2003) A Temporal-Abstraction Rule Language for Medical Databases. *IDAMAP '03*.

- Boaz, D. and Shahar, Y. (2003) Idan: A Distributed Temporal-Abstraction Mediator for Medical Databases. *Artificial Intelligence in Medicine-Europe (AIME)*.
- Boaz, D. and Shahar, Y. (2005) A framework for distributed mediation of temporal-abstraction queries to clinical databases. *Artificial Intelligence in Medicine*, 34, 3-24.
- Chakravarty, S. and Shahar, Y. (2000) CAPSUL: A constraint-based specification of repeating patterns in time-oriented data. *Annals of Mathematics and Artificial Intelligence*, 30, 3-22.
- Charbonnier, S. and Gentil, S. (2007) A trend-based alarm system to improve patient monitoring in intensive care units. *Control Engineering Practice*, 15, 1039-1050.
- Duchene, F., Garbay, C. and Rialle, V. (2007) Learning recurrent behaviors from heterogeneous multivariate time-series. *Artificial Intelligence in Medicine*, 39, 25-47.
- Foster, D. and McGregor, C. (2005) Overview of an Agent-based IDSS Framework for Neonatal Analysis and Trend Detection. *National Health Informatics Conference (13th : 2005 : Melbourne, Vic.)*. Brunswick East, Vic., Health Informatics Society of Australia.
- Fuchsberger, C., Hunter, J. and McCue, P. (2005) *Testing Asbru Guidelines and Protocols for Neonatal Intensive Care*.
- Griffin, M.P., Lake, D.E., O'Shea, T.M. and Moorman, J.R. (2007) Heart Rate Characteristics and Clinical Signs in Neonatal Sepsis. *Pediatric Research*, 61, 222-227.
- Heath, J. (2006) A Framework for an Intelligent Decision Support System (IDSS), Including a Data Mining Methodology, for Fetal-Maternal Clinical Practice and Research. *School of Computing and Mathematics*. Sydney, University of Western Sydney.
- Ho, T.B., Kawasaki, S., SQuang, L., Takabayashi, K. and Yokoi, H. (2004) Combining temporal abstraction and data mining to study hepatitis data. *SIG-KBS*, 64, 63-68.
- House, J.T., Schultetus, R.R. and Gravenstein, N. (1987) Continuous neonatal evaluation in the delivery room by pulse oximetry. *Journal of Clinical Monitoring and Computing*, 3, 96-100.
- Kramer, M.S., Platt, R., Yang, H., Joseph, K.S., Wen, S.W., Morin, L. and Usher, R.H. (1998) Secular Trends in Preterm Birth: A Hospital-Based Cohort Study. *JAMA*, 280, 1849-1854.
- Laws, P., Abeywardana, S., Walker, J. and Sullivan, E.A. (2007) Australia's mothers and babies 2005. *Perinatal Statistics*. Sydney, AIHW National Perinatal Statistics Unit.
- McGregor, C., Sow, D., James, A., Blount, M., Ebling, M., Eklund, J.M. and Smith, K. (2009) Collaborative Research on an Intensive Care Decision Support System utilizing Physiological Data Streams.
- Moskovitch, R., Stopel, D., Verduijn, M., Peek, N., Jonge, E.d. and Shahar, Y. (2007) Analysis of ICU Patients Using the Time Series Knowledge Mining Method. *IDAMAP 2007 Workshop*. Amsterdam.
- O'Connor, M.J., Grosso, W.E., Tu, S.W. and Musen, M.A. (2001) RASTA: a distributed temporal abstraction system to facilitate knowledge-driven monitoring of clinical databases. *Medinfo*.
- Post, A.R. and Harrison, J.H., Jr. (2007) PROTEMPA: A Method for Specifying and Identifying Temporal Sequences in Retrospective Data for Patient Selection. *J Am Med Inform Assoc*, 14, 674-683.
- Roddick, J.F. and Spiliopoulou, M. (2002) A survey of temporal knowledge discovery paradigms and methods. *Knowledge and Data Engineering, IEEE Transactions on*, 14, 750-767.
- Sacchi, L., Larizza, C., Combi, C. and Bellazzi, R. (2007) Data mining with Temporal Abstractions: learning rules from time series. *Data Mining and Knowledge Discovery*, 15, 217-247.
- Seyfang, A. and Miksch, S. (2004) Advanced Temporal Data Abstraction for Guideline Execution. *Stud Health Technol Inform*, 101, 88-102.
- Seyfang, A., Miksch, S., Horn, W., Urschitz, M.S., Popow, C. and Poets, C.F. (2001) Using Time-Oriented Data Abstraction Methods to Optimize Oxygen Supply for Neonates. *Lecture Notes in Computer Science*.
- Shahar, Y. (1997) A framework for knowledge-based temporal abstraction. *Artificial Intelligence*, 90, 79-133.
- Shahar, Y., Miksch, S. and Johnson, P. (1998) The Asgaard project: a task-specific framework for the application and critiquing of time-oriented clinical guidelines. *Artificial Intelligence in Medicine*, 14, 29-51.
- Silvent, A.-S., Dojat, M. and Garbay, C. (2004) Multi-level temporal abstraction for medical scenario construction. *International Journal of Adaptive Control and Signal Processing*, 19, 377-394.
- Stacey, M. and McGregor, C. (2007) Temporal abstraction in intelligent clinical data analysis: A survey. *Artificial Intelligence in Medicine*, 39, 1-24.
- Stacey, M., McGregor, C. and Tracy, M. (2007) An architecture for multi-dimensional temporal abstraction and its application to support neonatal intensive care. *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*.
- Stratti, H.B. (2008) A Framework for Temporal Abstractive Multidimensional Data Mining. *School of Computing and Mathematics*. Sydney, University of Western Sydney.
- Tusch G., O.C.M., Redmond T., Shankar R., Das A. (2007) SPOT - Utilizing Temporal Data for Data Mining in Medicine. *IDAMAP 2007 Workshop*.
- Verduijn, M., Sacchi, L., Peek, N., Bellazzi, R., de Jonge, E. and de Mol, B.A.J.M. (2007) Temporal abstraction for feature extraction: A comparative case study in prediction from intensive care monitoring data. *Artificial Intelligence in Medicine*, 41, 1-12.

Automatic sleep stage identification: difficulties and possible solutions

Sukhorukova, N^{1*} Stranieri, A¹ Ofoghi, B¹ Vamplew, P¹ Saleem, M¹ Ma, L¹ Ugon, A^{2,3} Ugon, J¹ Muecke, N¹ Amiel, H² Philippe, C² Bani-Mustafa, A¹ Huda, S¹ Bertoli, M¹ Lévy, P² Ganascia, J-G³

1 Centre for Informatics and Applied Optimisation, University of Ballarat, Australia

2 Tenon Hospital, Paris, France

3 Laboratoire d'Informatique de Paris 6, France

*Corresponding author: n.sukhorukova@ballarat.edu.au

Abstract

The diagnosis of many sleep disorders is a labor intensive task that involves the specialised interpretation of numerous signals including brain wave, breath and heart rate captured in overnight polysomnogram sessions. The automation of diagnoses is challenging for data mining algorithms because the data sets are extremely large and noisy, the signals are complex and specialist's analyses vary. This work reports on the adaptation of approaches from four fields; neural networks, mathematical optimisation, financial forecasting and frequency domain analysis to the problem of automatically determining a patient's stage of sleep. Results, though preliminary, are promising and indicate that combined approaches may prove more fruitful than the reliance on a single approach.

Keywords: Sleep stage identification, data mining.

1 Introduction

Sleep Stage Identification (SSI) is the first step in the process of modern sleep disorder diagnostics. Currently, the identification of stages 1, 2, 3, REM and Awake is performed manually using rules drafted for medical practitioners based on the frequency and amplitude of waves recorded during polysomnogram sleep sessions (PSG). A polysomnogram sleep session (PSG) includes measures of eye movement (EOG), brain wave fluctuations (EEG), heart rhythm (ECG), muscle activity (EMG), respiratory effort and other biophysiological characteristics while a patient is asleep.

SSI is a time consuming, manual process that requires a great deal of skill and expertise in scanning PSG graphs and applying SSI rules. Recent advances in computing performance has made computer-based automatic scoring of sleep stages very attractive.

Copyright © 2010, Australian Computer Society, Inc. This paper appeared at the Australasian Workshop on Health Informatics and Knowledge Management (HIKM 2010), Brisbane, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 108. Anthony Maeder and David Hansen, Eds. Reproduction for academic, not-for-profit purposes permitted provided this text is included.

However, sleep practitioners report that existing automated techniques are not accurate enough to be routinely used (Robert, Guilpin et al. 1998).

PSG scoring experts apply rules based on the visual appearance of frequencies and amplitudes of waves on screen rather than using quantitative data describing frequencies and amplitudes. A survey of existing automatic tools for SSI by Bashashati, Fatourehchi et al. (2007) and Rajeev and Gotman (2002) reveals that the majority of approaches apply signal processing (SP) methods (Bashashati, Fatourehchi et al. 2007), Artificial Neural Network (ANN) methods (Robert, Guilpin et al. 1998) or Wavelet Transformations (Virkkalaa, Hasan et al. 2007). Approaches based on Financial Forecasting, Mathematical Optimisation and Hidden Markov Models have been deployed with other time series data and could conceivably lead to accurate classifications of SSI.

Much of the challenge in automated SSI is due to the translation of open textured standards to mathematical models (Rajeev and Gotman 2002) and the dimension of the problem. Raw data for one patient for 10 hours results in a single file more than 300 MB large with over 3,600,000 observations. Further, over 65% of records are sleep stage 2 and less than 5% for sleep Stage 1 and 3. This adds to the complexity of the challenge. Further, PSG data contains a great deal of noise. With practice, experts are able to ignore noise to an extent that is challenging for automated scoring tools. In addition, SSI can be performed differently by two sleep practitioners with an 80% level of agreement.

Rules for SSI originally were standardized by Rechtschaffen and Kales (1968). Since then, the rules have been updated numerous times. The most recent version is reported by Iber, Ancoli-Israel et al. (2007). A comprehensive explanation on why this update was necessary can be found in Schulz (2008).

In SSI doctors rely on the visual presentation of waves. Recorded waves are signals and therefore it is natural to analyze them with existing SP techniques, especially given the theoretical advances in this field in recent decades. The drawback of this approach is that in many cases SP completely ignores manual scoring characteristics, which are not described in general scoring rules, but are often taken into account by medical doctors.

Since doctors are not experts in SP they learn the shapes of the waves through their visual characteristics rather than wave characteristics used in SP.

The main problem with ANN approaches is that the dimension of the problem challenges most learning algorithms. Several simplifications have been used including using fewer variables in order to overcome the dimensionality problem. However, simplified models are not accurate enough to meet the needs of sleep disorder specialists. Problems associated with the use of ANN and SP approaches suggest the need to explore combinations of ANN and SP with other approaches.

In this study, approaches based on financial forecasting, mathematical optimisation, frequency domain analysis and neural networks have been adapted for SSI. The approaches have been applied to data supplied and classified from overnight sleep records from 100 patients from the Tenon Hospital sleep research group in Paris. Each approach is described and results presented in the sections below, before providing a cross-approach analysis and concluding remarks.

2 Neural network approach

ANN is a network composed of artificial nodes that process input activation for transmission to connected nodes. Input vectors to the ANN are treated as a temporal sequence whose analysis requires consideration of a set of prior input vectors. (Waibel, Sawai, et al. 1989) used Time-Delay Neural Networks (TDNNs) for speech recognition. The delay-based methodology of TDNNs, which reduces the high dimensionality of the input data to the network, is very important in the SSI, due to the length of input sequences.

A TDNN is a type of dynamic ANN where the output of the network at time t_i is not only dependent on the input p_i at this time, but also on a range of previous inputs $p_{i-1}, p_{i-2}, \dots, p_{i-n}$ corresponding to $t_{i-1}, t_{i-2}, \dots, t_{i-n}$ where n is the delay length that is to be considered by the network. The main benefit obtained when using TDNN's is that there is no need for the network to contain many input nodes to deal with the whole set of delayed input vectors. The sequential data (original signal information) is presented to the network over time and the network is trained to deal with desired steps of delay.

A focused TDNN (delay only at the input layer) was configured with 1 input layer, 3 hidden layers, and 1 output layer using MATLAB's ANN package. Six input layer nodes represent PSG variables, EEG Curve 1, 2 & 3, EOG Curve 1 & 2 and EMG. Each hidden layer included 6 nodes, and the output layer comprised 6 nodes corresponding to sleep stage classes, Awake, Stage 1, 2, 3 and REM. The input signals were first normalized to the range of $[-1, +1]$ and then converted to time sequences. The delay length of the network was set to 1 second (equal to 100 input vectors).

We have implemented a focused TDNN (delay only at the input layer) with an input layer, 3 hidden layers, and an output layer using MATLAB's ANN package. The training procedure is carried out with 500 epochs. The classification accuracy obtained as a baseline for comparison with other approaches is 76.15% of correctly classified records. A total number of records=33,407 were used to train the network.

3 Financial forecasting

A forecasting approach applied to financial market predictions by Bertoli and Stranieri (2004) was adapted in this study to predict sleep stages using data on six PSG variables. Like the TDNN, the approach is based on the intuition that a classification at a point in a series depends on classifications on previous sequences. The approach combines subsequence conditional probabilities to perform a classification in a way that is scalable to large data sets. The adapted forecasting algorithm was applied to data collected from the same patient in an overnight sleep session which included over 3.5 million records on six real-valued PSG signals. The size of the data makes this dataset challenging for any algorithm.

The real valued raw data was first converted to five point interval data labelled BI (big increase), SI (small increase), N (no change), SD (small decrease) and BD (big decrease). Threshold values for the intervals derived from percentiles. The algorithm was applied to discover all unique sequences shorter than 7. The intuition being that a sequence such as BI, BI, BI, SI, N, SD and SI could be discovered on each variable that could discriminate one sleep stage from another.

The confusion matrix (CM) represented in Table 1 depicts classifications made by the Tenon Hospital sleep experts against classifications predicted using the FF approach. This CM illustrates the forecasting approach has some promise given the large and noisy data set, however the prevalence of Stage 2 classifications in the training set led to relatively high mis-classifications. It was also found that the thresholds used in the transformation of real values to interval data for the classification labels (BI, SI, N, SD, BD) had significant impact so further work is required to identify optimal mappings. Unexpectedly, the experiments also found that the length of the pattern used to make the prediction did not need to be particularly long and that a pattern length of 6 or more did not result in any improvement to the predictions but did have a detrimental effect on the processing time.

Actual		Predicted				
		A	S 1	S 2	S 3	REM
A		70,000 45%	35,000 23%	44,000 28%	1,000 1%	5,000 3%
S1		9 4%	195 81%	28 12%	3 1%	5 2%
S2		317,000 2%	1,700,000 9%	13,000,000 60%	1,300,000 7%	2,400,000 13%
S3		0	1 2%	1 2%	45 96%	0
REM		23 4%	22 4%	103 20%	2 0%	374 71%

Table 1. Confusion matrix for financial forecasting approach

4 Non-smooth optimisation

The adaptation of non-smooth optimisation to SSI is based on minimising the deviation between the actual PSG curve and modelled wave patterns. This approach extracts wave characteristics (similar to SP), but these characteristics are more flexible than “standard” SP characteristics and are targeting wave shape descriptions. These characteristics can be used for explicit description of wave shape patterns (similar to ANN), but the dimension of the problem is considerably lower.

The EEG curves are taken to be the sum of two sine curves. The first curve (lower frequency) represents a general trend which is passing through the whole observation sub-period. The second one (higher frequency) is the actual behaviour of the curve along the general trend. The amplitude of each curve is modelled as a piece-wise linear function. This approach allows more precise curve patterns than in the case of classical sine curves where the amplitude is scalar. Additionally, it

allows for abrupt changes in the wave patterns with the piecewise linear function (non-smoothness). In our experiments we use non-smooth optimisation techniques from the GANSO library (Ganso 2006).

All the experiments have been performed on an EEG curve with the horizontal axis corresponding to time. First, the higher frequency sine curve was obtained (Figure 1). This curve is the first approximation of EEG data. The accuracy of approximation is improved by taking into account the general trend of the curve. In Figure 2, the general trend is plotted against the data which represents the difference between the original data and the first trend. Finally, Figure 3 represents the final pattern which follows the original EEG data quite well. In these experiments the subinterval corresponds to 5 seconds of sleep, therefore for each epoch we construct 6 patterns. The dimension of this problem is 12. The dimension of an ANN problem would be 500. This suggests the use of the output of the optimisation problem as an input for ANN could lead to good results.

Figure 1 Actual behaviour and main frequency

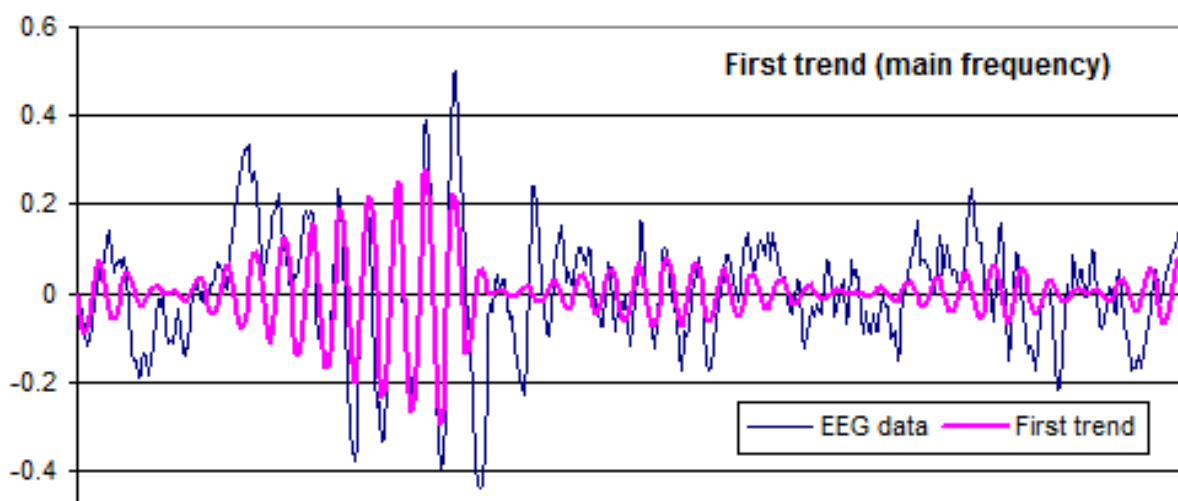


Figure2 General trend

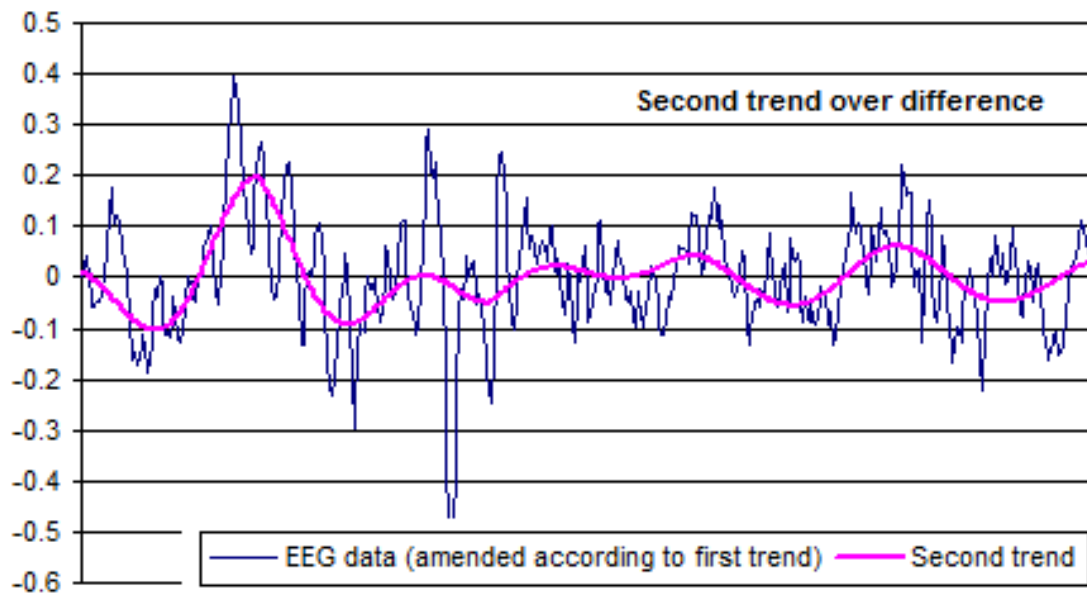
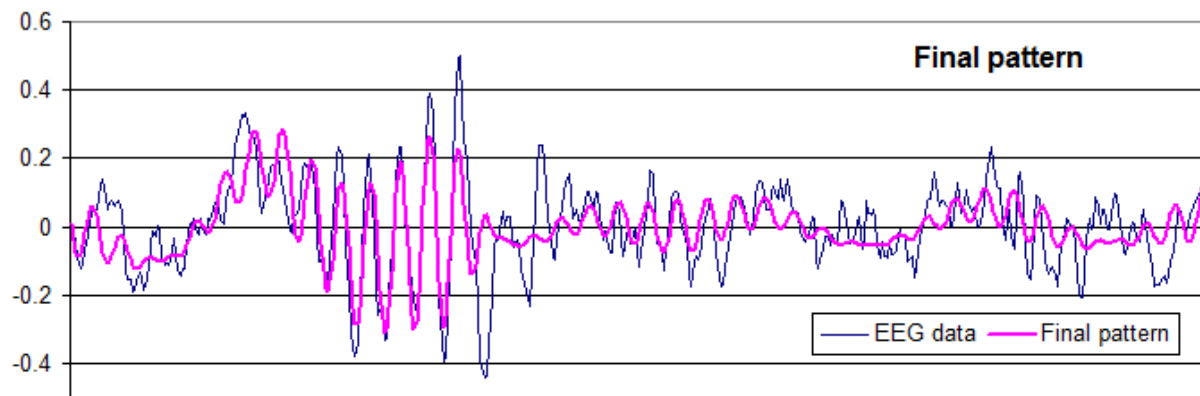


Figure 3 Final pattern (sum of the first and the second trend)



5 Frequency domain analysis

The frequency domain analysis (FDA) approach is based on the basic concept of windowing the signal in the time domain and then taking it into the frequency domain, also called Short-Time Fourier Transform (STFT). The resultant signal is mapped into a two dimensional function of time and frequency. As signals of EEG, EOG, and EMG are not stationary, hence, such techniques give limited precision over this conversion. EEG power spectra has been used in the literature for detecting behavioural microsleeps and estimating the alertness (Jung, Makeig et al. 1997), (Peiris, Jones et al. 2006).

EEG, EMG, and EOG signals are taken into the frequency domain with a window of 30 seconds. The frequency components can be divided into four bands: δ (< 4 Hz), θ (4 – 7 Hz), α (8 – 13 Hz) and β (> 13 Hz) (Carney, Berry et al. 2005). Once these frequency components are separated, the power spectral density is plotted for the window. The power spectral density can be calculated as $\Phi(\omega) = \frac{(F(\omega)F^*(\omega))}{2\pi}$, where $F^*(\omega)$ is the complex conjugate of the frequency matrix.

REMs are generally characterized by a number of features (Pressman 2007), i.e., a low voltage, fast frequency EEG. This is marked by an increase in $\Phi(\beta)$

and relative decrease in the spectral densities of low frequency components. These characteristics will be exploited to detect REM.

According to (Pressman 2007) it is not essential that all the characteristics described before for REM detection are present simultaneously. The presence of only two features out of three can be accepted as a valid REM stage. Figure 4 illustrates the Short Time Fourier Transform (STFT) plot of EMG and Figure 5 represents the STFT for EOG signals. Figure 6 presents the STFT plot of EEG signal and Figure 7 depicts the manual scoring of sleep stages where Stage 5 is REM. One case is described below to explain the REM detection.

Figure 4 EMG in frequency domain against time and sleep stages

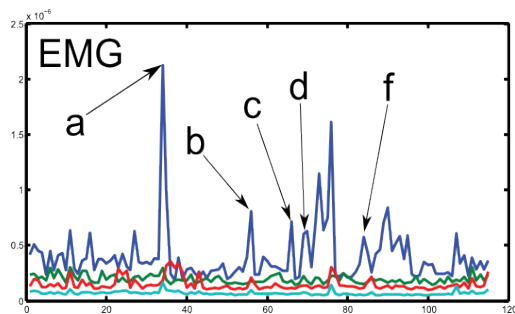


Figure 5 EOG in frequency domain against time and sleep stages

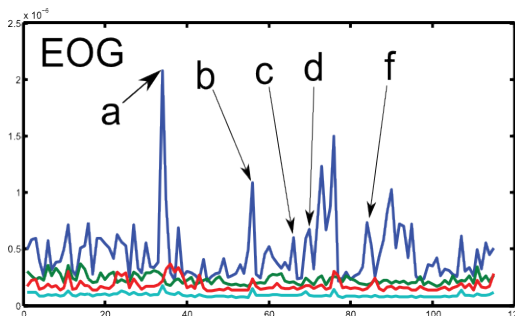


Figure 6 EEG in frequency domain against time and sleep stages

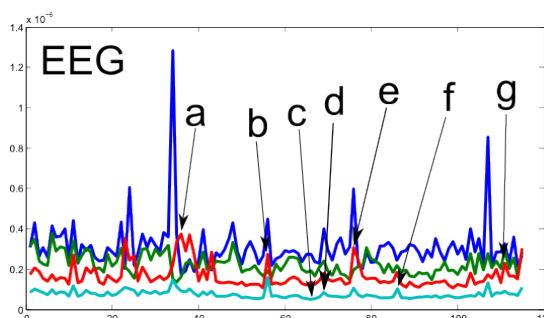
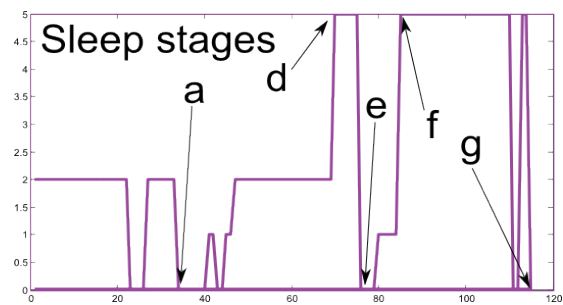


Figure 7 Sleep stages



REM starts with rise in EMG $\Phi(\delta)$, as $\Phi(\delta)$ corresponds to very low frequency components. 'a' and 'b' cannot be considered the start of REM stage because EEG $\Phi(\alpha)$ also increases sharply with EMG $\Phi(\delta)$. 'c' is also not the start of REM because there should be a small rise in EEG $\Phi(\beta)$. 'd' is the start of REM as there is an increase in EMG $\Phi(\delta)$ and a small increase in EEG $\Phi(\beta)$. There is also a rise in all the frequencies of EOG. At 'e', the REM finishes at a sharp increase of EEG $\Phi(\alpha)$, marking an awake stage. 'f' marks the start of REM as there is an increase in EMG $\Phi(\delta)$ accompanied by small increase in EEG $\Phi(\beta)$. 'g' marks the finish of this REM stage as there is an increase in EEG $\Phi(\alpha)$. The last REM for a short time duration has not been detected. All these rules, which are inferred from the characteristics of REM stages described by doctors, can be elegantly implemented using a state machine. However, thresholds of different frequency components that would trigger the state change varies from case to case.

6 Analysis and discussion

In this project it has been shown that the automated SSI procedure is a complex process which cannot readily be achieved without employing a number of diverse methods. This diversity allows one to overcome the problem of "translating" manual scoring rules into automated algorithms.

In our study we used TDNN which can handle higher dimension data better than other types of ANN. The accuracy of 76% is quite good since 2 manual scorers may also produce different classification results (the level of agreement is around 80%). One possible way to enhance the obtained accuracy is to apply TDNN after dimension reduction using NOM. Another possible way is to detect different sleep stages with different approaches, e.g., to identify REM using FDA and Stage 3 using FF. It was also found that the correct detection of Stage 2 is a challenging task for several methods. This is mainly due to the presence of short lasting events (K-complexes), which are difficult to detect by our methods. One future research direction involves the identification of these events using NOM.

7 Conclusions and further research directions

This project is an attempt to build an automated SSI procedure as a meta-classifier, which involves different methods to solve the problem. We identify strengths of particular methods and distribute the “roles”.

In the future we are planning to incorporate these methods in a single procedure. Basing on the research findings of this paper the procedure can be organised as follows:

1. NOM is used as a specific preprocessing tool to convert raw data into a lower dimensional space.
2. Apply ANN methods (or other classification method) to a lower dimensional space of extracted features, obtained on the previous stage.
3. Refine our classification results using FDA and FF for some specific sleep stages (REM and Stage 3 respectively).

7.1 Further research directions

Our future research directions include the meta-classifier building and testing on available data. Also, we are planning to conduct a study on how this meta-learner would learn from two medical experts scored the same data. As it was mentioned before, the level of agreement between two experts can be as low as 80%.

Another promising method for SSI is the Hidden Markov Model (HMM). HMM has a powerful ability to model signals statistically and represent arbitrarily complex probability density functions of the underlying systems. Previous attempts on sleep stage identification (Flexer A., et.al 2002) using HMM did not have much success. One of the important reasons is that these approaches consider modelling sleep stage sequences using a single HMM with a small number of HMM states. The elaboration of this method is another future research direction.

8 References

- Bashashati, A., M. Fatourehchi, et al. (2007). 'A survey of signal processing algorithms in brain-computer interfaces based on electrical brain signals.' *J. Neural Eng.*(4): R32-R57.
- Bertoli and Stranieri (2004). 'Forecasting on complex datasets with association rules'. 8th International Conference on Knowledge-Based Intelligent Information & Engineering Systems. Wellington. Springer. 1170-80.
- Carney, P. R., R. B. Berry, et al. (2005). *Clinical Sleep Disorders*, Lippincott Williams & Wilkins.
- Flexer, A., G. Gruber, et al. (2005). "A reliable probabilistic sleep stager based on a single EEG signal." *Artificial intelligence in Medicine* 33(3): 199-207.
- Ganso (2006) <http://www.ganso.com.au/>
- Iber, C., C. Ancoli-Israel, et al. (2007). *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Technology and Technical Specifications*, Westchester: American Academy of Sleep Medicine.
- Jung, T. P., S. Makeig, et al. (1997). "Estimating alertness from the EEG power spectrum." *IEEE Transactions on Biomedical Engineering* 44(1): 60--69.
- Peiris, M. T. R., R. D. Jones, et al. (2006). Detecting Behavioral Microsleeps from EEG Power Spectra. 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society.
- Pressman, M. (2007). Stages and architecture of normal sleep, UpToDate
- Rajeev and Gotman (2002). 'Digital tools in polysomnography.' *Journal of clinical neurophysiology* 12(2): 136-143.
- Rechtschaffenand, A. and A. Kales (1968). *A Manual of Standardized Terminology, Techniques, and Scoring System for Sleep Stages of Human Subjects*. U. G. P. O. US Public Health Service. Washington, DC.
- Robert, C., C. Guilpin, et al. (1998). 'Review of neural network applications in sleep research.' *Journal of Neuroscience Methods* (79): 187-193.
- Schulz, H. (2008). 'Rethinking Sleep Analysis.' *Journal of Clinical Sleep Medicine* 4(4): 99-103.
- Virkkalaa, J., J. Hasan, et al. (2007). 'Automatic sleep stage classification using two-channel electro-oculography.' *Journal of Neuroscience Methods* 166(1): 109-115.
- Waibel, A., H. Sawai, et al. (1989) 'Modularity and Scaling in Large Phonemic Neural Networks', *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(12): 1888-1898.
- Zobel, J. and Dart, P. (2000): Partitioning number sequences into optimal subsequences. *Journal of Research and Practice in Information Technology* 32(2):121-129.

Visualising a State-wide Patient Data Collection: A Case Study to Expand the Audience for Healthcare Data

Wei Luo¹ Marcus Gallagher¹ Di O’Kane² Jason Connor³ Mark Dooris⁴
Col Roberts² Lachlan Mortimer² Janet Wiles¹

¹ School of Information Technology & Electrical Engineering
The University of Queensland
Brisbane, Australia
Email: {luo, marcusg, wiles}@itee.uq.edu.au

² Clinical Practice Improvement Centre
Centre for Healthcare Improvement
Queensland Health
Brisbane, Australia

³ Department of Psychiatry
The University of Queensland
Brisbane, Australia

⁴ Department of Cardiology
Royal Brisbane and Women’s Hospital
Brisbane, Australia

Abstract

This paper describes the application of existing and novel adaptations of visualisation techniques to routinely collected health data. The aim of this case study is to examine the capacity for visualisation approaches to quickly and effectively inform clinical, policy, and fiscal decision making to improve healthcare provision. We demonstrate the use of interactive graphics, fluctuation plots, mosaic plots, time plots, heatmaps, and disease maps to visualise patient admission, transfer, in-hospital mortality, morbidity coding, execution of diagnosis and treatment guidelines, and the temporal and spatial variations of diseases. The relative effectiveness of these techniques and associated challenges are discussed.

Keywords: Visualisation, Exploratory Data Analysis, Routine Data Collection

1 Introduction

The state of Queensland has the third largest population in Australia [20]. In the financial year 2006-2007, public hospitals in Queensland treated more than 780,000 inpatients [11, page 11-12]. All such inpatient encounters are routinely collected in the Queensland Hospital Admitted Patient Data Collection (QHAPDC) [4]. This centralized database setup represents an invaluable resource for knowledge discovery and evidence based medicine. Since 2005, the health department of the state government, Queensland Health, has implemented a series of initiatives to improve performance monitoring and governance. As an example, the VLAD (Variable Life Adjusted Display) system is in place to detect extraordinary trends and occurrences [1], using data from QHAPDC. Significant challenges exist in developing efficient and ef-

fective ways of maximizing the utility of this data resource. A visualisation toolkit tailored for health data such as QHAPDC is likely to have significant benefits to both Queensland Health and the broader medical community. This article describes a step towards developing such a toolkit.

In the following sections, we describe various techniques used to visualise QHAPDC data. Our visualisation is exploratory in nature, with the overall aim of expanding the audience for healthcare data, and the following specific aims guiding the selection of techniques:

1. To assess data quality, and hence to identify potential improvements to the data collection process. (See Section 3.1 for an example where coding issues were identified through a simple histogram.)
2. To detect anomalies (both positive and negative) in clinical practices, and hence to promote clinical practice improvement. (See Section 3.2 for such an attempt.)
3. To identify temporal trends and spatial variation in the data for better allocation of health care resources. (See Section 3.5 and 3.6.)
4. To identify the potential research value of the routinely collected data; to generate medical hypotheses that lead to further research projects. (See Section 3.4.)

Visualisation of public health data has been demonstrated to enhance knowledge and support decision making (see for example [12, 22, 23, 24]). A unique challenge of the state-wide QHAPDC database is that it is essentially a repository for a number of largely independently generated data collection sites, typically different hospital campuses. While this provides comprehensive and rich data for visualisation techniques, there are key challenges of “noisy” and missing data associated with possible non-uniformity of data coding practices across hospital sites.

This article is organised as follows: Section 2 describes the dataset to be used and briefly explains

terms such as International Classification of Diseases, 10th Revision (ICD-10) and Diagnosis-Related Group (DRG). Section 3 reports on a collection of visualisation plots within specific areas of health data. Section 4 summarises the findings, highlights the unique contributions of these techniques and discusses opportunities for further research in this field.

2 Data Description

QHAPDC provides comprehensive data on patient demographic and clinical/treatment occurrences. In this preliminary study, we focus on the clinical data as coded by the International Classification of Diseases.

2.1 International Classification of Diseases

The morbidity information in QHAPDC is encoded following the *International Classification of Diseases and Related Health Problems, 10th Revision, Australian Modification* (ICD-10-AM, [6]). ICD is published by the World Health Organization (WHO) as a standard way of coding morbidity and mortality statistics [21]. The 10th Revision (ICD-10) is the current standard. ICD-10-AM includes the Australian extensions to ICD-10 and contains more than 20,000 codes in total.

ICD-10 codes are organized in 22 chapters, each denoted by a capital letter. A chapter is further divided into blocks, each denoted by a number from 0 to 99. A block again can be further refined by appending a fraction number¹. Take the code I21.4 as an example. The leading letter I indicates it belongs to the chapter of *diseases of the circulatory systems*; I21 is the block for *Acute Myocardial Infarction* (AMI); The fraction .4 indicates the infarction is *subendocardial* [21].

Based on diagnoses and procedures, hospital cases are classified into over 600 *Diagnosis-Related groups* (DRGs)². DRGs can be further grouped into 25 *Major Diagnostic Categories* (MDCs) [17]. For example, MDC 05 *Diseases and disorders of the circulatory system* contains DRGs ranging from F01A—*Implantation or Replacement of AICD, Total System with Catastrophic complication* to F76B—*Arrhythmia, Cardiac Arrest and Conduction Disorders without Catastrophic or Severe complication*. In later discussions, we focus on cases in MDC 05 that have a principal ICD code in the I21 (AMI) block.

2.2 AMI data

Ischaemic heart disease is the number one cause of death for Australians [14, page 44]. Acute myocardial infarction (AMI), also known as a *heart attack*, is a major clinical form of ischaemic heart disease that affects many Australians. The Australian Institute of Health and Welfare, citing the 2004-2005 National Health Survey, suggests that about 1.8% of Australians reported a history of AMI [16, page 183].

In this article, all visualisations are demonstrated using an AMI dataset from QHAPDC³. Specifically, the dataset was created by extracting QHAPDC records that met the following criteria:

1. The patient was treated in one of 6 *principal referral and specialised* public hospitals.

¹In ICD-10-AM, morphology codes (for identifying the morphology of neoplasms) follow a different convention: a morphology code consists of 4 digits following the letter M.

²DRGs are used in the casemix funding model in Australia.

³The visualisation techniques we present are to gain understanding of the data. Further confirmatory studies are necessary before definitive conclusions can be made.

2. The principal diagnosis was “acute myocardial infarction” (i.e., the ICD is in the I21 group)⁴.
3. Records were extracted from January 2005 to December 2008.

3 Visualisation

3.1 Histograms for data cleaning

As a preliminary step prior to involved statistical analysis, it is desirable to “feel” the structure of data and assess the data quality to avoid “garbage in, garbage out”.

We start out by showing how simple graphics such as a histogram can expose anomalies in data, and how interactions such as colour brushing [26] can help discover unexpected relations.

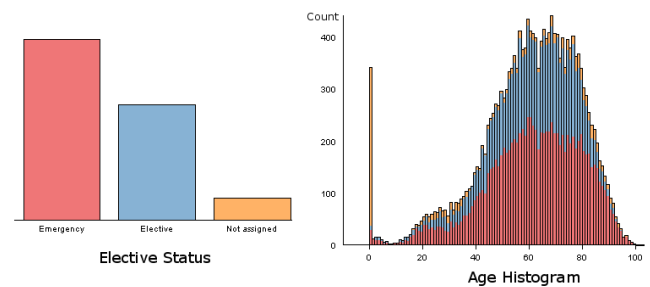


Figure 1: Histogram of the age of the AMI patients described in Section 2.2.

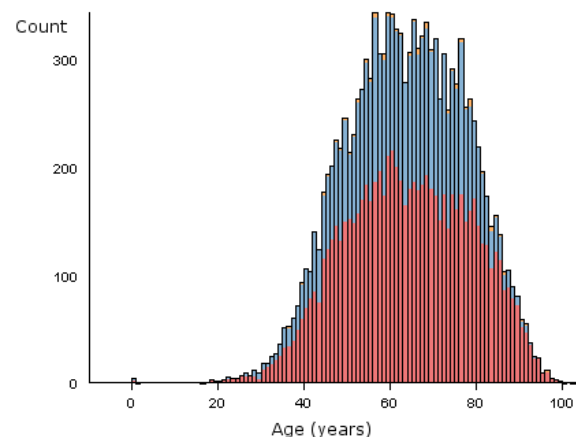


Figure 2: Histogram of the age of patients in the circulatory disease MDC. The shape is less skewed than in Figure 1. The colour scheme for elective status is the same as in Figure 1.

As the AMI dataset has about 20,000 cases, we expect certain continuous variables to approximately follow a unimodal distribution⁵; otherwise we may suspect the data come from multiple distinct sources.

As an example, Figure 1 displays the histogram of the age of all patients. One distinctive feature of the histogram is the spike at age 0. Colour-brushing various discrete variables in the data revealed a connection between elective status and patient’s age.

⁴Late in Section 3.1 we shall see that extraction based on the principal diagnosis may not be a good choice, as some coding practice is either unreliable or unintuitive.

⁵It is useful here to distinguish measurement variables from nominal variables. For example, a patient’s age is a measurement variable. But his or her length of hospital stay is not a measurement variable, as it is an aggregated function of whether the patient is discharged on a particular day, which is nominal.

Namely, most newborns in the dataset did not have elective status assigned. Moreover, the histogram also shows a group of patients of age 16 to 37 who did not have elective status. Most of these patients were female and in obstetric DRGs. On subsequent investigation we found in the QHAPDC manual ([4, Section 7.30]) that admissions for normal delivery and admissions which begin with the birth of the patient do not have elective status assigned. Why did these patients in obstetric DRGs have AMI as their principal diagnosis? Consultation with an obstetrician and coding staff may help answer the question.

Figure 1 has highlighted the heterogeneous nature of the dataset. Upon further examination, we restricted the data to cases in the MDC 05—Diseases and disorders of the circulatory system [15], which account for 70% of the original cohort. The age distribution of the selected subgroup is shown in Figure 2. The new histogram has a shape more closely resembling a normal distribution, indicating a more homogeneous group of patients.

3.2 Fluctuation plots: visualising patient transfer among hospitals

With a state-wide database, we can visualise patient flow among hospitals within the state, which can potentially help to facilitate the logistics involved in patient transfer. A fluctuation plot serves this purpose well. A *fluctuation plot* is used to visualise the contingency table of two discrete variables, where the count of each combination is represented by a square of proportional area.

Figure 3 shows AMI patient transfers since 2005. The size of a square indicates the relative number of transfers from one hospital to another. One can see that most patients were transferred to hospital **A**, as the vertical column **A** has most large squares. One can also infer that hospital **B** and hospital **D** are referral destinations for hospital **C** and hospital **F**, respectively, as indicated by the two large squares at the coordinates (B,C) and (D,F).

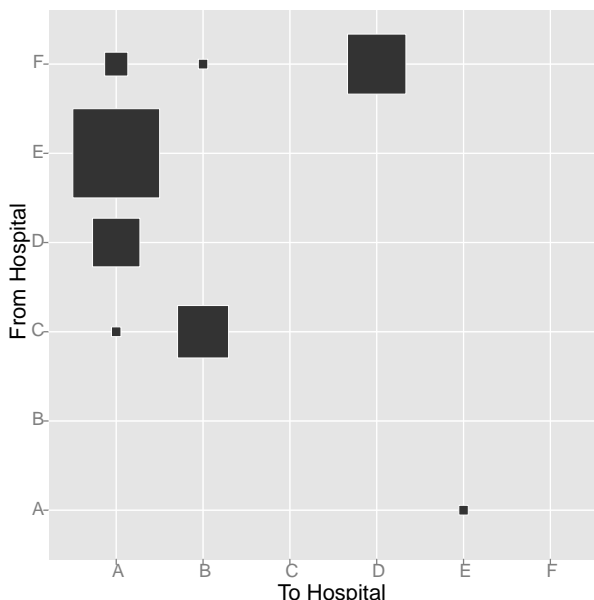


Figure 3: Fluctuation plot showing number of patients transferred among 6 Queensland hospitals. The size of a square indicates the number of transfers. It shows that hospital A is a major transfer destination for AMI patients.

For patients being transferred, it would be ex-

pected that they would receive the same principal diagnosis in the second hospital. We used a fluctuation plot to verify this assumption. In Figure 4, we see that principal diagnoses were largely identical, as the squares along the diagonal contain much of the total area of all the squares. However, small squares off the diagonal indicate potential inconsistencies in diagnosis or coding, or a logical clinical phenomenon not reflected in the current data set. As the principal diagnosis affects how a patient is treated, these data anomalies warrant further investigation.

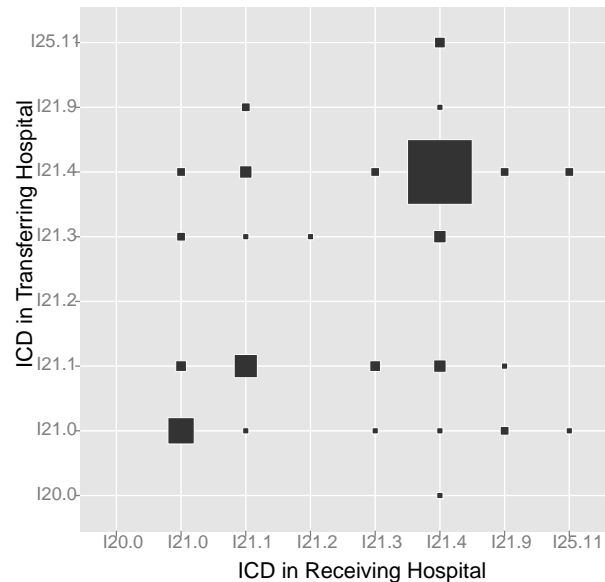


Figure 4: Fluctuation plot showing principal diagnoses before and after transfer. Squares along the diagonal correspond to cases with an unchanged principal diagnosis after transfer. Squares off the diagonal correspond to cases with different principal diagnoses before and after transfer. Size of a square reflects the relative number of particular combination.

3.3 Mosaic plots: visualising clinical pathways

The *Clinical Practice Improvement Centre* (CPIC) at Queensland Health has developed a set of state-wide Cardiac pathways⁶.)

As a simplified example, suppose a patient presented at an emergency department with chest pain. Based on her ECG result, a medical officer decided whether she had a ST-segment elevation myocardial infarction (STEMI) case or a Non-ST-Segment Elevation case. If it was a STEMI case, then percutaneous coronary intervention (PCI, insertion of a catheter into coronary vessels to remove blockages or improve blood flow) was needed; otherwise PCI might not be necessary. With a mosaic plot, we can gain insight into how these clinical pathways have been followed in hospitals⁷.

The mosaic plot is closely related to the fluctuation plot. In a mosaic plot of two discrete variables, each

⁶A clinical pathway is “a document outlining a standardised, evidence-based multidisciplinary management plan, which identifies the appropriate sequence of clinical interventions, timeframes, milestones and expected outcomes for a homogenous patient group” [10]

⁷A more data-driven approach to visualise the potential causal relation between comorbidities and procedures is to learn a Bayesian network using the database. One such learning algorithm can be found at [25]).

major cell shows the relative frequency of data observations corresponding to values of the two variables. With an additional discrete variable, each cell is further divided according to the conditional frequency of the variable in the cell. Hence at both the local (cell) and global level, the visualisation shows the degree of non-uniformity across variable values.

Here we use mosaic plots to visualize the conditional relationship between PCI and ST-segment elevation. Unfortunately ST-segment elevation is not explicitly encoded in ICD-10-AM or anywhere else in QHAPDC⁸. After consulting coding staff from a Queensland hospital, we used the following rule to estimate ST-segment-elevation status: For patients with diagnoses I20.0-I20.3, we assumed ST-segment elevation (STEMI in Figure 5); for patients with diagnosis I20.4, we assumed no ST-segment elevation (NSTEMI in Figure 5); for patients with diagnosis I20.9, we assume the ST-segment elevation status was unknown (UNKNOWN in Figure 5). In addition, we assume that PCI was performed if and only if the case is in one of the DRGs: F10Z, F15Z, or F16Z.

Figure 5 shows that practices are relatively consistent across 3 hospitals of similar sizes. It also shows that not every patient in the STEMI group received PCI, whereas some patients in Non-STEMI group did receive PCI.

A possible explanation for the above inconsistency between pathways and data is that we have not correctly estimated STEMI status for patients. If that is the case, then Figure 5 shows the importance of coding the STEMI status in ICD-10-AM.

Mosaic plots have been used in Queensland Health for displaying the relationships between risk and the first and second Troponins, for the chest-pain management protocol. People found it “the most appropriate and easiest” way to present the relationships and information, as one could clearly see the “size” of the issues from presenting the information in this manner.

In general, the health-policy workers often need to present two or three way tables. It is suggested that the mosaic plot and its variants have a high applicability for these tables. The main limitation of the mosaic plot is that basic software such as Excel does not have the capability and not many people are aware of this style of presentation method.

3.4 Heatmaps: visualising the connection between severity and morbidity

As mentioned in Section 2, the QHAPDC database consists of two parts: socio-demographic data and clinical data, which includes morbidity coded data. How to use morbidity data is an important but challenging problem. For example, all past INFORMS data mining contests centered on mining the morbidity data to predict health care outcomes [8, 9]. Here we introduce a use of a heatmap to visualise morbidity to effectively identify risk factors and protective factors for both mortality and hospital long-stay.

Heatmaps have been primarily used in genomics to visualise DNA microarrays [5]. As a DNA microarray can be regarded as a 2-dimensional matrix, its rows and columns can be permuted to highlight clusters. Similarly, morbidity data coded in ICDs can be thought as a matrix M :

1. Each patient is a row;
2. Each ICD code is a column;

⁸In ICD-10-CA (Canadian enhancement of ICD-10), STEMI has the code R94.30 (see [7, page 26]).

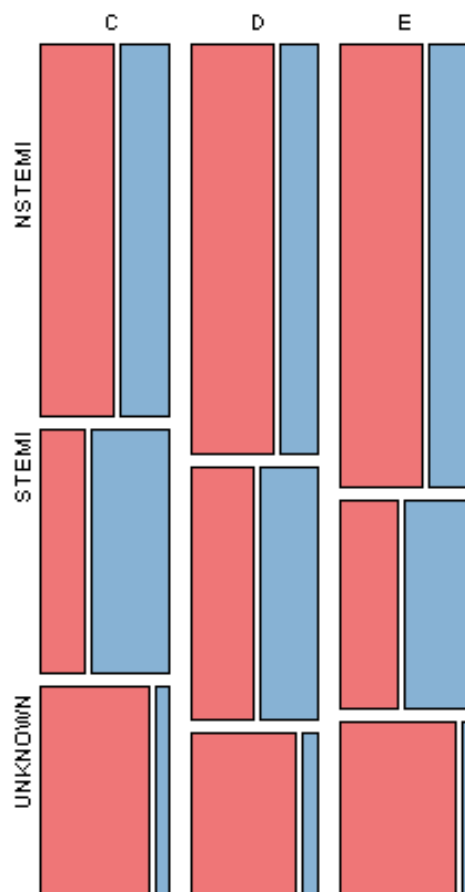


Figure 5: Mosaic plot showing the relation between ST-segment elevated MI (STEMI) and percutaneous coronary intervention (PCI). Three hospitals of similar sizes (C, D, and E) are compared. Blue shading in the plot corresponds to patients who received PCI. The clinical pathways suggest that STEMI patients should receive PCI, whereas non-STEMI patients may not need PCI. The plot shows that STEMI patients are more likely to receive PCI, but that relation cannot be taken as absolute in practice.

3. A matrix cell $M(i, j)$ takes value 1 if patient i was diagnosed with morbidity j , and it takes value 0 otherwise.

Therefore, we consider visualising morbidity data with a heatmap as a natural choice.

In a heatmap, ICDs can be clustered just as in a DNA microarray. Patients could also be clustered with respect to their morbidity information. To identify risk and protective factors, however, we sorted the patients (rows) with the following measure of severity⁹:

$$\text{SEVERITY}(x) = \begin{cases} C - \text{LOS}(x) & \text{if } x \text{ died in hospital} \\ \text{LOS}(x) & \text{otherwise} \end{cases} \quad (1)$$

where $\text{LOS}(x)$ is the length of hospital stay (in hours) of patient x , $C = 2 \max_x \text{LOS}(x)$.

Figure 6 shows such a heatmap. Records are ordered from top to bottom with increasing severity. It shows that a cardiogenic shock has strong correlation with AMI mortality.

In Figure 6, we have included only diagnosis codes; a heatmap could also be generated with both diagnosis and procedure codes to identify correlations (1) between diagnoses and procedures, and (2) between procedures and reduced/prolonged hospital stay. Also the severity measure could be adjusted by age and gender.

3.5 Time plot: visualising temporal trends and seasonal patterns

A *time plot*, where observations are plotted over time [2], displays temporal variation of the data. It helps us

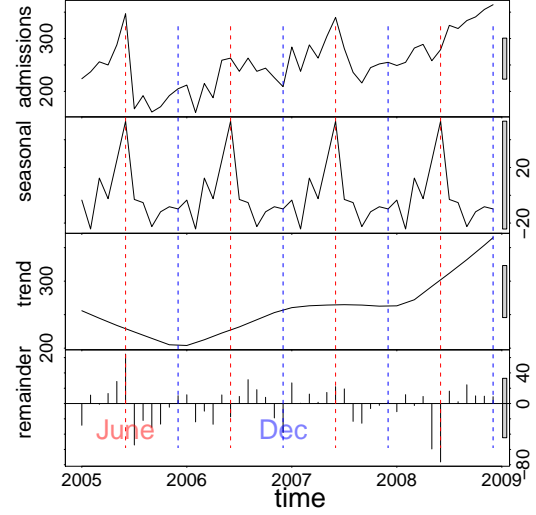
1. recognize trends and seasonal patterns in the data, which is potentially useful for resource allocation and planning, and
2. check the consistency of health care quality and spot outliers that warrant further investigation.

Regarding the second point, a time plot can complement the VLADS system (see Section 1) to provide visualisations that are easier to understand by people with no knowledge of statistical process control.

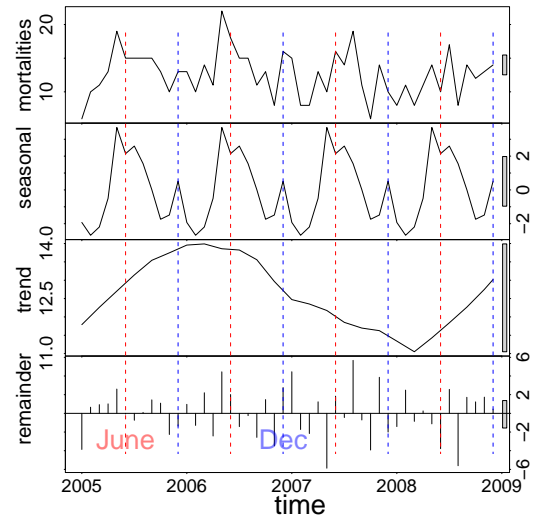
Here we give a simple example of comparing the admission and mortality of AMI patients. Figure 7 plots the number of admissions and deaths each month from January 2005 to December 2008. We use the method described in [3] to decompose the two time series into seasonal, trend and irregular components. Figure 7(a) shows more patients were admitted during the month of June each year, a winter month in the southern hemisphere. It also shows a trend of increasing numbers of admissions since 2006. The seasonal plot in Figure 7(b) shows that most AMI deaths occurred during the winter months in Australia. Comparing the trend plots in the two figures, we see that in 2005, the number of deaths increased while the number of admissions decreased. But in the years 2006 and 2007, according to the available data, the mortality rate decreased monotonically.

3.6 Disease map: visualising spatial variation

As a complementary technique to time plots, which are useful for visualising temporal variation of the data, disease maps are useful for visualising spatial variation. A *Disease map* is a popular type of visualisation for public health data (see for example [12, 22, 23]).



(a) number of admissions



(b) in-hospital mortality

Figure 7: Time plots showing seasonal, trend and irregular decomposition of admission and mortality using loess smoothing [3].

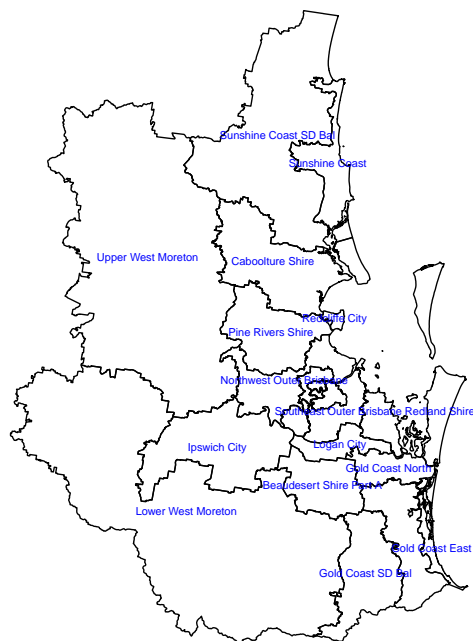


Figure 8: Statistical subdivisions of South East Queensland in 2006. Map data is courtesy of the Australian Bureau of Statistics.

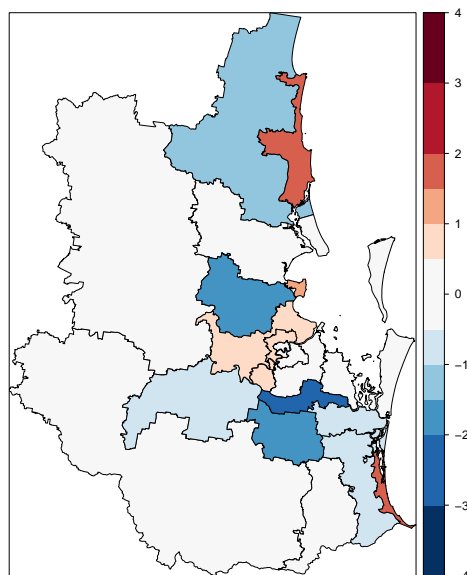


Figure 9: Residual map for the population aged 85 and over in South-East Queensland. A quasi-Poisson model was fitted to the estimated number of residents aged 85 and over in each statistical subdivision. The standardized residuals are plotted to highlight extreme values. The map shows that dense coastal population centres have a higher percentage of seniors.

For a geographic division, let X be the number of observed cases, and E be the expected number of cases (for example by adjusting for its population and age distribution). Then a disease map will highlight the discrepancy between X and E , for example, by plotting the standardized mortality/morbidity ratio (SMR) X/E for each geographic division. Both AMI incidence and mortality rate in Queensland can be mapped in this way to understand their geographic variation, which itself reflects the distribution of indigenous population or older demographic in particular areas.

Due to the sensitive nature of mortality data, however, maps of AMI incidence and mortality rate are not shown here. To still demonstrate the technique, we use the population data from the 2006 Australian Census [19]. Figure 8 shows a map of statistical subdivisions in South East Queensland. The region accommodates more than 66% of Queensland population [18]. Figure 9 shows a residual map for the percentage of the population aged 85 or over. From the map, we see that in the year 2006, both the Sunshine Coast and the Gold Coast East have higher proportions of older population.

The health-policy workers consider the thematic map an “excellent” way to visualize the allocation of resources, and to assess areas of need, provided the appropriate data and technique is used in displaying the information. The availability of mapping layers, however, limits the applicability of this technique.

4 Discussion

This paper has demonstrated that visualisation can provide important clinical insights into various aspects of the routinely collected health data. Our results show that visualisation is helpful in every stage of the data life-cycle, from collection and validation (see Section 3.1), to reporting (see Section 3.2), to knowledge discovery (see Section 3.4), and to anomaly detection (see Section 3.5 and 3.6). Visualisation complements more formal statistical analysis with its flexibility, and generates final products easily understood by data managers and clinicians, who have not necessarily received formal statistical training.

The overarching goal of our project is to expand the audience for healthcare data. This paper makes the following contributions to health-care industry in general and to health informatics & knowledge management in particular:

1. It assesses the applicability and value of various visualisation techniques in health and medical data.
2. It proposes a new way to visualise diagnosis coding with respect to case severity.
3. It constitutes an important step toward a visualisation tool-kit for healthcare workers.

To further develop the toolkit, we are extending the visualisation techniques and applying these techniques to a wider range of data, including a Cesarean section dataset and an Orthopaedic hip-replacement dataset from Queensland Health. At the same time, we have been collaborating with clinicians from Queensland Health to access the utility of these visualisation techniques and to develop types of visualisations that are informative in clinical settings. We also continue to work with CPIC to understand the nature and quality of the Queensland Health data, and to further develop the insights that can be gained with these unique resources.

⁹This severity measure is rather coarse. Currently we are undertaking research for a more accurate estimate of severity [13].

Acknowledgment

This work is supported by an ARC Linkage Grant to MG, DO, JC, and JW (LO 0776417) and the Clinical Practice Improvement Centre, Queensland Health. We thank Michael Milford and Chris Nolan for help on graphics manipulation, and Karen Borchardt for help on locating the ABS census data.

References

- [1] Clinical Practice Improvement Centre. *VLADs for dummies*. Wiley, Queensland Health edition, 2008.
- [2] Chris Chatfield. *The Analysis of Time Series: An Introduction, Sixth Edition*. Chapman & Hall/CRC, 2003.
- [3] Robert B. Cleveland, William S. Cleveland, Jean E. Mcrae, and Irma Terpenning. STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6(1):3–73, 1990.
- [4] Sue Cornes. *Queensland Hospital Admitted Patient Data Collection (QHAPDC) Manual*. Health Information Centre, Queensland Health, 2005–06 edition, 2005.
- [5] Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, and David Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95(25):14863–14868, 1998.
- [6] The National Centre for Classification in Health (Australia). ICD-10-AM. http://nis-web.fhs.usyd.edu.au/ncch_new/2.aspx.
- [7] Canadian Institute for Health Information. Canadian coding standards for ICD-10-CA and CCI. http://secure.cihi.ca/cihiweb/dispPage.jsp?cw_page=RC_382_E.
- [8] The Institute for Operations Research and the Management Sciences. INFORMS data mining contest 2008. <http://informdataminingcontest.googlepages.com/>, 2008.
- [9] The Institute for Operations Research and the Management Sciences. 2009 INFORMS data mining contest. <http://www.informsdmcontest2009.org/>, 2009.
- [10] Queensland Health. Queensland health implementation standard: Clinical pathways. http://www.health.qld.gov.au/cpic/pdf/clin_path_imp_strd.pdf.
- [11] Queensland Health. Annual public hospitals performance report 2006-07. http://www.health.qld.gov.au/performance/performance_report.asp, November 2008.
- [12] Nada Lavrac, Marko Bohanec, Aleksander Pur, Bojan Cestnik, Marko Debeljak, and Andrej Kobler. Data mining and visualization for decision support and modeling of public health-care resources. *Journal of Biomedical Informatics*, 40(4):438 – 447, 2007.
- [13] Wei Luo, Marcus Gallagher, and Janet Wiles. Bayesian modelling of inpatient severity. In preparation.
- [14] P. Magnus and K. Sadkowsky. *Mortality over the twentieth century in Australia: trends and patterns in major causes of death*. Australian Institute of Health and Welfare, 2006. <http://www.aihw.gov.au/publications/phe/motca/motca.pdf>.
- [15] Australian Institute of Health and Welfare. Separation, patient day and average length of stay statistics by Australian Refined Diagnosis Related Group (AR-DRG) version 5.0/5.1, Australia, 1998-99 to 2006-07. www.aihw.gov.au.
- [16] Australian Institute of Health and Welfare. *Australia's Health 2008*. Australian Institute of Health and Welfare, 2008.
- [17] Department of Health and Australia Ageing. MDC-Partition-DRG structure. [http://www.health.gov.au/internet/main/publishing.nsf/Content/2A68FBBD47DC69DOCA25753E00032FC2/\\$File/MDC-Partition-DRG%20Structure.pdf](http://www.health.gov.au/internet/main/publishing.nsf/Content/2A68FBBD47DC69DOCA25753E00032FC2/$File/MDC-Partition-DRG%20Structure.pdf).
- [18] Department of Infasture and Queensland Government Planning. South East Queensland. <http://www.dip.qld.gov.au/seq>.
- [19] Australian Bureau of Statistics. 3235.0 - population by age and sex, Australia, 2006. <http://www.abs.gov.au/ausstats/abs@.nsf/Products/3235.0~2006~Main+Features~Queensland?OpenDocument>, August 2008.
- [20] Australian Bureau of Statistics. Australian demographic statistics, dec 2008. <http://www.abs.gov.au/ausstats/abs@.nsf/mf/3101.0>, April 2009. Catalogue number 3101.0.
- [21] World Health Organization. International statistical classification of diseases and related health problems, 10th revision, version for 2007. <http://apps.who.int/classifications/apps/icd/icd10online>.
- [22] Bambang Parmanto, Maria Paramita, Wayan Sugiantara, Gede Pramana, Matthew Scotch, and Donald Burke. Spatial and multidimensional visualization of Indonesia's village health statistics. *International Journal of Health Geographics*, 7(1):30+, June 2008.
- [23] Mohsen Rezaeian. How to visualize public health data? Part one: Box plot and map. *Middle East Journal of Family Medicine*, 6(10):19–24, December 2008.
- [24] Mohsen Rezaeian. How to visualize public health data? Part two: Direct and indirect standardization methods. *Middle East Journal of Family Medicine*, 7(1):42–44, January 2009.
- [25] Oliver Schulte, Gustavo Frigo, Russell Greiner, Wei Luo, and Hassan Khosravi. A new hybrid method for Bayesian network learning with dependency constraints. In *CIDM*, pages 53–60, 2009.
- [26] Martin Theus and Simon Urbanek. *Interactive Graphics for Data Analysis: Principles and Examples (Computer Science and Data Analysis)*. Chapman & Hall/CRC, 2008.

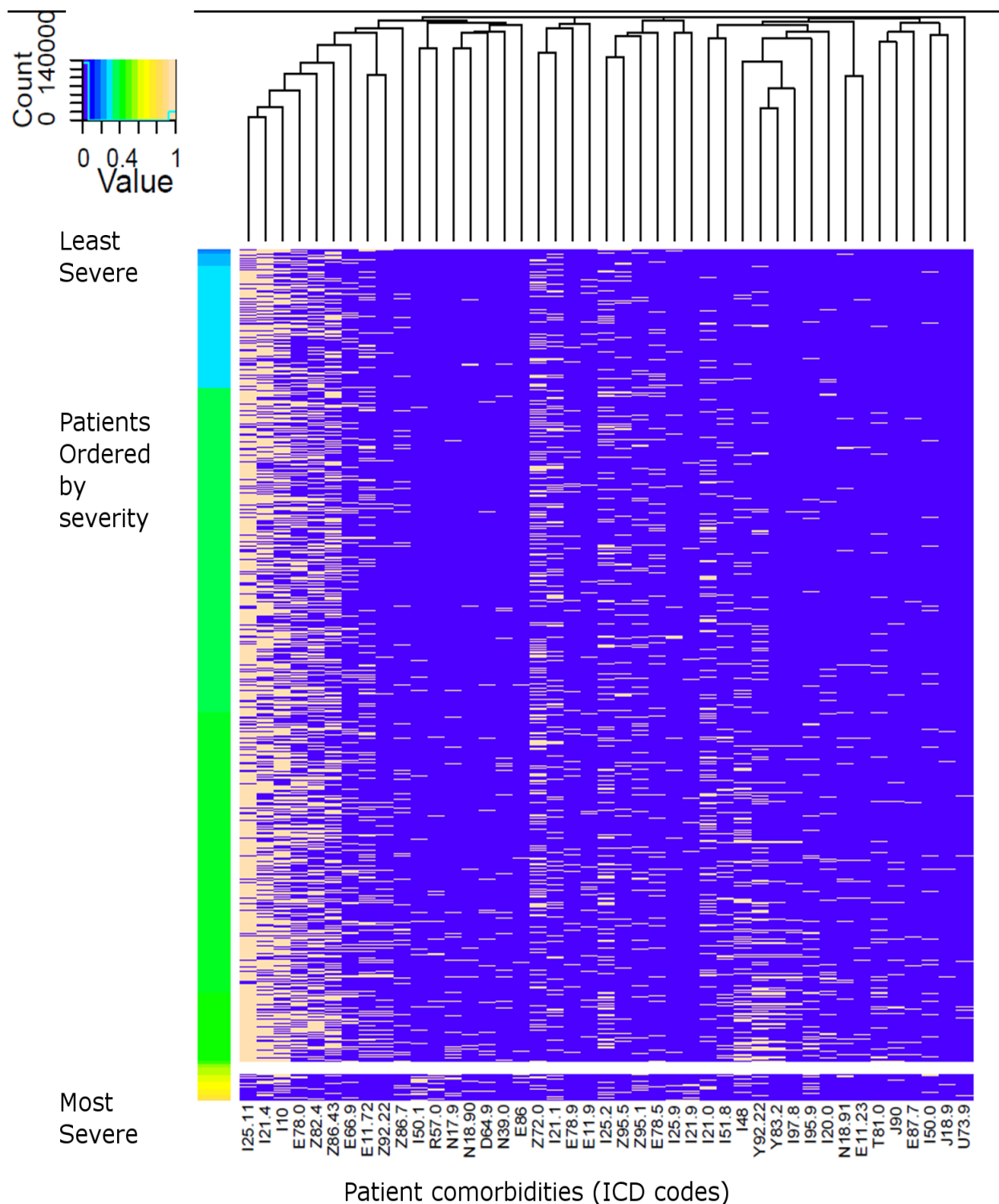


Figure 6: Heatmap showing morbidity codes from a single hospital. Each row represents a patient's morbidity codes; Rows are ordered from top to bottom with increasing severity. A horizontal gap near the bottom separates patients who survived (above the white line) from those who died in hospital (below the line). The vertical side bar on the left shows each patient's severity estimate (normalized to a number between 0 and 1, and then colour-coded with the colour key at the top left corner). Each column represents the distribution of an ICD code across patients. (Columns have been automatically clustered based on similarity of their distribution. The dendrogram at the top shows the clustering.) The plot shows that most patients had the code I21.4 (acute subendocardial myocardial infarction), which is interpreted as non-ST-segmented Elevated MI in the hospital under study. It also shows that conditions R57.0 (cardiogenic shock) and I50.1 (left ventricular failure) are strongly correlated with mortality, while conditions T81.0 (haemorrhage and haematoma complicating a procedure, not elsewhere classified) and Z92.22 (personal history of long-term (current) use insulin) often occur among patients with long hospital stays, but not so often among patients who died.

High Accuracy Information Retrieval and Information Extraction System for Electronic Clinical Notes

Jon Patrick and Min Li

Health Information Technology Research Laboratory, School of IT
The University of Sydney
Sydney, NSW 2006, Australia

jonpat@it.usyd.edu.au

mili9528@uni.sydney.edu.au

Abstract

There is a great demand for highly accurate and timely Information Retrieval and Information Extraction in medicine and health care. To meet this need, we have developed a novel system, Intelligent Clinical Notes System (ICNS) to assist doctors in retrieving clinical notes based on concept searching. This has required dealing with the both the software engineering and natural language processing aspects of the task. This system has been installed and integrated into the existing clinical information system in the Intensive Care Unit, Royal Prince Alfred Hospital, Sydney.

Keywords: Information Extraction, SNOMED CT (SCT), Text to Snomed, Intelligent Clinical Notes System (ICNS).

1 Introduction

Clinical notes is a new domain for information extraction, where health professionals have a demand for high accuracy access to information, while the demand for intelligent techniques in their information seeking tasks has become much stronger. However, a good information extraction system in health care should not only retrieve information accurately but obtain it quickly in time critical situations.

Although the advance of clinical information systems is self-evident, serious limitations in semantic retrieval in these systems are still present. The principal reason is that clinical retrieval, in which correct semantics are crucial, are quite different from the traditional keyword based search, as performed by Google. In particular, a clinical concept may be known by a large variety of different names, consequently the traditional keyword can't retrieve all instances of the same concepts.

In this paper, we present an Intelligent Clinical Notes System (ICNS), built to Intensivists requirements that can retrieve patient notes and extract useful information from them. The main objective for this system is to use natural language processing (NLP) to serve clinicians and improve their productivity and efficiency thus contributing to patient quality and safety. For example, the concept based search engine can automatically identify synonyms for use in a search request. In addition, automatic spelling correction, as well as, most of the abbreviations and

acronyms are identified and expanded, which makes the clinical notes more readable for the non-author.

During the development of this system, several external resources are used. They are a medical ontology SNOMED CT, a process for converting text into SNOMED CT terminology, gazetteers and dictionaries.

1.1 SNOMED CT

SNOMED CT (SCT) is a comprehensive medical ontology constituting a reference terminology in a relationship hierarchy with approximately 350,000 concepts and 1.4 million relationships (Wua et al., 2004). The computable concept definitions and relationships, which it provides can help determine the semantic categories in collected data. With the help of its reference terminology, clinical notes can be codified automatically, and data extraction and analysis relating to the causes of disease, the treatment of patients, and the outcomes of the overall health care process can be much more easily researched (Spackman, et al., 1997).

1.2 Text to SNOMED CT Converter (TTSCT)

In order to index all the medical terminology in the patient notes, an existing algorithm [Text To SNOMED CT (TTSCT) (Patrick, et al., 2007)] for mapping text to SCT is used in this system. The implementation of TTSCT is based on the strategy that the input is collected into "chunks" based on their meaning by utilizing natural language processing then the text tokens are matched to SCT concept tokens. Finally, a matching algorithm uses a pre-computed matching matrix to rank SCT concept descriptions against the phraseology of the chunks to identify the highest ranking match for the longest phrase. Subsequently negations and qualifications are separately recognised.

TTSCT is utilised in the ICNS by scanning through the whole patient note, identifying all the medical concepts within the free text and mapping them into concept IDs of SCT. This function then enables searches to be made on a broad range of phrases equivalent to SCT concepts rather than literal strings of descriptions.

1.3 Gazetteers

A gazetteer is a word list of one class of content, such as a list of staff names, new medical terminology, etc which is used to improve the quality of information extraction. In ICNS, we use words extracted from clinical notes consisting of terminology not in SCT to build gazetteers (such as the abbreviations gazetteer, acronym gazetteer, clinical staff gazetteer, etc), and then recognize them in the patient notes.

Gazetteers serve a different purpose to SCT terminology. The accurate encoding of all content in the patient notes is not possible due to limitations of natural language processing functions and the failure of SCT to provide 100% coverage of clinically relevant content. The gazetteers have been trained, from a 60 million token ICU corpus, with the content that is unknown to SCT or expressed in forms that are not readily identifiable in SCT. One important advantage of the gazetteers is that they support clinicians' shorthand expressions and idiosyncratic forms of expression not otherwise recognisable, as well as non-medical categories of information, e.g. occupations, named institutions, etc.

1.4 Dictionaries

Besides SNOMED CT and the gazetteers, several other external resources have also been adopted to identify medical terms and common words. They are:

1. Unified Medical Language System (UMLS) dictionary (Zieman and Bleich 1997).
2. Common words (Moby) dictionary.
3. Besides the above dictionaries, some other custom dictionaries have also been adopted, which were generated from the clinical notes mentioned previously. One of the largest is the abbreviation dictionary. It contains nearly 1000 abbreviations with their matching full expressions. Moreover, the misspellings dictionary (a list of misspelt words and their corrected spellings) and the unknown-words dictionary (a collection of unknown words) also have been derived from the clinical notes. The unknown words lexicon is important, as the recognition of unverifiable words is a progressive activity, which is never completed. An external process for verifying unknown words is in place but it always lags behind the generation of new unknown words, which need to be recognised in the presentation layer.

2 System Architecture and Implementation

The ICNS is based on a Client/Server/Data Warehouse

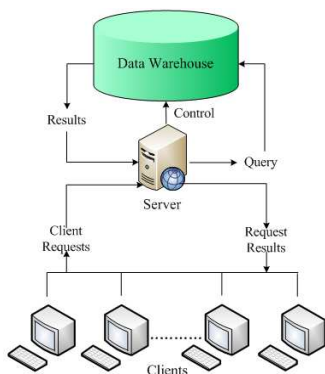


Figure 1. System Hierarchy

configuration. All the client requests are processed at the server side. In other words, the client computer has only one main task, which is to submit the user request, such as the retrieval request, annotation request etc. Meanwhile, all the substantive tasks, such as indexing, retrieval, proof reading and so on, are processed on the server side.

Furthermore, the server is also responsible for controlling the data stores. Figure 1 shows the system configuration.

The system architecture is divided into three main components. They are the data warehousing component, server system component, and client system component.

2.1 Data Warehousing

Figure 2 shows the workflow of the data warehouse processing. This component is responsible for the data transfer, index and annotation.

The basic process is: firstly, the original notes undergo the proof reading process. Proof reading provides correction for lexical formation, spelling corrections and canonicalization of non-words such as measurements. It also adds attributes to each token such as its semantic category.

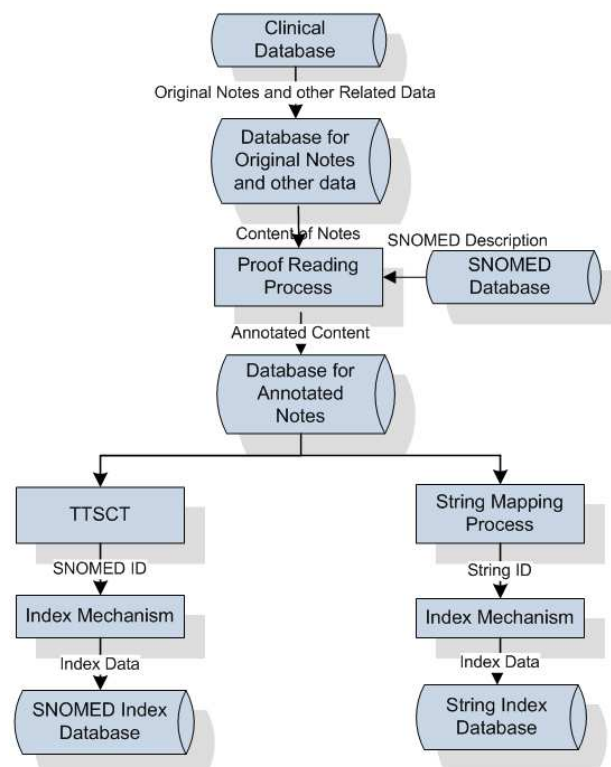


Figure 2. Data Warehouse Processing

Figure 3 presents the preprocessing required for the proof reading. The main purpose of this preprocessing is to extract four custom dictionaries from the manually appraised list of unknown words, which come from the ICU clinical corpus. The generated dictionaries will be used in the proof reading process. There are three main steps in the pre-processing:

1. The manual evaluation of unknown words is done in spreadsheets compiled to have a certain number of context examples for each unknown word. Manually, in one column the correct word is inserted and in another the gazetteer name is placed.
2. Subsequently, the columns are filtered for reuse where for example, duplicated rows and empty rows are deleted.
3. Finally, the reorganized rows are used to generate three dictionaries and approximately 50 gazetteers, which are used in the next stage.

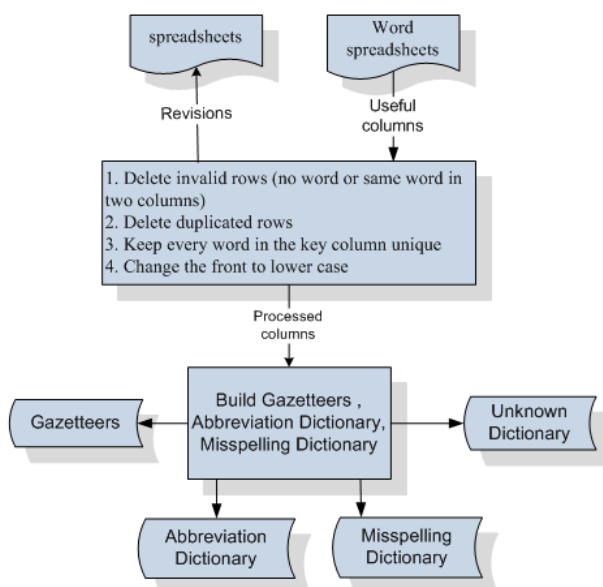


Figure 3. Proof Reading Proprocessing

The flow chart for the next stage of the proof reading is presented in Figure 4. Besides the unknowns lexicon, abbreviations lexicon, misspellings lexicon and gazetteers, three external resources also are used, they are the SCT (SNOMED CT) ontology, UMLS lexicon and Moby lexicon. The basic workflow for this process is:

1. The plain token is checked as to whether it belongs to the abbreviations lexicon. If not an abbreviation it will pass to the next step otherwise, this abbreviation will be expanded to the full name. Meanwhile, each word in the expanded name will go to the next step.
2. Next, if the token belongs to a gazetteer, the gazetteer class name will be added to that word's attribute list as an output for the annotated content. Otherwise, it will be passed to the next step.
3. Next, if it belongs to the SCT ontology, an SCT tag, Concept Id and description is attached as a tag attribute. This task is achieved by a query to the SCT terminology server.
4. In the fifth and sixth steps, the word will be checked with the UMLS dictionary and Moby dictionary respectively. If it matches either it will be tagged and exported.
5. In this step, the Misspellings lexicon is used to verify if it is misspelt in which case the corrected spelling will be returned. Subsequently, every word in this corrected form will go back to the second step. This is a recursion point in this proof reading processing.
6. Finally, if the word does not belong to any of the resources mentioned above, it is checked with the Unknown Dictionary. If it exists, an unknown tag will be added to it, otherwise, it will be stored in the Unknown dictionary and given an unknown tag. The strategy of keeping an explicit lexicon of unknown words is aimed at progressively identifying and reducing the list of unknowns though manual intervention.

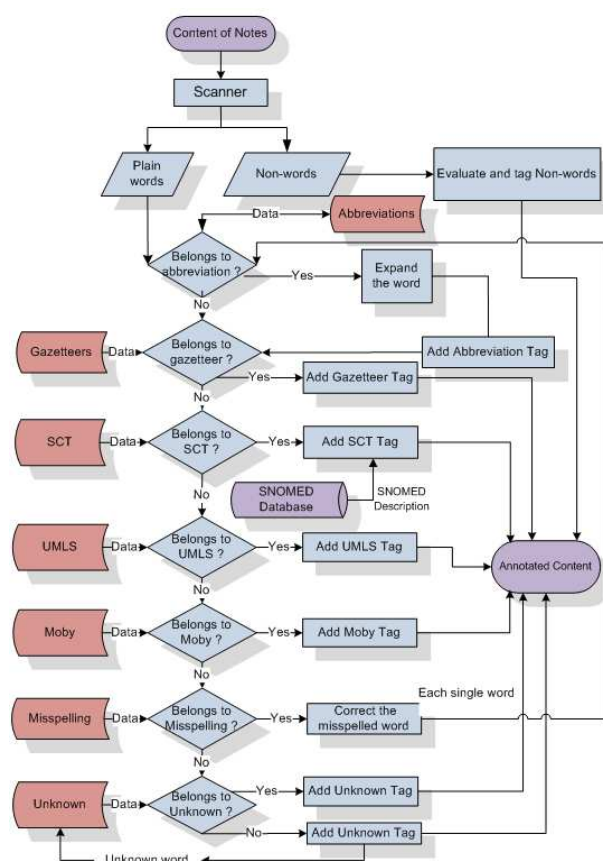


Figure 4. Proof Reading Processing

When the content of the notes is processed by proof reading, the output (annotated content) is stored in another database (see Figure 2). Subsequently, the annotated content is sent to the SCT index mechanism and String index mechanism, and the indices stored in their respective databases.

The SCT index is one of the most important parts of this system since the retrieval function is based on it. The main objective for the SCT index mechanism is to identify all the SCT concepts in the patient notes, and use the unique ID of the SCT concepts to build an index table pointing into the patient notes. In this mechanism, we use the TTSCT algorithm (introduced in section 1) to identify SCT concepts. When a document is processed by TTSCT, for each word or phrase in the notes the SNOMED fully specified name and unique concept ID is returned. Figure 5 shows an original note and the returned content after the TTSCT process is completed. As TTSCT is a statistical processor it will lead to returning some false positives as shown in the example. This leads to a level of error in the indexing of the notes and improvement in this error is the subject of separate research.

In order to utilize the content returned from TTSCT to build the index, some pre-processing is completed first:

1. For each returned text word/phrase strip out the SCT descriptions to leave just the Concept IDs. One text word/phrase may have more than one SCT concept match.
2. Delete the duplicate concept IDs from the index for a single document.

Once the concept IDs are isolated, the substantive task of building the SCT index can be completed.

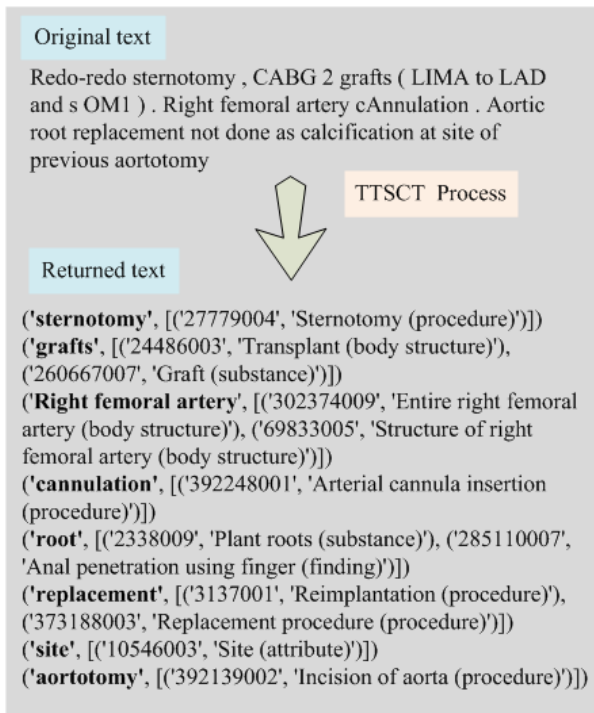


Figure 5. Input and Output of the TTST Process

2.2 Server System and Client System

The other two system components are the server system and the client system. The main task for the client system is to submit client requests. These include the retrieval requests and annotation requests. In contrast, the role of the server system is to process those requests and return the results from the data warehouse. As well, the user accounts management function and auto-completion of search expressions function are also fulfilled by the server system. Figure 6 is a diagram for the Client and Server system.

There are three main boxes in the diagram, which represent the three system components. The data warehouse component has already been introduced in the previous section. In this section, the server system component and client system component are discussed.

2.2.1 Client System

Generally speaking, the client system is a user interface to submit client requests and receive the reorganized results. The client requests include the retrieval requests, annotation requests, user registration requests and login/logout requests. Five retrieval keywords are available in the retrieval request, namely patients' MRN (medical reference number), SCT Concept, String, Care Provider Name and Storage Date. Alternatively, five types of annotation are included in the annotation request, namely, SCT Concepts, Abbreviation, Acronym, Gazetteer, and Unknown words annotation.

2.2.2 Server System

The box in the middle of Figure 6 is the server system component, which is mainly responsible for processing the client requests, managing the user accounts and organizing the results for presentation. Meanwhile, the server system

is also a bridge to connect the data warehouse component and client system component.

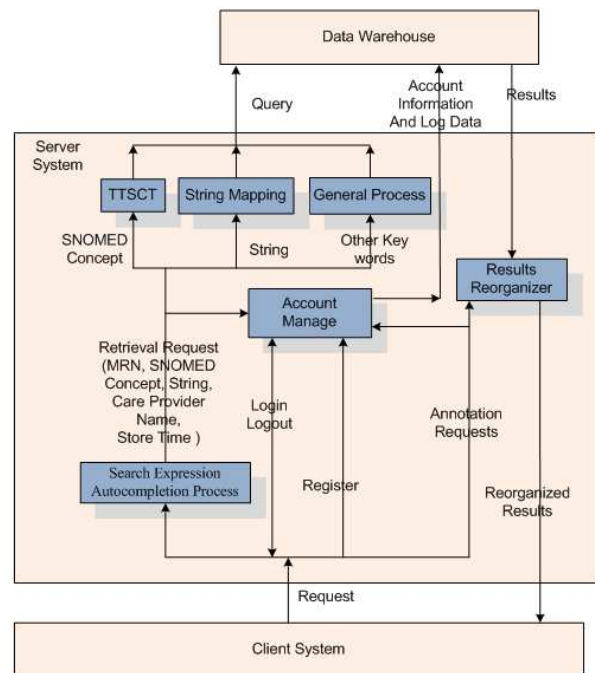


Figure 6 Client System and Server System

The basic workflow for the server system is that after it receives the registration and login requests from the client system, the retrieval requests will be processed by the search expression auto-completion process.

After input of the request data, the retrieval request will be divided into three categories, such as SCT Concept, String, and Other keywords (MRN, Care Provider Name and Store Time). The reason for this classification is that the SCT Concept retrieval and String retrieval cannot be processed by an SQL query directly. For example, the SCT Concept needs to go to the TTST first. When the concept ID is returned from the TTST, an SQL query can be used to read the SCT index and return the list of notes ID which contain this SNOMED concept. The String retrieval is similar, which is processed by the string mapping and index lookup initially. After the indices are retrieved an SQL query will be used to retrieve the relevant notes from the data warehouse.

1. Once the original results are returned from the data warehouse, the user can then request annotations to the text. These requests will be processed in the Results Reorganizer. Finally, the reorganized results will be returned to the client system for screen presentation to the user.
2. The server also has the process to record every user retrieval request and annotation request in a log file. This file is stored in the data warehouse.

3 The Interface and Functionality

Four web pages are designed as the user interface in the ICNS, namely the Main Page, User Registration Page, User Login Page and Retrieval Page. In this section, the Retrieval Page will be introduced in detail, as the main functions are accessed in this page. Figure 7 is a screen shot of the retrieval page.

In Figure 7, the retrieval page is divided into three main parts. The top of the image is the retrieval specification area. This area is used to submit the client retrieval request. The next area is the annotation area, which is located in the centre of the interface (beneath the purple line with “Corrected Spelling”). It is used to submit annotation requests. The remaining part is the results area. The notes index, retrieval results and annotation results are displayed in this area.

In the retrieval area, user friendly functions are available which are used to assist clients to submit their requests. The most important is the auto-completion for search expressions. With the help of this function, users don't need to know the exact spelling of a word (like SCT Concept) or entire number (like patient MRN). This function can predict these when users only input a part of the full spelling or digits. This function can be used in the MRN, SCT Concept, String, and Care Provider Name field.

Another user friendly function is implemented in the Storage Date field. Where users don't need to type the storage time themselves. Rather, a calendar will pop up for users to choose the desired date. There is auto-completion for search expressions in the SCT Concept field and the pop-up calendar in the Storage Date field.

Besides these functions, the retrieval function is one of the most significant functions in the retrieval area. There are two types of retrieval methods.

The first method is searching patient notes focused on one patient. In this retrieval method, a patient MRN is input, and hence, all the notes which are retrieved only belong to that particular patient. On input a list of encounters with their time periods for the patient is populated in the encounter panel just below the MRN window. Meanwhile, an index of notes for the selected encounter period, with their storage time, is displayed in the notes index window on the left side of the retrieval panel. In the encounter panel, multiple encounter periods can be chosen according the user's retrieval requirements. When these operations are finished, the user can go to “step2” of the retrieval area and fill in some input fields (See Figure 7 Retrieval notes for one patient). When the above operations are completed, the search button is clicked and the results will be displayed in the results panel.

The second retrieval type is the global and it searches the complete collection of notes in the database, so the patient MRN is not required. Clients can go to “step2” of the retrieval area directly.

Another function is the notes annotation, which is used to help clinicians extract useful information from the contents in the results area. This function is used in the annotation area. In the first instance, the original content of the notes will be displayed in the results area. However, when the corrected spelling checkbox is checked, the original notes are replaced by the spelling corrected notes. See Figure 8, where all misspellings are corrected.

After spelling correction, the notes can be annotated for the different categories, such as Abbreviations, Acronyms, SCT Concepts, Gazetteers and Unknown-words. This operation is done by clicking on the checkboxes that appear under the corrected spelling checkbox, thus providing multiple annotations in the annotation area.

Figure 8 is the screen shots for Abbreviation annotation, Acronym annotation and Unknown-words annotation.

The final function for this interface is the log function. The purpose of this function is to record every user request, which includes the retrieval requests and annotation requests. This information is stored into the database and can be exported as a log file. With the help of this function, we will learn which functions are used mostly and receive feedback from the users on the functions that are and are not successful by their assessment.

4 Evaluation

Currently, comprehensive evaluation for the ICNS has not been completed. However, a small user interview was held for an initial evaluation. During the evaluation, 2 doctors, one junior and one senior, were interviewed while they were operating the ICNS. Both of them believe the ICNS brings worthwhile convenience to their daily work, since if they want to search for a concept or keyword in the patients' notes without using the ICNS they need to read all notes for a given patient until they find the desired content. A sample demonstration given by the senior doctor was a question he had about the justification for a patient being on a certain antibiotic. He needed to find the laboratory results from previous “blood cultures” without knowing the time range of the search but excluding all notes not about the topic. The system recognised 5 notes out of 100+ notes with reference to blood cultures and automatically found the same content referenced by the abbreviation “bc” and orthographic variant “BC”.

Also, many good suggestions for improvements were given by the two doctors. The junior doctor mentioned some ideas for improving the user interface, for instance:

1. A search field for the patient name, since sometimes the doctor can only remember the patient's name.
2. Combination of the store time in the ‘Note Index’ with other information, such as the patient name and his/her care provider for each note. This would be much more user friendly giving better aid for identifying the desired note.

On the other hand, the feedback from the senior doctor was focused on the results from the retrieval functions:

1. When we search multiple words in the string search field or SCT search field, the issue of whether the multiple words should be treated as a single phrase or separate words needs to be considered.
2. Ambiguity of abbreviations/acronyms needs to be resolved, e.g. FROM (Free Range of Movement) is also a standard preposition.

Since these interviews a special window for collecting feedback has been built into the User Interface to gather the users' written feedback. Based on this information, we can do further investigations and research. We realize that only when the user interface is friendly enough, will the users be willing to use the new technology actively and persistently. In this way, it's very important to know the functions that the users like and dislike. We believe with the continuous help of this feedback, a sophisticated and comprehensive system can be persistently improved to serve the clinicians well.

5 Conclusion

In this paper, a universal system for clinical information extraction is presented. The Intelligent Clinical Notes System can be easily adapted for different clinical departments. The only change for the system is to connect to the data source. Furthermore, this system is easy to change due to its three independent components. For example, besides the introduced annotation categories, many other categories can be easily added into the interface (like UMLS, Digit, etc.) without changing the Data Warehouse and Server System components, because all these categories have already been annotated in the Data Warehouse processing. Meanwhile, the criterion for annotation can be easily modified by changing the external resources (such as the Abbreviation dictionary, Gazetteers, SNOMED Dictionary, etc.) without modifying the whole system. Currently, the Intelligent Clinical Notes System is installed in the Intensive Care Unit of Royal Prince Alfred Hospital for testing purposes.

6 Future Work

Although this system has many advantages, there are still many improvements that are needed in the future. Currently, we are designing the function to allow users to revise the annotated data, since sometimes there are some mistakes in the external resources (like abbreviations, gazetteers) which will lead to inaccurate annotation in the patient notes. This function will enable processing the revised data automatically. In other words, once the corrected data is submitted by the users, the system can modify the annotation resources which it is using immediately. Finally, this system is a part of a Health Information Technologies Research Laboratory (HITRL) large project. Some of the external resources, such as the custom dictionaries and gazetteers come from HITRL's other projects. We plan to change the system architecture so we can fetch the newest data from these projects and pass them to this system's data resources (such as the log file and unknown words collection) to improve the services automatically.

7 Acknowledgement

We would like to thank Drs. Robert Herkes, Michael O'Lerie, and Angela Ryan and other staff in the Intensive Care Unit of Royal Prince Alfred Hospital for providing the development and evaluation environment.

8 References

- Spackman, K. A., Campbell, K. E., & et, a. (1997): "snomed rt: A reference terminology for health care." *Proc AMIA Annu Fall Symp* 640(4).
- Patrick, J., Wang, Y., & Budd, P. (2007): An automated system for conversion of clinical notes into snomed clinical terminology. *In proceedings of the fifth australasian symposium on acsw frontiers*. 68: 219-226.
- Wua, C., Xie, Z., Todd, J., Johnson, B., Kuoa, G. M., & Jeffrey, R. (2004): Development of web-based snomed-ct post-coordinated code searching tool *MEDINFO*

Zieman, Y. L., & Bleich, H. L. (1997): Conceptual mapping of user's queries to medical subject headings *Proc AMIA*

Welcome admin! | [logout](#) | [history](#)

Intelligent Clinical Notes System - RPAH(ICU)

Step 1: Patient MRN:

Please choose one or more encounters:

Step 2:

Care Provider Name: SNOMED Concept:

String: Store Date:

Display Gazetteer: Return notes

NOTES INDEX

ChartTime	MRN	CareProvider
2009-02-20 14:09	1511731993	David Jones (Physic
2009-02-19 22:00	1511731993	David Jones (Physic
2009-02-19 10:00	1511731993	David Jones (Physic
2009-02-17 22:00	1511731993	David Jones (Physic

Corrected Spelling ☐

Patient Name: 1534 1535 1536 (1511731993) Care Provider Name: David Jones Create Time: 2009-02-19 10:00

NURSING :

Day7 admission with GI bleed 2 * ESLD.Care taken over at 1930hrs.Stable night .

NEURO:Sedated with fentanyl and midazolam.Propofol ceased.Fentanyl **infusion decreased** from 40 to 20mcg/hr.GCS-3-4.Hyperextension of upper and lower limbs with stimulus.PEARL 2mm brisk with hippus .

RESP:Chest clear.AE **decreased** bases.PSV 8/5/0.30 oral ETI.Loose sputum small amounts.RR 8-10.Tv>1000ml.Saturating>98%.Gas exchange good .

CVS:Afebrile.Peripherally warm well perfused.nsr rate 80-100.Normotensive MAP>65.Mil inotropes.Potassium replaced.CVP~3.Rij CVC and R)radial arterial line insitu.Maintenance fluids at 50ml/hr .

GII:Abdomen slightly distended.Bowel sounds present.No PR bleeding.Bowels not open.Linton tube off traction,balloon inflated.Promote feeds at 20ml/hr via gastric port.Aspirates minimal and appearance normal.Octreotide and pantoprazole **infusions** continue.BSL stable.Not requiring insulin .

RENAL:IDC draining clear yellow urine volumes>40ml/hr .

Please add your feedback here (and then click send) (Max 1000 characters)

Figure 7. Retrieved notes for one patient. “SNOMED Concept” retrieves any text related to the specified concept, whereas “String” retrieves only literal strings.

Corrected Spelling ☒

Annotate: **SNOMED** ☐ **Abbreviation** ☒ **Acronym** ☒ **Gazetteer** ☒

NURSING

Received care of **pt[patient]** @ 0730hrs . I/V/S .

T/F **gicu** ~ 0700hrs , post-op emergency **LSCS[lower segment**

36/52 gest.) . 5L blood **loss** in **OT[occupational therapy]**

extubated ~ 1130hrs . NIL complications .

GCS 14 . PEARL . Drowsy yet **rousable[rouse]** . Obeying com

Fent . Requirements high this am , weaned throughout am . P

currently 2mls/hr . For **PCA[patient controlled analgesia]**

SR~75bpm . Required Gel filling for sluggish **MAP[mean arterial pressure]** . Otherwise **MAP[**

mean arterial pressure]>70 . **CVP[central venous pressure]**~6 . Afebrile . Peripherally warm

. **Maintenance[maintenance]** @ 60mls/hr . **K + & Mg** replaced . **2xperipheral[peripheral]** **IVC[**

inferior vena cava] removed . Oxytocin ceased @ 1200hrs . 4/24 bloods . Latest **Hb[heart**

block] 109 , **INR[international normalization ratio]** 1.3 .

AE L=R . Currently self-ventilating with O2 2L via NP . **ABG[arterial blood gases]** stable .

Sats>98 % .

Polyuric . **BSL[blood sugar level]** stable .

Abdc soft , tender . Permitted **CF[cystic fibrosis]** diet as per O&G . Tolerated sip test .

Min PV **loss** . Combine **insitu[in situ]** .

TEDS & calf-compressors . **PA[posterior-anterior]** intact .

ALL ☒ **alt-orthography** ☒ **colloquial** ☒ **compound** ☒ **d_chartitems** ☒ **ex** ☒ **foreign** ☒ **infusions** ☒

Unknowns ☒

Figure 8. User Interface selections for Abbreviations (pink), acronym (purple) and unknown-word (blue) annotations, auto-corrected misspellings (in grey), and dropdown list of Gazetteers.

A Study on the Use of Search Engines for Answering Clinical Questions

Andreea Tutos¹

Diego Mollá²

Department of Computing, Faculty of Science
Macquarie University,
Sydney, NSW 2109

¹ Email: andreea.tutos@students.mq.edu.au

² Email: diego.molla-aliod@mq.edu.au

Abstract

This paper describes an evaluation of the answerability of a set of clinical questions posed by physicians. The clinical questions belong to two categories of the five-leaf high-level hierarchical Evidence Taxonomy created by Ely and his colleagues: Intervention and Non Intervention. The questions are passed to two search engines (PubMed, Google), two question-answering systems (MedQA, Answers.com's BrainBoost), and a dictionary (OneLook) for locating the answers to the question corpus. The output of the systems is judged by a human and scored according to the Mean Reciprocal Rank (MRR). The results show the need for question modification and analyse the impact of specific types of modifications. The results also show that No Intervention questions are easier to answer than Intervention questions. Further, generic search engines like Google obtain higher MRR than specialised systems and even higher than a version of Google based on specialised literature (PubMed) only. In addition, an analysis of the location of the answer in the returned documents is provided.

Keywords: Question Answering, Evidence Based Medicine, Search, Evaluation.

1 Introduction

Latest clinical guidelines urge physicians to practise Evidence Based Medicine (EBM) when providing care for their patients (Yu et al. 2005). Evidence Based Medicine implies referring to the best evidence from scientific and medical research that can assist in making decisions about patient care (Sackett et al. 1996). However, current practise of EBM is challenged by the large amounts of external evidence information available to the medical practitioner. The number of biomedical publications is increasing to the point where thousands of new articles are published daily world wide, and no human can keep up to date without help. Lowering the barriers to the use of evidence based knowledge has the potential of improving the quality of patient consultation at the point of care.

This work forms part of a student project in Macquarie University's masters unit ITEC810.

Copyright ©2010, Australian Computer Society, Inc. This paper appeared at the Australian Workshop on Health Informatics and Knowledge Management (HIKM2010), Brisbane, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 108, Anthony Maeder and David Hansen, Ed. Reproduction for academic, not-for profit purposes permitted provided this text is included.

A study about the major obstacles to answering doctor's questions about patient care with evidence (Ely et al. 2002) highlighted, among other factors, the excessive time required to find the information, the difficulty in formulating an adequate question according to recommended practise in EBM, and the difficulty of synthesising multiple bits of evidence into a clinically useful statement. All of these issues are targets of current research in text-based question answering. We envision a scenario whereby the practitioner would ask a question using his or her words, and the system would search for the evidence and present it in the most effective way.

Our project is a step towards assessing the potential of the use of question-answering technology to access external evidence stored in the Internet by studying the answerability of a set of 50 medical questions sourced from the Parkhurst Exchange¹ website. We study the relevance of answers located through two selected search engines: PubMed² and Google, two question-answering systems: MedQA³ and Answers.com's BrainBoost⁴, and a dictionary: OneLook.⁵ In the process we perform an initial study of the modifications required for the questions to facilitate the retrieval of the answers by the above tools.

Our work is related to the study by Yu & Kaufman (2007) who conducted a cognitive evaluation of four online engines on answering definitional questions. Yu and Kaufman's evaluation criteria included quality of answers, ease of use, time spent and number of actions taken to locate an answer. Their results showed that PubMed performed poorly, Google was the preferred system for quality of answer and ease of use, and MedQA surpassed Google in time spent and number of actions. Our study does not limit the input questions to definitional questions only. We use a wider range of questions belonging to the 'Evidence' node in the Evidence Taxonomy introduced by Ely et al. (2002). Further, we study the ability of freely available systems to provide documents containing the answer, and the relative position of the answer-bearing documents in the ranked list presented to the user. In addition, we explore specific types of query modifications that can be made to find the documents.

The structure of this document is as follows: Section 2 describes studies and concepts related to our project. Section 3 introduces the evaluation methodology employed in the study. The methodology details the corpus of questions and how it has been selected. It details the classification of candidate questions according to the 'Evidence' node in the Evidence Taxonomy. It also describes the selected systems and

¹<http://www.parkhurstexchange.com>

²<http://www.ncbi.nlm.nih.gov/pubmed/>

³http://monkey.ims.uwm.edu:8080/MedQA/query_qa.cgi

⁴<http://www.answers.com/bb>

⁵<http://www.onelook.com>

the reasons behind their selection. A description of question processing follows together with a section on answer extraction. Section 4 presents the results of the evaluation. Section 5 analyses the results. Finally, Section 6 provides a summary and an indication of lines of future work.

2 Background

2.1 Question Answering

There has been considerable research in the area of open-domain Question Answering (QA). This research has been mainly driven by the Text REtrieval Conference (TREC) (Voorhees 2001), and more recently by the Cross Language Evaluation Forum (CLEF) (Vallin et al. 2005), the workshops by the NII Text Collection for IR Project (NTCIR) (Kando 2005), the Document Understanding Conference (DUC) (Dang 2006), and the Text Analysis Conference (TAC) (Dang 2008). Open-domain QA initially focused on fact-based questions that expected short answers, but more recently (e.g. in DUC and TAC) questions allowed more complex answers that are the result of combining information from multiple documents. This is the sort of questions that are applicable to the biomedical domain.

The biomedical domain is a specialised domain that presents challenges and opportunities that make it a very useful area for researchers, together with the potential of being very beneficial to the users (Zweigenbaum 2003, Zweigenbaum et al. 2007, Mollá & Vicedo 2007). In particular, there are collections of documents which can be used as corpora for searching the answers. For example, MEDLINE is a collection of abstracts maintained by the US National Library of Medicine (NLM) that contains more than 17 million records dating back to 1966. There are also terminological resources such as NLM's Medical Subject Headings (MeSH), which contains an extensive list of diseases, drugs and treatments. And there are tools like PubMed Central⁶ which provides an interface to MEDLINE and incorporates query expansion using MeSH in an attempt to find documents that are related to the question. The time is ripe for the development of question-answering technology for the biomedical domain.

There have been some attempts to integrate question-answering technology to the medical domain. Some methods are based on the so-called PICO format to formulate the questions. The PICO format (Niu et al. 2003) has four components that reflect key aspects of patient care and which are recommended for the practise of Evidence Based Medicine: **P**rimarily **P**roblem, **I**ntervention, **C**omparison, and **O**utcome of intervention. Current systems presume a preliminary stage that converts the question to the PICO format that can be easily processed by the computer (Niu et al. 2003, Demner-Fushman & Lin 2007). However, not all clinical questions (even among those that are strictly evidence-based questions) can be mapped in terms of PICO elements (Huang et al. 2006). There is also evidence that even doctors may find it difficult to formulate the question in terms of the PICO format (Ely et al. 2002). Therefore, research focusing on the PICO format will first need to show that it is possible to automatise the analysis of questions into the PICO format, or at least to provide tools that would help the practitioner to formulate PICO questions. This work falls outside the scope of this paper and therefore we do not use PICO in our experiments.

⁶<http://www.ncbi.nlm.nih.gov/pmc/>

- I. Clinical (n=193)
 - A. General (n=141)
 - 1. Evidence (n=106)
 - a. Intervention (n=71)
 - What is the drug of choice for epididymitis?*
 - b. No Intervention (n=35)
 - How common is depression after infectious mononucleosis?*
 - 2. No Evidence (n=35)
 - What is the name of that rash that diabetics get on their legs?*
 - B. Specific (n=52)
 - What is causing her anaemia?*
- II. Non-clinical (n=7)
 - How do you stop somebody with five problems, when their appointment is only long enough for one?*

Figure 1: Evidence Taxonomy used to classify 200 questions from family doctors

MedQA (Yu et al. 2007) is a recent medical answering system that responds to definitional questions by accessing the MEDLINE records and other World Wide Web collections. It automatically analyses a large number of electronic documents in order to generate short and coherent answers in response to the input questions. The reason behind using definitional questions is that they are 'more clear-cut' as opposed to other types of clinical questions that can have large variations in their expected answers. MedQA relies on the IMRAD (Introduction, Methods, Results and Discussion) structure of biomedical articles to determine the relevance of an article to the search query. MedQA is the first system to integrate end-to-end QA technology including question analysis, information retrieval, answer extraction and summarisation techniques (Lee et al. 2006). The system includes a Web demo, but unfortunately the demo was often not functional when the experiments reported in the present study were carried out.

2.2 Evidence Taxonomy

Our work uses the Evidence Taxonomy created by Ely et al. (2002). This high-level, five-leaf hierarchy categorises medical questions that are potentially answerable with evidence. The hierarchy is presented in Figure 1, with the examples given in the original paper.

Ely et al. (2002) concluded that the 'Non-clinical', 'Specific' and 'No Evidence' questions are not answerable with evidence, while both categories of 'Evidence' ('Intervention' and 'No Intervention') are potentially answerable. 'Non-clinical' questions do not address the specific medical domain and 'Specific' questions require information from the patient personal record.

We have focused on the two evidence categories confirmed as being answerable with evidence according to Ely et al. (2002): 'Intervention' and 'No Intervention' questions. According to the Evidence Taxonomy, 'Intervention' questions are scenario-based, quite complex and they require complex answers that provide descriptions of possible treatments or recommended drugs. 'No Intervention' questions usually enquire about medical conditions or drugs, without asking for directions in managing a disease. They generally belong to the family of factoid questions for which short answers are usually expected.

TITLE: Is watermelon allergenic?

QUESTION: "A 16-year-old female patient had an urticarial reaction from watermelon. She now avoids eating it," writes ABDULRAHEM LAFTAH, MD, of Watson Lake, Yukon. "What substance in watermelon would have caused the attack, and are there other related foods she should now stay away from?"

ANSWER: Watermelon does contain allergenic proteins that could provoke an IgE-dependent urticarial response. You can refer the patient for allergy skin testing to determine if this fruit was indeed the culprit. Watermelon belongs to a family of foods associated with ragweed pollen. These include cantaloupe, honeydew, zucchini, banana, cucumber and chamomile tea. Individuals suffering from ragweed allergic rhinitis may develop symptoms, often mild, after eating these foods. This is particularly true during or following hay fever season, when their IgE to ragweed is the highest. PK

Figure 2: Sample of question and answer

3 Evaluation Methodology

3.1 Corpus of Questions

The corpus of questions of our study has been constructed from the question and answer list available on the Parkhurst Exchange website.¹ Parkhurst Exchange is a medical publishing website based in Canada that includes a collection of over 4,800 clinical questions and their answers provided by physicians. Since 1983 when it first started, it reportedly continues to develop strong relationships with top physicians across many medical disciplines.

To determine whether the output of a system contains the answer we relied on human judgement (the first author of this paper). To facilitate this judgement, we have selected clinical questions that address relatively simple health issues and have no complicated medical language. Figure 2 shows an example of a question and answer that we used in our study. As the figure shows, the answers are not simple factoids typical of current QA systems.

The website's medical questions appear grouped in over 30 categories such as Psychiatry, Oncology, Pediatrics, Endocrinology, etc. In our selection process we have opted for the 'Browse All' option which lists all questions sorted descending based on the date they have entered the collection. We have then picked questions that addressed areas that presented relatively straightforward enquiries. A list of examples is included in Table 1. We admit that the question selection process might have introduced bias in our corpus of questions and therefore the results presented in this study are of a preliminary manner and need to be verified with a larger set of questions.

Parkhurst Exchange contains mainly clinical questions asked by family doctors. We have mapped our selection to the Evidence Taxonomy tree. All of the questions were classified as belonging to the 'Intervention' (46%) and 'No Intervention' (54%) categories.⁷ This distribution is relatively close to the percentages of the study by Ely et al. (2000).

⁷Whereas all the questions we looked at were classified as either 'Intervention' or 'No Intervention', we didn't check whether *all* questions in Parkhurst Exchange could be classified this way.

Question	Category
Is watermelon allergenic	No Intervention
When to introduce solids to infants	Intervention
Should family doctors be immunized with Pneumovax and Menactra or Menjugate	Intervention
Can cell phones cause cancer	No Intervention
How much folic acid — 400 g, 1 mg, 5 mg — is recommended before conception and during pregnancy	Intervention
How to beat recurrent UTIs	Intervention
How to recognize autism in adults	No Intervention
Does skin colour affect vitamin D requirements	No Intervention

Table 1: Example of questions classified according to the Evidence Taxonomy

Is watermelon allergenic?

("citrullus"[MeSH terms] OR "citrullus"[All Fields] OR "watermelon"[All Fields]) AND allergenic[All Fields]

Figure 3: A simple PubMed query and its expanded form

3.2 Search Engines and Question Answering Systems

We have selected the systems to test based on a few guidelines. They needed to be available online and free of charge and also be able to accept natural language questions. We initially considered the possibility of transforming the questions into PICO format. However, this idea was later postponed due to the intrinsic problems of formulating a query into PICO as described in Section 2.1. Without the option of mapping the input questions to the PICO format, selecting systems that accepted natural language questions became a must.

PubMed is a search engine that accesses a reputable medical repository (MEDLINE) maintained by the US National Library of Medicine (Demner-Fushman & Lin 2007). The MEDLINE database includes over 19 million medical articles and is a well recognised knowledge source across medical question answering studies. PubMed uses MeSH to expand the query with related terms. Figure 3 shows an example of a query and its expanded form.

Google is a popular web search engine that uses text matching techniques to locate web pages relevant to a user's search. Google's architecture includes a list of features that make it an effective search engine. First to be mentioned is the ability to determine quality rankings or PageRanks for each web page based on the link structure of the Web. Another characteristic of Google is that it establishes a relation between the text of links and the pages the links point to (Brin & Page 1998).

Two variants of Google were included in our study:

the standard Google and Google pointed towards the PubMed database. The reason for the second variation was the observation that quite often Google returned information from consumer-oriented web sites rather than scientific articles and publications. To make the results easier to compare, Google was pointed to search for information against PubMed (MEDLINE) database, ensuring compatibility with the results provided by the PubMed search engine itself. Using Google on PubMed only also addresses any possible concerns about the quality of the information provided by user-oriented sites indexed by Google.

MedQA (Yu et al. 2007) is one of the first developed end-to-end medical answering systems and responds to definitional questions accessing the MEDLINE records and other World Wide Web collections. It automatically analyses a large number of electronic documents in order to generate short and coherent answers in response to the input questions.

The MedQA system proved to be quite unstable, producing parse errors or simply becoming frozen during an answer search cycle. As a result, the evaluation of its performance is not entirely relevant. A subsequent attempt to rerun all questions through the answering system proved even more unsuccessful, as we were unable to obtain any answers due to a 404 HTTP error ("the requested resource is not available").

Answers.com is a website that offers useful answers to categories of questions like business, health, travel, technology, science, entertainment, arts, etc. Their collection includes over four million answers drawn from over 180 titles from brand-name publishers, together with content created by their own editorial team. Apart from its repository of questions and answers, Answers.com also hosts BrainBoost, a generic end-to-end question answering system that highlights the answer to the user's questions. In our experiments we used BrainBoost⁴ rather than the general answers.com site.

OneLook is a dictionary and translation meta-search engine that accesses more than 900 online dictionaries in order to locate the desired definition. It offers the ability to decide on the dictionary to focus on, with choices of domains as medical, art, business, etc, though we did not use this feature in our experiments.

3.3 Question Processing

Turning knowledge into specific requests for information is not always an easy task. Some information needs are difficult to express and when they can be expressed, the way the question is interpreted influences the delivered answers. Yu et al. (2005) calculated an average of 2.7 different ways of expressing generic General Practitioner's clinical questions. The same study mentions the difficulty of explaining the context of the questions to the information source.

Question processing is therefore an important and difficult task in QA. The specific task of automated classification of clinical questions still has room for improvement, as illustrated by the results reported by Yu et al. (2005) on the classification of questions according to the Evidence Taxonomy (less than 60% accuracy for the five-category classification), and the results by Yu & Cao (2008) on a different taxonomy by the National Library of Medicine (76% F-score). These results are below those of generic question classification systems such as the one by Li & Roth (2002)

(up to 98.80% accuracy for a six-category classification). Question analysis is the largest source of errors in generic question-answering systems, with over 50% of the errors attributed to this stage by Moldovan et al. (2003). We therefore expect an even larger impact of question processing for medical question answering.

To mitigate the difficulties of question processing in our study, we applied query modification to every question that did not produce any relevant results when run unmodified through all systems. Then, exactly the same (possibly modified) question was sent to all systems. We did this to obtain results that are comparable across all systems.

We applied five levels of processing, in this order, until a system returned relevant results:

1. Introduce synonyms, hyponyms, and hypernyms of the medical terms in an attempt to improve the performance of the search. Example: we replaced *infectious conjunctivitis* with *bacterial conjunctivitis*.
2. Expand any abbreviations that might decrease the system's ability to find answers. Example: we replaced *BP* with *blood pressure*.
3. Add general medical terms such as *disease*, *syndrome* or *condition* to help clarify the target of the search query. Example: *What is shoulder frozen* was replaced with *What is frozen shoulder syndrome*.
4. Eliminate additional grammatical terms such as adverbs and prepositions from the original question. Example: the original question *Are there any contraindications to dental office visits in pregnancy* was modified to *Dental office visits in pregnancy*.
5. Use external knowledge to transform the question as an attempt to express the medical context. Example: *What is the evidence that antibiotics change the course of the disease in infectious conjunctivitis* became *Are antibiotics recommended for bacterial conjunctivitis*.

The query modifications were made manually. To source relevant words we used the online dictionary MedLinePlus.⁸ MedLinePlus has extensive information from the National Institutes of Health and other trusted sources on over 750 diseases and conditions and is a service offered by the US National Library of Medicine.

A summary of the five levels of question processing is shown in Table 2.

In order to evaluate the efficiency of our question processing and the degree to which each defined level of transformation had a positive impact on the search results, we have analysed the questions that did not produce any relevant answers when run through the systems in their original form. We have then determined which level of transformation has been applied in order to get a relevant answer. If after applying a particular level of processing, we have obtained a relevant answer or link, we have flagged that question as being improved by Level *x* of transformation. In order to quantify if there was an improvement, we did not consider the position of the relevant link on the results page and did not try to improve the relevant link position in the list by applying a subsequent level of processing. After computing the results we have observed that Level 4 of processing "Eliminate additional terms" was applied with the highest frequency

⁸<http://medlineplus.gov/>

Level	Description	Original Question	Processed Question
1	Introduce synonyms/hypernyms	infectious	bacterial
2	Replace abbreviations	BP	blood pressure
3	Introduce general medical terms	What is shoulder frozen	What is shoulder frozen syndrome
4	Eliminate additional terms	Are there any contraindications to dental office visits in pregnancy	Dental office visits in pregnancy
5	Express medical context	What is the evidence that antibiotics change the course of the disease in infectious conjunctivitis	Are antibiotics recommended for bacterial conjunctivitis

Table 2: Question processing levels

(45.95% of total successful transformations), followed by Level 5 “Express medical context” (27.03% of total successful transformations). The results are presented in Table 3.

3.4 Answer Extraction

In our attempt to locate answers to our corpus of questions, we have established a limit of 10 first links returned in response to a query. Any other links past this limit, relevant or irrelevant, have been ignored. Any relevant links that refer to a scientific article but do not have an abstract available have been ignored. We have set this rule as usually, if the abstract of the article is not available, the attempt of viewing the full text of the publication fails, requesting a registered username and password.

Most of the systems included in the study would return a list of links that then need to be evaluated in order to determine their relevance to the query. This is a time consuming process that MedQA, as a question answering system, manages to overcome by providing a summarised and concise answer. For some instances of our searches, when PubMed returned only one link in response to a query, the abstract was automatically displayed and we were able to locate the answer.

4 Results

In order to evaluate the results of our retrieval systems, we have used the Mean Reciprocal Rank (MRR), an evaluation measure frequently used in question-answering evaluations and first introduced in TREC (Voorhees 2001). If a link returned by a search was in the n th ($n \leq 10$) position in the list of results, and it was evaluated as being relevant to the question using the Parkhurst Exchange answers as a benchmark, it was given a score of $1/n$. We have adopted this methodology in order to assess the ranking system of each system. The further down the list, the more effort required from the user to locate the answer. Our evaluation includes the “ease of use” component in our scoring system.

In order to evaluate whether a summarised answer or a link returned in response to a search query is relevant, a human judge (the first author of this paper) has referred to the answer provided by the Parkhurst Exchange website. We initially opted for a lenient evaluation, in the sense that a link or summarised answer that was relatively relevant to the question received a score that was giving them a credit lower

than the 10th position of a relevant answer in the top 10 list: $1/11$. However we have later revised this scoring system as we came to the conclusion that it was possible that this methodology was introducing bias in our evaluation. We have decided to stick with the strict evaluation that only gives credit to links or summarised answers that express the same ideas as the Parkhurst Exchange benchmark answer. This decision was also supported by the limited medical knowledge of the human judge, which diffculted a comprehensive evaluation and judgement of diagnosis, drugs and treatments that are related to the search question.

After processing the 50 medical questions through all the selected systems, we have obtained a total of 119 answers.

The results of our evaluation are presented in Table 4, for the two evidence categories our corpus of questions was mapped to. They have been calculated as an average of scores, per question category and system.

The results of the actual location of the answer in a scientific article are shown in Table 6. This table shows the percentage of occurrences of the answer in a specific section (note that a document did not necessarily have all the sections listed in the table). Our results show that the answer can be located in one of the following sections: abstract, results, conclusions, recommendations, purpose or methods. The abstract was the section that most likely contained the answer.

The results of Table 6 do not refer to answers located in consumer oriented websites which do not follow a set document structure. They have been obtained after analysing the answers extracted from medical scientific articles which represent 34% of our total number of answers.

5 Discussion of Results

The results are summarised in Table 5 and show the following:

Google performed better than the other systems tested for both Intervention and Non-Intervention questions. Google on PubMed also has the second place for both Intervention and No Intervention questions, showing that Google still seems to be a comparatively good system.

PubMed was outperformed by Google on PubMed. Analysing the detailed results, we concluded that Google’s advantage was mainly due to PubMed returning the relevant links further down in the list and consequently obtaining a lower score than Google on

Processing level	1	2	3	4	5
How often level was applied	5.41%	10.81%	10.81%	45.95%	27.03%

Table 3: Question Processing Results

	PubMed	OneLook	Answers.com	MedQA	Google	Google on PubMed
No Intervention	0.27	0.04	0.38	0.04	0.80	0.41
Intervention	0.24	0.04	0.10	0.04	0.54	0.35

Table 4: Question Scores (MRR@10)

	Source	Position
Intervention	Google	1
	Google on PubMed	2
	PubMed	3
	OneLook	4
	MedQA	4
	Answers.com	6
No Intervention	Google	1
	Google on PubMed	2
	Answers.com	3
	PubMed	4
	OneLook	5
	MedQA	5

Table 5: Overall scores

PubMed. Apparently the ranking algorithm adopted by PubMed is not performing as well as Google's. This is in line with the observation by Plikus et al. (2006) that concluded that PubMed does not produce well classified search outputs and proposed PubFocus to help ranking by adding publication quality attributes. Table 7 provides some examples of rankings for the same link in PubMed as opposed to Google on PubMed. It is quite obvious that in those instances Google assigned a better score than PubMed for the same relevant link.

The systems were generally bad at detecting acronyms. Apparently any possible advantages of PubMed's automated query expansion were offset by our manual modification of the query. We indeed observed that PubMed did not detect acronyms, and often questions presenting acronyms were processed satisfactorily only after manual acronym expansion. Analysing the PubMed search engine behaviour we noticed that even if MEDLINE benefits from MeSH, the controlled vocabulary thesaurus, by expanding the query with related terminology, it still underperforms in retrieving relevant answers for questions using acronyms. An example is the original question *Is it a good idea to take ASA before an extended period of air travels* in which ASA is the medical acronym for *acetylsalicylic acid*. We expected that Google, as a generic search engine, would not be able to handle the acronym, but PubMed was also not able to translate it and could not provide any answers until we manually processed the question and replaced the abbreviation in the question with the explicit chemical substance name.

A rather surprising observation was that Google outperformed Google on PubMed. We attribute this to the much larger corpus of text indexed by Google as compared with Google on PubMed. Google's search system and ranking of results is designed for large volumes of highly hyperlinked data and therefore the reduced data in PubMed may affect its ability to rank the results effectively. However, even though Google obtained the best MRR scores, we still need to evalu-

ate whether Google's returned text would be acceptable to a medical practitioner.

Generally, systems were better on 'No Intervention' questions. Analysing the performance of the generic search engines and question answering systems, we observed that Google performed better on 'No Intervention' questions with an average score of 0.80 as opposed to 0.54 for 'Intervention' questions. Answers.com also proved better on 'No Intervention' questions than 'Intervention' questions. The results show that the systems have more difficulties on producing answers for scenario-based, complex medical questions.

Overall Answers.com performed much better than OneLook and this could be explained by the fact that Answers.com is designed to answer questions and incorporates question-answering techniques (Brain-Boost), whereas OneLook is basically a dictionary and therefore only able to handle definitions. In particular, OneLook only managed to answer two questions out of the 50 included in our corpus questions: one 'Intervention' question and one 'No Intervention' question. These results show that OneLook is currently not suitable as a potential technology for medical answering systems.

MedQA obtained one of the worst scores, but as mentioned earlier, this was mainly due to the fact that the online link was not always up and running.

We have also found out that, after manual query modification, all the questions in our corpus of questions were answerable with current technology, which we consider to be an important finding for future medical question answering systems as it indicates the potential benefit of developing question answering systems. The most effective query transformation consisted of eliminating noise from the query (45.98% of questions), followed by using expert knowledge to transform the query (27.03%). An example is the original question *What's the best antihistaminic for mild acute urticaria in infants and children?* for which PubMed could not locate an answer until we have transformed it into *antihistaminic for mild acute urticaria in children*. The impact of introducing synonyms was relatively low (5.41%).

Going further to the actual location of the answer in medical articles, we have determined that the probability of the answer to be located in the Abstract section of an article is 50%, Conclusions section 26.19% and Results section 14.29%. This gives a good indication on the areas a question answering medical system should look most of the time for answers to ad-hoc queries.

Our study results have been compiled on a small set of 50 questions and we admit this might introduce some bias in our process. Our results will have to be confirmed and compared to the performance obtained on a larger corpus of questions. For a more confident evaluation, we recommend random sampling of the Parkhurst Exchange or another corpus that more closely reflects the characteristics of questions asked

	Abstract	Results	Conclusions	Recommendations	Purpose	Methods
Non Intervention	43.48%	17.39%	26.09%	0.00%	8.70%	4.35%
Intervention	57.89%	10.53%	26.32%	5.26%	0.00%	0.00%

Table 6: Percentage of answer location in scientific articles

Question No.	Question	Category	PubMed	Google on PubMed
19	When should moles be removed?	Intervention	7	2
43	What can be done for a patient with persistent (non-typhoid) Salmonella in stool, despite 2 antibiotics	Intervention	7	4
48	Is it a good idea to take an ASA before an extended period of air travel?	Intervention	5	4

Table 7: Google on PubMed vs. PubMed ranking for the same relevant link

by medical doctors.

6 Summary and Future Work

We have presented a study of the answerability of documents returned by current freely available technology within the domain of clinical question answering. Our study shows that current technology is able to find the answers to the asked questions, though the questions need to be transformed. We have applied a set of question transformations and evaluated the impact of transformation on the answerability of results returned. It is our intention to explore methods to automatically perform such transformations to increase recall.

Our study also includes an analysis of the location of the answer. This analysis needs to be extended by considering the type of question and other factors and narrow down the actual zoning that could be done to find the answer.

We obtained the surprising result that Google performs best than any other systems, including PubMed which is specialised on medical text, and even including Google on PubMed documents only.

Future work includes an evaluation of PubMed enhanced with an optimised ranking system such as the one provided by PubFocus. We would then compare these results with the previous PubMed performance and of Google on PubMed.

Another line of future work is determining the actual quality of the answers returned by a particular search engine or question answering system. The study presented was based on MRR as the measure to compare all systems. Note, however, that MRR is only concerned with the location of the document containing the answer but it does not measure the quality of the answers presented or the impact that the erroneous answers may produce in the judgement of the medical doctor. It is therefore desirable to evaluate the acceptance of the answer returned by the medical practitioner, and any possible errors of judgement that non-relevant texts could introduce. The study of answer quality will also help to determine the best technology to extract the answer and present it to the user.

References

Brin, S. & Page, L. (1998), The anatomy of a large-scale hypertextual web search engine, in 'Proc. WWW-7', Brisbane, Australia.

Dang, H. T. (2006), DUC 2005: Evaluation of question-focused summarization systems, in 'Proceedings of the Workshop on Task-Focused Summarization and Question Answering', Association for Computational Linguistics, Sydney, pp. 48–55.

Dang, H. T. (2008), Overview of the tac 2008 opinion question answering and summarization tasks, in 'Proc. TAC 2008'.

Demner-Fushman, D. & Lin, J. J. (2007), 'Answering clinical questions with knowledge-based and statistical techniques.', *Computational Linguistics* **33**(1), 63–103.

Ely, J., Osherooff, J. A., Ebell, M. H., Chambliss, M. L., Vinson, D., Stevermer, J. J. & Pifer, E. A. (2002), 'Obstacles to answering doctors' questions about patient care with evidence: Qualitative study', *BMJ* **324**(7339), 710.

Ely, J., Osherooff, J. A., Gorman, P. N., Ebell, M. H., Chambliss, M. L., Pifer, E. A. & Stavri, P. Z. (2000), 'A taxonomy of generic clinical questions: Classification study', *British Medical Journal* **321**(7258), 429–432.

Huang, X., Lin, J. & Demner-Fushman, D. (2006), Evaluation of PICO as a knowledge representation for clinical questions, in 'AMIA Annu Symp Proc.', pp. 359–363.

Kando, N. (2005), Overview of the fifth NTCIR workshop, in 'Proceedings NTCIR 2005'.

Lee, M., Cimino, J., Zhu, H. R., Sable, C., Shanker, V., Ely, J. & Yu, H. (2006), Beyond information retrieval — medical question answering, in 'Proc. AMIA 2006', p. 6 pages.

Li, X. & Roth, D. (2002), 'Learning question classifiers', *Proc. COLING 02*.

Moldovan, D., Pasca, M., Harabagiu, S. & Surdeanu, M. (2003), 'Performance issues and error analysis in an open-domain question answering system', *ACM Transactions on Information Systems* **21**(2), 133–154.

Mollá, D. & Vicedo, J. L. (2007), 'Question answering in restricted domains: An overview', *Computational Linguistics* **33**(1), 41–61.

Niu, Y., Hirst, G., McArthur, G. & Rodriguez-Gianolli, P. (2003), Answering clinical questions with role identification, in 'Proc. ACL, Workshop on Natural Language Processing in Biomedicine'.

- Plikus, M., Zhang, Z. & Chuong, C. M. (2006), 'PubFocus: Semantic MEDLINE/PubMed citations analysis through integration of controlled biomedical dictionaries and ranking algorithm', *BMC Bioinformatics* **7**(1), 424.
- Sackett, D. L., Rosenberg, W. M., Gray, J., Haynes, R. B. & Richardson, W. S. (1996), 'Evidence based medicine: What it is and what it isn't', *BMJ* **312**(7023), 71–72.
- Vallin, A., Magnini, B., Giampiccolo, D., Aunimo, L., Ayache, C., Osenova, P., nas, A. P., de Rijke, M., Sacaleanu, B., Santos, D. & Sutcliffe, R. (2005), Overview of the CLEF 2005 multilingual question answering track, in 'Proceedings CLEF 2005'. Working note.
- Voorhees, E. M. (2001), 'The TREC question answering track', *Natural Language Engineering* **7**(4), 361–378.
- Yu, H. & Cao, Y.-g. (2008), Automatically extracting information needs from ad hoc clinical questions, in 'AMIA Annu Symp Proc.', pp. 96–100.
- Yu, H. & Kaufman, D. (2007), A cognitive evaluation of four online search engines for answering definitional questions posed by physicians, in 'Proc. Pacific Symposium on Biocomputing', pp. 328–339.
- Yu, H., Lee, M., Kaufman, D., Ely, J., Osheroff, J. A., Hripcsak, G. & Cimino, J. J. (2007), 'Development, implementation, and a cognitive evaluation of a definitional question answering system for physicians.', *Journal of Biomedical Informatics* **40**(3), 236–251.
- Yu, H., Sable, C. & Zhu, H. R. (2005), Classifying medical questions based on an evidence taxonomy, in 'Proc. AAAI'05 Workshop on Question Answering in Restricted Domains'.
- Zweigenbaum, P. (2003), Question answering in biomedicine, in 'Proc. EACL2003, workshop on NLP for Question Answering', Budapest.
- Zweigenbaum, P., Demner-Fushman, D., Yu, H. & Cohen, K. B. (2007), 'Frontiers of biomedical text mining: current progress.', *Briefings in Bioinformatics* **8**(5), 358–375.

Customisable Query Resolution in Biology and Medicine

Peter Ansell¹, James Hogan¹ and Paul Roe¹

¹ School of Information Technology
Queensland University of Technology,

2 George Street, Brisbane, Queensland, 4000

Email: p.ansell@qut.edu.au, j.hogan@qut.edu.au and p.roe@qut.edu.au

Abstract

Scientists and healthcare workers regularly use data from a number of sources as part of their research and professional work. The current mechanisms for providing combined access to multiple datasources are either closed or not easily extensible, with some requiring users to locally load and query each datasource independently. In this work we introduce a new model for transparent querying across multiple datasources which relies on the single unifying format of RDF to merge information before returning it to users. The use of normalised, resolvable URI's, combined with the SPARQL RDF query language, enables common queries to be executed across multiple public and private datasources, including those not initially designed or represented using RDF. In order to accommodate a range of users, the implemented system has been set up to enable customisation of queries and datasources based on RDF formatted configuration files. This breadth of data and configurability allows scientists and healthcare workers to more efficiently find and communicate semantic references, supporting research, professional practice and dissemination of knowledge across communities and disciplines.

1 Introduction

The areas of science and medicine rely on information transfer between organisations to make sure each organisation is taking advantage of the latest innovations and discoveries. These sources of information are varied in nature and diverse in location, with users sometimes requiring data from many locations to make the best decisions. The ability to cross between disciplines, for example from medicine through to genomics and chemistry, generally requires a large amount of expertise, including an understanding of each of the relevant data formats. Particularly in medicine, organisations need to be able to utilise both external and local knowledge bases as part of their decision making processes. The combination of distributed, cross-discipline, and potentially private knowledge provides a case for a system designed around a single knowledge representation format. In this system users who are unfamiliar with a particular discipline can utilise a single query method to explore the information and decide on the importance of the information without requiring the intervention of a domain expert. The use of a single extensible format

enables organisations to insert their own, potentially private, information into documents without requiring a redesign of the file format or any external data disclosure.

Recently, there have been a number of datasources, representing science, medicine, and other areas (Auer, Bizer, Kobilarov, Lehmann, Cyganiak & Ives 2007, Belleau, Nolin, Tourigny, Rigault & Morissette 2008, Ruttenberg, Rees, Samwald & Marshall 2009), which have been republished using RDF (Resource Description Format). RDF is a domain-neutral information format that can be easily extended by organisations to include references to their own data where necessary without requiring them to customise the file structure to their particular needs. Queries across RDF datasources can be performed without a knowledge of the particular properties used by any of the relevant datasources. Knowledge of the properties used by particular datasources may, however, be used to integrate knowledge from multiple datasources into homogeneous documents. The ease of querying provided by RDF query languages such as SPARQL (SPARQL Query Language for RDF)¹, enables users to share and customise queries easily, and in some cases perform the same queries across datasources from different disciplines.

This paper presents a novel design for a cross-database, customisable query system based on RDF. It includes a description of the model and a prototype implementation, before attempting to show how they could be used and adapted to fit health informatics requirements. Section 2 provides some background into both RDF and non-RDF projects that attempt to integrate or describe large datasources. A brief description of the elements that make up the distributed query model, including ways to integrate other sources of information, is given in Section 3. The applicability of the distributed RDF query model to health informatics and clinical biologists is described in Section 4, along with a case study that starts with a drug and explores the corresponding links to genomics, cheminformatics, and clinical trial datasources. The interlinked datasources provide use cases and future research paths for both the drug and patients who may benefit from the drug. Section 5 discusses issues that relate to the model and its use in the context of health informatics.

2 Background

RDF versions of scientific datasources, have been created by projects such as Bio2RDF (Belleau et al. 2008), Neurocommons (Ruttenberg et al. 2009), Flyweb (Zhao, Klyne & Shotton 2008), and Linked Open

Copyright 2010, Australian Computer Society, Inc. This paper appeared at the Australasian Workshop on Health Informatics and Knowledge Management (HIKM 2010), Brisbane, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 108. Anthony Maeder and David Hansen, Eds. Reproduction for academic, not-for-profit purposes permitted provided this text is included.

¹<http://www.w3.org/TR/rdf-sparql-query/>

Drug Data (LODD)². Where possible, the RDF documents produced by these organisations utilise HTTP URI's to link to RDF documents produced by the other organisations. This is useful, as it matches the basic Linked Data³ goals which are designed to ensure that data represented in RDF is accessible and contextually linked to other data as required.

SPARQL queries that are required for complex investigations are in most cases limited to execution on a single RDF database. This project aims to simplify the distribution of queries across multiple datasources, given that it is impractical to expect every datasource to be copied to a single local database for complex queries. The majority of systems which attempt to distribute SPARQL queries across a number of endpoints, convert single SPARQL queries into multiple SPARQL queries, before joining and filtering the results to match the original query. These systems generally require that users configure the system with specific knowledge of properties used by each datasource, and in some cases the join may require a large amount of information to be moved, rendering the method relatively inefficient (Adamku & Stuckenschmidt 2005, Langeegger, Blochl & Woss 2007, Quilitz & Leser 2008). Other similar systems also require query designers to insert the URL's of each of the datasource endpoints into their queries by redefining the meaning of a SPARQL keyword, making complex datasources inaccessible (Zemánek, Schenk & Svátek 2008). Systems that focus on RDF query performance improvements from a parallelism point of view do not generally require users to use specific predicates, but they may require a suitable distribution of information across a local set of RDF endpoints in order to facilitate fairly random access to specific RDF statements (Harth, Umbrich, Hogan & Decker 2007). The number of requirements given by these query systems highlight the need for a simple method of customising and extending results, particularly in the case where private datasources must be accessed using a unique, secure, method that cannot be disclosed to other users.

The BioGUID project (Page 2009) provides a single point of entry for users to obtain RDF descriptions from a range of datasets. However, it does not provide a generalised mechanism for resolution, as the RDF resolvers are implemented as a set of tools rather than a single configurable implementation. Other projects such as the Distributed Annotation System (DAS) (Prlić, Birney, Cox, Down, Finn, Grf, Jackson, Khri, Kulesha, Pettett, Smith, Stalker & Hubbard 2006) allow distributed customised querying but do not use RDF, so discipline specific file formats must be understood by any software utilising the resulting documents. In comparison to an RDF based query solution, the current DAS implementation suffers in that it requires software updates to support any new classes of information being added to the system. In comparison, RDF based systems can be extended without having multiple data models implemented in software. RDF based solutions enable users to extend an official implementation in a completely valid way using their own predicates, enabling customised configuration-based additions as well as the up to date alternative annotation data that DAS was designed to provide.

The SRS system (Etzold, Harris & Beulah 2003) provides an integrated set of biological datasources, with a custom query language and internal addressing scheme. Although the internal identifiers are unambiguous in the context of the SRS system, they do not have a clear meaning when used in other con-

texts. In comparison to the many formats offered by SRS and the native document formats of particular scientific datasources, the use of RDF for both documents and query results provides a single method, URI's, to reference items from any of the involved datasources. The locally integrated model that SRS relies on for its queries is not sustainable as the size and number of datasources grows. The approximate number of RDF statements—similar in nature to SQL database records—that are required to represent each of the largest 14 databases in the Bio2RDF project as shown in Table 1, illustrating the scale of the information provided currently in distributed RDF datasources. SRS provides a generic query language, that makes use of the localised database, giving it performance advantages over the distributed RDF query system described in this paper.

Database	Approximate RDF statements
PDB	10,000,000,000
Genbank	5,000,000,000
Refseq	2,600,000,000
Pubmed	1,000,000,000
Uniprot Uniref	800,000,000
Uniprot Uniparc	710,000,000
Uniprot Uniprot	220,000,000
IProClass	182,000,000
NCBI Entrez Geneid	156,000,000
Kegg Pathway	52,000,000
Biocyc	34,000,000
Gene Ontology (GO)	7,400,000
Chebi	5,000,000
NCBI Homologene	4,500,000

Table 1: Bio2RDF datasource sizes

3 Model

In order to allow a simple method of performing and customising queries across the many potential datasources, an easily extensible model was designed and implemented. The model consists of a chain of elements starting with a user query, typically given as part of a URL, and ending with a set of RDF statements. The chain is started by matching the user query against a set of Query Types, any of which could be used in parallel to respond to the query. Each Query Type can be configured to identify the relevant namespace, as required, and utilise these namespaces as the basis of distributing the query across any applicable providers. For each provider, normalisation rules may be configured to be applied to the parts of the query that have been defined to be specific to each endpoint.

The resulting RDF information from each provider is then transformed back using the output part of the normalisation rules which were configured for the provider. The normalised information is then merged into an overall pool of RDF statements that will be returned in a single document to the user. The model focuses on pooling information, as this method provides the simplest way of retrieving information from datasources that may not all use the same query structure or interface. In order to provide for different users of a widely distributed set of configuration information, profiles have been included in order to allow varying levels of flexibility with respect to the inclusion or exclusion of Query Types, Providers and Normalisation Rules based on a user's goals. Profiles make it simple for users to customise the local configuration by adding their own RDF configuration

²<http://esw.w3.org/topic/HCLSIG/LODD>

³<http://www.w3.org/DesignIssues/LinkedData.html>

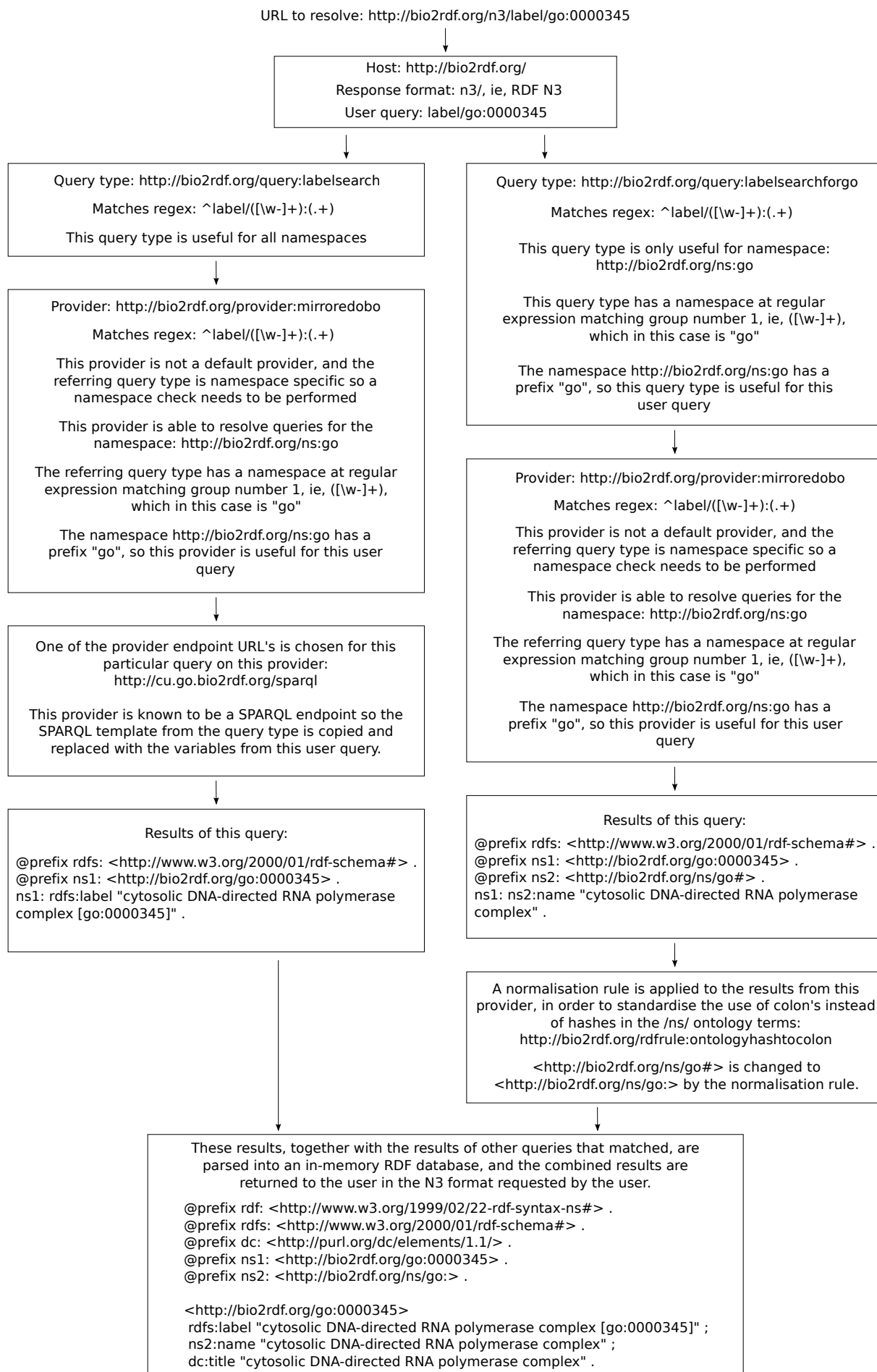


Figure 1: URL resolution using model

snippets.

An example illustrating the steps required by the model, to retrieve a list of labels for the Gene Ontology (GO) item with identifier “0000345”, known as “cytosolic DNA-directed RNA polymerase complex”, is shown in Figure 1. It illustrates the combination of a generic query, along with a query that is customised for the GO datasource. The queries are designed so that the generic query will be used on any information provider, while the custom GO query will be restricted to providers that contain GO information. If another datasource was available to retrieve labels for GO terms using RDF, then a custom query definition could be added in parallel to these two queries without any side effects.

A prototype server was implemented using Java and JSP and is currently in use by resolvers of the <http://bio2rdf.org/> website. The implementation allows users to select between the different RDF file formats, including an HTML page displaying a list of resources in the corresponding RDF document. The implementation was used to query across all of the Bio2RDF datasets, all of the LODD datasets, most of the Neurocommons (Ruttenberg et al. 2009) datasets, and the DBpedia (Auer et al. 2007) dataset (including the pagelinks set), using namespaces created using the <http://bio2rdf.org/> authority. The model defines the RDF statements required for implementations to use when creating configurations, meaning other implementations can utilise configurations created using the same version of the model vocabulary without reference to the implementation they were originally created or used on.

The simplest possible configuration consists of a single Query Type and a single Provider, as shown in Figure 2. The Query Type needs to be configured with a regular expression that matches user queries. The provider needs to be configured with both a reference to the Query Type, and an endpoint URL that can be used to resolve queries matching Query Type. Although the example trivial, in that the user’s query is directly passed to another location, it provides an overview of the features that make up a configuration. One particular feature to be noted is the use of the profile directive to process profile exclude instructions first, and then include in all other cases. In this example, there are no profiles defined, resulting in the items being included in the processing of queries that match the definitions.

```
@prefix query: <http://purl.org/queryall/query:> .
@prefix provider: <http://purl.org/queryall/provider:> .
@prefix profile: <http://purl.org/queryall/profile:> .
@prefix : <http://example.org/> .

:myquery a query:Query ;
  query:inputRegex "(.*)";
  profile:profileIncludeExcludeOrder
    profile:excludeThenInclude .

:myprovider a provider:Provider ;
  provider:resolutionStrategy provider:proxy ;
  provider:resolutionMethod provider:httpgeturl ;
  provider:isDefaultSource "true"^^<http://www.w3.org/2001/
XMLSchema#boolean> ;
  provider:endpointUrl "http://myhost.org/${input_1}";
  provider:includedInQuery :myquery ;
  profile:profileIncludeExcludeOrder
    profile:excludeThenInclude .
```

Figure 2: Simple system configuration in Turtle RDF file format

4 Applicability to Health Informatics

The model is designed so that it can be easily customised by users. Extensions can range from additional sources and queries to removal of sources or queries for efficiency or other reasons. This functionality provides a simple way for users to select which sources of information they want to use without having to make choices about every published information source.

In the context of Health Informatics, a hospital may want to utilise information from a drug information site, such as DrugBank and DailyMed, together with their private medical files. In order to do this, the hospital could map references in their medical files to DrugBank and/or DailyMed identifiers and publish the resulting information into RDF. The RDF statements could then be integrated with the DrugBank information without any further changes. They hospital may then create a mapping between the terminologies stored in their internal database and those used by DrugBank and related datasources. These mappings could be made using one of the available SQL to SPARQL converters such as the Virtuoso RDF Views mechanism (Erling & Mikhailov 2007) or the D2RQ server (Bizer & Seaborne 2004).

To distinguish private records from external public datasources, hospitals should create a namespace for their internal records, along with providers matching the internal addresses used for queries about their records. This novel private information, is able to be safely included in the model through the use of private provider configurations indicating the source and the particular RDF formats in which the information is available.

If the hospital then wanted to map a list of diseases into their files, they could find a source for disease descriptions, such as Diseaseome, and either find existing links through DrugBank and DailyMed, or they could attempt to use text mining to discover common disease names between their records and Diseaseome. For scientific research, resources like Diseaseome are linked to bioinformatics databases such as the NCBI Entrez GeneID, PDB, PFAM, and OMIM databases. This linkage is defined explicitly in RDF and enables links to be discovered, for instance there may be links from patients and clinical trials to genetic factors. Patients could be directly linked to genes using RDF syntax without reference to diseases, and new diseases could be described internally without requiring outside publication.

Potential side effects are accessible through the Sider database, enabling patients and doctors to both have access to equal information about the potential side effects of a particular course of medication. Side effects that are discovered by the hospital could be recorded using references to the public Sider record, reducing the possibility that the effect would be missed in future cases.

4.1 Case Study

The continuous use of RDF enables users to transition between databases using URI’s without having to register a new file format for each database. For ease of reading, the URI’s, which can be resolved to retrieve the relevant information used in the following case study, are footnoted. The links between the datasets were observed by resolving the URI’s and finding RDF statements inside the document that link to other URI’s. This case study utilises a range of datasets that are shown in Figure 3. These datasets are sourced from Bio2RDF, LODD, Neurocommons, and DBpedia. The original image can be found at <http://www4.wiwiw.fu-berlin>.

de/lodd/lodd-datasets_2009-08-06.png. A portion of the case study, highlighting the links between items in the relevant datasources, can be seen in Figure 4.

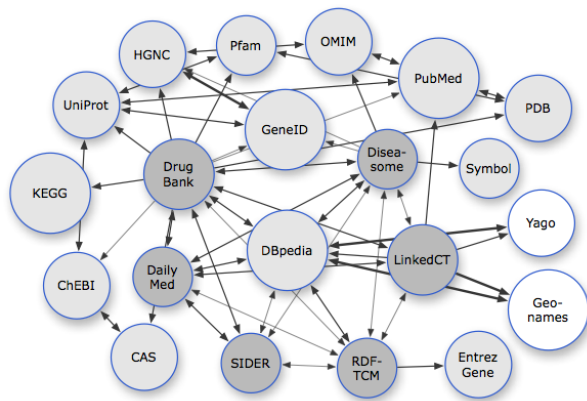


Figure 3: Medicine related RDF datasets

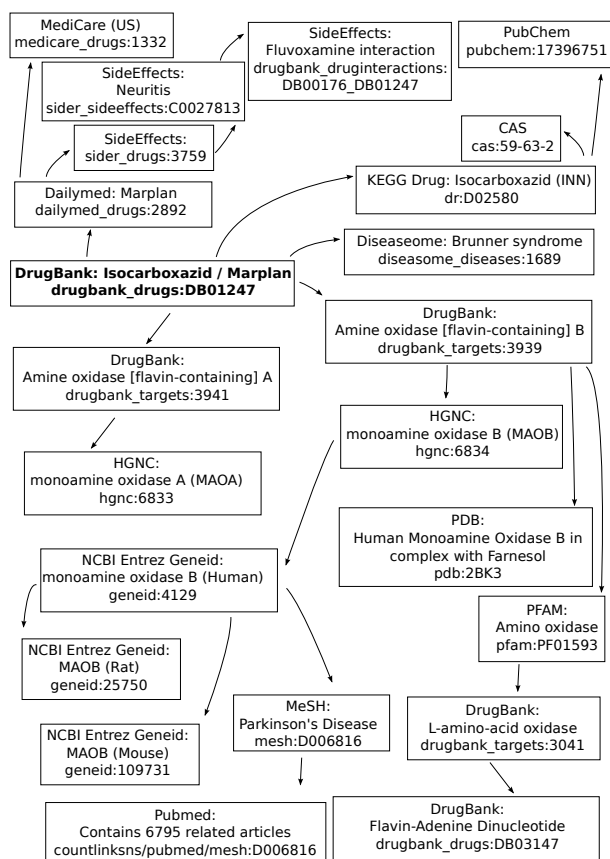


Figure 4: Inter-datasource links in Isocarboxazid case study

This case study is founded around a drug known generically as “Isocarboxazid”. It is also known by the brand name “Marplan”. The aims of this case study are to discover potential relationships between this drug and patients with reference to publications, genes and proteins that may affect the course of their treatment. In cases where patients are known to have adverse reactions or they do not respond positively to treatment, alternatives may be found by examining the usefulness of drugs that are designed for similar purposes. For the purposes of this case study, the relevant entry in DrugBank is known⁴. According to

this DrugBank record, Isocarboxazid is “[a]n MAO inhibitor that is effective in the treatment of major depression, dysthymic disorder, and atypical depression”.

The DrugBank entry for Isocarboxazid contains links to the CAS (Chemical Abstracts Service) registry⁵, which in turn contains links to the KEGG (Kyoto Encyclopedia of Genes and Genomes) Drug database⁶. The link back to DrugBank from the KEGG Drug database and others in this case could also have been discovered using only the original DrugBank namespace and identifier⁷. The brand name drug database, Dailymed, also contains a description for Marplan⁸, which is linked from Sider⁹ and the US MediCare database¹⁰. These alternative URI's could be used to identify more datasources with information about the drug.

The record for Isocarboxazid in the Sider database has a number of typical depression side-effects to watch for, but it also has a potential link to Neuritis¹¹ a symptom which is different to most of the other 39 side effects that are more clearly depression related. Along with side effects, there are also known drug interactions available using the DrugBank database. An example of these is an indication of a possible adverse reaction between Isocarboxazid and Fluvoxamine¹². If Fluvoxamine was already being given to the patient, other drugs may need to be investigated, as alternatives to prevent the possibility of a more serious Neuritis side effect. DrugBank contains a simple categorisation system that might reveal useful alternative Antidepressants¹³, in this case, such as Nor-tryptiline¹⁴.

Dailymed contains a list of typically inactive ingredients in each brand-name drug, such as Lactose¹⁵, which may factor into a decision to use one version of a drug over others. The Drugbank entry for Isocarboxazid also contains links to Diseaseome, for example, Brunner syndrome¹⁶, which are linked to the OMIM (Online Mendelian Inheritance in Man) entry for Monoamine oxidase A (MAOA)¹⁷.

DrugBank also contains a list of biological targets that Isocarboxazid is known to effect¹⁸. If Isocarboxazid was not suitable, drugs which also affect this gene; Monoamine oxidase B (MOAB)^{19,20,21,22}, the protein²³, or the protein family²⁴, might also cause a similar reaction. The negative link (in this case derived using text mining techniques) between the target gene, monoamine oxidase B,²⁵ and Huntington's Disease^{26,27}, might cause a doctor to decide not to give the drug to a patient with a history of Huntington's.

⁵<http://bio2rdf.org/cas:59-63-2>

⁶<http://bio2rdf.org/dr:D02580>

⁷http://bio2rdf.org/links/drugbank_drugs:DB01247

⁸http://bio2rdf.org/dailymed_drugs:2892

⁹http://bio2rdf.org/sider_drugs:3759

¹⁰http://bio2rdf.org/medicare_drugs:13323

¹¹http://bio2rdf.org/sider_sideeffects:C0027813

¹²http://bio2rdf.org/drugbank_druginteractions:DB00176_DB01247

¹³http://bio2rdf.org/drugbank_drugcategory:antidepressants

¹⁴http://bio2rdf.org/drugbank_drugs:DB00540

¹⁵http://bio2rdf.org/dailymed_ingredient:lactose

¹⁶http://bio2rdf.org/diseaseome_diseases:1689

¹⁷<http://bio2rdf.org/omim:309850>

¹⁸http://bio2rdf.org/drugbank_targets:3939

¹⁹<http://bio2rdf.org/symbol:MAOB>

²⁰<http://bio2rdf.org/hgnc:6834>

²¹<http://bio2rdf.org/geneid:4129>

²²<http://bio2rdf.org/mgi:96916>

²³<http://bio2rdf.org/uniprot:P27338>

²⁴<http://bio2rdf.org/pfam:PF01593>

²⁵<http://bio2rdf.org/geneid:4129>

²⁶http://bio2rdf.org/mesh:Huntington_Disease

²⁷<http://bio2rdf.org/mesh:D006816>

⁴http://bio2rdf.org/drugbank_drugs:DB01247

The location of the MAOB gene on the X chromosome in Humans might warrant an investigation into gender related issues related to the original drug, Isocarboxazid. The homologous MAOB genes in Mice²⁸ and Rats²⁹, are also located on chromosome X, indicating that they might be useful targets for non-Human trials studying gender related differences in the effects of the drug.

The Human gene MAOA, can be found in the Traditional Chinese Medicine (TCM) database,³⁰ as can MAOB³¹, although there were no direct links from the Entrez Geneid database to the TCM database. TCM has a range of herbal remedies listed as being relevant to the MAOB gene³² including *Psoralea corylifolia*³³. *Psoralea corylifolia* is also listed as being relevant to another gene, Superoxide dismutase 1 (SOD1)^{34,35}. SOD1 is known to be related to Amyotrophic Lateral Sclerosis^{36,37}, although the relationship back to the Brunner Syndrome and Isocarboxazid, if any, may only be exploratory given the range of datasources in between.

LinkedCT is an RDF version of the ClinicalTrials.gov website that was setup to register basic information about clinical trials. It provides access to clinical information, and consequently is a rough guide to the level of testing that various treatments have had. The drug and disease databases mentioned above link to individual clinical interventions in LinkedCT, enabling a path between the drugs, affected genes and trials relating to the drugs. Although there are no direct links from LinkedCT to Marplan at the time of publication, a namespace based text search returns a list of potentially interesting items³⁸. An example of a result from this search is a trial³⁹ conducted by John S. March, MD, MPH⁴⁰ of Duke University School of Medicine and overseen by the US Government⁴¹. The trial references published articles, including one titled "The case for practical clinical trials in psychiatry"^{42,43}. These articles are linked to textual MeSH (Medical Subject Headings) terms such as "Psychiatry - methods"⁴⁴, indicating an area that the study may be related to. The trial is linked to specific primary outcomes and the frequency with which the outcomes were tested, giving information about the scientific methods in use⁴⁵.

Although LinkedCT is a useful resource, as with any other resource, there are difficulties with the data being complete and correct. An example of this are recent studies about the use of ClinicalTrials.gov which indicate that a reasonable percentage of clinical trials either do not publish results, register with ClinicalTrials.gov, or reference the ClinicalTrials record in publications resulting from the research (Ross, Mulvey, Hines, Nissen & Krumholz 2009, Mathieu, Boutron, Moher, Altman & Ravaut 2009). These issues may be reduced if people were required to register all drug trials and reference the entry in any

publications.

Doctors and patients do not have to know what the URI for a particular resource is, as there is a search functionality available. This searching can either be focused on particular namespaces or it can be performed over the entire known set of RDF datasources, although the latter will inevitably be slower than a focused search as some datasources are up to hundreds of gigabytes in size, representing billions of RDF statements. An example of this may be a search for "MAOB"⁴⁶, which reveals resources that were not included in this brief case study.

5 Discussion

Given that the system uses URI's for all of the universal identifiers internally, and they are designed to all be resolvable, it is possible to show labels to humans, and have URI's for computers to use without prejudicing the system to either party. The RDF datasets have been designed with this in mind and the majority should include triples that indicate what the best label for a given resource is after it has been resolved. If an application recognises a URI as fitting the Bio2RDF system it can get labels for a URI using "label/namespace:identifier"⁴⁷. In order to reduce the latency involved with resolving a set of URI's, a list of labels can be resolved using the format "multiplelabel/namespace1:identifier1 / namespace2:identifier2..."⁴⁸.

Health Informatics requires that multiple systems be integrated in order to answer questions such as, "what observations were made for patients with heart disease, in the past 2 months, who were not on drugs that have current clinical trials". In order to answer this question multiple datasources may be required, including medical terminology repositories, patient databases, drug databases, and clinical trial databases. The mapping process may be an imprecise operation, as doctors may use ambiguous shorthand notations for their observations, and medical terminology repositories may not contain an exact term for a particular condition. This integration process may still benefit from the use of RDF, as a mapped URI, even if it is incorrect, is unambiguous, and can be identified immediately by resolving the URI in order to verify its suitability.

The ability to easily mix arbitrary sources of information provides both benefits and complications. The major benefits come from the ability to traverse the mixed dataset using a single file structure, RDF, and from being able to publish novel and mixed datasets in the same form for others to use. In the context of science and medicine this provides the ability to annotate studies and factual databases with extra information and provide that information to other users without having to republish the entire database or extend the file format originally used for the database. These benefits, however, also bring complications relating to provenance, privacy and reliability if any of the datasources or users are not trusted. The risk of these reliability complications can be reduced by only including queries and providers that are reasonably trusted.

The query model described here is designed to easily provide access to public datasources using configurations published by organisations such as Bio2RDF, while simultaneously allowing access to private internal datasources using unpublished configurations. Queries are executed using whatever permissions the

²⁸<http://bio2rdf.org/geneid:109731>

²⁹<http://bio2rdf.org/geneid:25750>

³⁰http://bio2rdf.org/linksns/tcm_gene/geneid:4128

³¹http://bio2rdf.org/linksns/tcm_gene/geneid:4129

³²http://bio2rdf.org/linksns/tcm_medicine/tcm_gene:MAOB

³³http://bio2rdf.org/tcm_medicine:Psoralea_corylifolia

³⁴http://bio2rdf.org/tcm_gene:SOD1

³⁵<http://bio2rdf.org/geneid:6647>

³⁶http://bio2rdf.org/mesh:Amyotrophic_Lateral_Sclerosis

³⁷<http://bio2rdf.org/mesh:D000690>

³⁸http://bio2rdf.org/searchns/linkedct_trials/marplan

³⁹http://bio2rdf.org/linkedct_trials:NCT00395213

⁴⁰http://bio2rdf.org/linkedct_overall_official:12333

⁴¹http://bio2rdf.org/linkedct_oversight:2283

⁴²http://bio2rdf.org/linkedct_reference:22113

⁴³<http://bio2rdf.org/pubmed:15863782>

⁴⁴<http://bio2rdf.org/mesh:D011570Q000379>

⁴⁵http://bio2rdf.org/linkedct_primary_outcomes:55439

⁴⁶<http://bio2rdf.org/search/MAOB>

⁴⁷<http://bio2rdf.org/label/pubmed:15863782>

⁴⁸<http://bio2rdf.org/multiplelabel/pubmed:15863782/omim:309860>

resolving server has, although authentication systems appropriate to each site could be used by modifying the code used for the resolving server to perform the authentication prior to performing the query.

Although some researchers deny that the use of RDF with URI's for entities such as humans and closely related records will cause new privacy issues, the ease and effectiveness with which different RDF documents can be merged has to be considered as a potential privacy issue (Feigenbaum, Herman, Hongsermeier, Neumann & Stephens 2007). They argue that the potential privacy issues are related to the applications that use the data, and that the issues are countered by the benefits that are obtained through the use of the integrated information. In the case of RDF though, the ability to merge documents is a feature, meaning that the merge of a patient's record with their credit history, for instance, may be far simpler than would be acceptable for patients. The URI's assigned to humans inside of the system should however be opaque and not contain identifying information such as age, location, or gender, which are particularly easy to map using RDF based reasoning tools in order to re-identify things. The overall RDF scheme has no recognised standard for stating what the rights and restrictions attached to a piece of information are, although there are a few schemes that attempt to perform this operation in limited circumstances. The model and implementation described in this work do not introduce new privacy issues, as the most important privacy issue in both cases is to prevent unauthorised bulk access to patient records, something which must be restricted using authentication and authorisation policies.

6 Conclusion

Knowledge management and informatics scenarios require that contextual information be provided to give evidence and background for particular sets of information. The aim of the distributed query system described here is to easily provide links into each of the relevant sets of information. The links are designed to make it easy for applications to cross traditional knowledge boundaries, such as the boundary between chemical theory and clinical pathology observations. The information required to convert a link into a document is all contained within the link reference, so applications only need to know how to resolve HTTP URL's in order to retrieve the related information. The document returned will always be in a known model, RDF, with a choice of file formats that represent the model, so applications can easily interpret the content without having to implement each known data structure in program code.

The steps required to integrate new sources of information into the query model include mapping current datasources to RDF; mapping relevant references to other datasources using URI's; and finally creation of configuration definitions describing the queries and locations that can be used to access information from the datasource.

The case study showed the diversity of links between the different health and biology related datasources currently available in RDF. It showed that the datasources were transparently accessible using the distributed query model and that URI based links between datasources can provide new insights that may not be easy to discover in current systems. Public datasources currently provide information about drugs, public drug trials, diseases, genetic information, and publications; while private datasources may provide linked information about patients, treatments, and private drug trials, in the

future. These can all be integrated by customising the distributed query model for each particular need through the use of simple RDF configuration files.

7 Acknowledgements

This research was funded through a Smart State National & International Research Alliance Scholarship by the Queensland State Government and Microsoft Research. It was supported by the Microsoft Queensland University of Technology eResearch Centre.

References

- Adamku, G. & Stuckenschmidt, H. (2005), Implementation and evaluation of a distributed rdf storage and retrieval system, in 'Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)', IEEE Computer Society, Los Alamitos, CA, USA, pp. 393–396.
- Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R. & Ives, Z. (2007), 'Dbpedia: A nucleus for a web of open data', *Lecture Notes in Computer Science* **4825**, 722.
- Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P. & Morissette, J. (2008), 'Bio2rdf: Towards a mashup to build bioinformatics knowledge systems', *Journal of Biomedical Informatics* **41**(5), 706–716.
- Bizer, C. & Seaborne, A. (2004), D2rq-treating non-rdf databases as virtual rdf graphs, in 'Proceedings of the 3rd International Semantic Web Conference (ISWC2004)', Citeseer.
- Erling, O. & Mikhailov, I. (2007), Rdf support in the virtuoso dbms, in 'Proceedings of the 1st Conference on Social Semantic Web (CSSW)', Springer, pp. 7–24.
- Etzold, T., Harris, H. & Beulah, S. (2003), 'SRS: An integration platform for databanks and analysis tools in bioinformatics', *Bioinformatics Managing Scientific Data* pp. 35–74.
- Feigenbaum, L., Herman, I., Hongsermeier, T., Neumann, E. & Stephens, S. (2007), 'The semantic web in action', *Scientific American* **297**, 90–97.
- Harth, A., Umbrich, J., Hogan, A. & Decker, S. (2007), 'Yars2: A federated repository for querying graph structured data from the webs', *Lecture Notes in Computer Science* **4825**, 211.
- Langegger, A., Blochl, M. & Woss, W. (2007), Sharing data on the grid using ontologies and distributed sparql queries, in '18th International Conference on Database and Expert Systems Applications, 2007. DEXA '07', pp. 450–454.
- Mathieu, S., Boutron, I., Moher, D., Altman, D. G. & Ravaud, P. (2009), 'Comparison of Registered and Published Primary Outcomes in Randomized Controlled Trials', *JAMA* **302**(9), 977–984.
- Page, R. (2009), Bioguid: resolving, discovering, and minting identifiers for biodiversity informatics. Available from Nature Precedings <http://hdl.handle.net/10101/npre.2009.3079.11>.
- Prlić, A., Birney, E., Cox, T., Down, T., Finn, R., Grf, S., Jackson, D., Khri, A., Kulesha, E., Pettett, R., Smith, J., Stalker, J. & Hubbard, T. (2006), *The Distributed Annotation System for Integration of Biological Data*, pp. 195–203.

- Quilitz, B. & Leser, U. (2008), 'Querying distributed rdf data sources with sparql', *Lecture Notes in Computer Science* **5021**, 524.
- Ross, J. S., Mulvey, G. K., Hines, E. M., Nissen, S. E. & Krumholz, H. M. (2009), 'Trial publication after registration in clinicaltrials.gov: A cross-sectional analysis', *PLoS Med* **6**(9), e1000144.
- Ruttenberg, A., Rees, J., Samwald, M. & Marshall, M. (2009), 'Life sciences on the semantic web: the neurocommons and beyond', *Briefings in Bioinformatics* **10**(2), 193.
- Zemánek, J., Schenk, S. & Svátek, V. (2008), Optimizing sparql queries over disparate rdf data sources through distributed semi-joins.
- Zhao, J., Klyne, G. & Shotton, D. (2008), Provenance and linked data in biological data webs, in 'Linked Open Data Workshop at The 17th International World Wide Web Conference'.

Assessing Text Characteristics of Electronic Discharge Summaries and their Implications for Patient Readability

Mehnaz Adnan¹, Jim Warren^{1,2}, Martin Orr^{2,3}

¹Department of Computer Science – Tamaki

²School of Population Health

The University of Auckland, New Zealand

³Waitemata District Health Board, Auckland, New Zealand

madn002@aucklanduni.ac.nz

Abstract

A Discharge Summary provides critical information to patients for managing their post-discharge care. This study analyzes the characteristics of a corpus of Electronic Discharge Summaries (EDSs) with respect to content and readability of its sections in terms of text length and grade level complexity, use of abbreviations and noun phrase complexity based on the Open Access Consumer Health Vocabulary. We find that the Advice to Patient section has acceptable readability but is brief, and does not tend to lengthen in proportion to the Clinical Management section of the EDS. Conversely, the Clinical Management section, while acceptable by traditional readability measures, has a higher density of abbreviations than Advice to Patient and considerable density of noun phrases that are unlikely to be understood by consumers. If patients are intended to be a primary audience of the EDS, then efforts should be made to improve readability for ordinary health consumers.

Keywords: electronic discharge summary, consumer vocabulary, consumer health informatics, abbreviations, readability.

1 Introduction

The healthcare system serving the 1.5M residents of the Auckland metropolitan area is largely electronic, both in the public hospital environment and in community based General Practice. Hospital discharge summaries are authored online by a health professional and transmitted to General Practice via HL7 messages. The patient is given a hard copy printout.

A Discharge Summary is usually created by a health professional for a number of audience including patients and their families to provide a snapshot of a patient's condition at the time of discharge (Barretto, Chu et al. 2006) and to provide a post-discharge framework of care for a patient (Walraven 1999). It usually includes a synopsis of care provided along with the advice of ongoing management of clinical condition, appropriate use of medications, relevant laboratory results and required follow-up.

Consumers play an important role in managing their own care, especially post-discharge (Maloney and Weiss 2008), hence the availability of easily understandable discharge information becomes critical. Engel et al (Engel KG, Heisler M et al. 2008 Jul) emphasize that the patients should understand "both" the care that they received and their discharge instructions.

While many patients have access to their discharge summaries, some studies (Heng, Tham et al. 2007; Clarke, Friedman et al.) have raised the comprehension issues of discharge instructions by patients with respect to ineffective care (Clarke, Friedman et al.), and lower compliance rates (Enguidanos and Rosen 1997). According to Makaryus et al (Makaryus and Friedman 2005), better understanding of diagnosis and treatment plans helps in enhancing patients' education and compliance, therefore reducing the likelihood of hospital readmission.

To disseminate the discharge summary information optimally, it is necessary that the written information provided is clear, free from errors and sources of confusion and, importantly, it must be easy to read and comprehend. While, a significant number of patients have low literary and/or health literacy levels (Zeng-Treitler, Goryachev et al. 2007), language is recognized as a factor affecting compliance rates of discharge instructions (Enguidanos and Rosen 1997). Clarke et al (Clarke, Friedman et al. Jan 2005) identifies the vocabulary and medical terminologies employed by electronic discharge summaries (EDSs) as key factors affecting

their comprehension. In addition, it has been observed that the use of abbreviations causes “severe shortcoming” in the clinical data of EDS and serves to confuse the communication between providers and patients (Walsh and Gurwitz 2008).

Keeping in view of the comprehension issues of discharge instructions by the patients and language employed by EDSs, we have embarked on a project to produce more readable EDS contents through interactive computer-based support, both at the authoring and in the reading of discharge summaries. As a first phase in this research, we are conducting analysis of current EDS content in terms of the text characteristics, associated readability, use of abbreviations and language familiarity. Herein we present findings from measurements on a corpus of EDS text.

2 Readability Evaluation of Health Information

Readability of a text, is usually expressed as grade level and refers to the ease with which it can be read (Zakaluk and Samuels 1988). However, standard measures of readability are insufficient to evaluate the difficulty of medical texts (Rosembat, Logan et al. 2006; Kim, Goryachev et al. 2007; Zeng-Treitler, Kim et al. 2007). Rosembat et al identified the ‘vocabulary’ as one of the factor that need to be considered in readability of medical text (Rosembat, Logan et al. 2006).

Many researchers have used common readability metrics and medical terminologies to examine the difficulty level of medical text for a lay reader. In 2006, Elhadad (Elhadad 2006) presented a method for health consumers to automatically predict difficult terms in medical literature. Also in 2006, Leory et al. (Leory, Eryilmaz et al. 2006) analysed and compared the text characteristics of easy and difficult WebMD documents, patient education material and patient blogs. In a follow-up study (Leory, Helmreich et al. 2008) Leory and colleagues analysed and compared the text characteristics of disease specific web contents, Medline and patients blogs. Furthermore, in 2007, Zeng et al (Zeng-Treitler, Kim et al. 2007) showed that Electronic Health Records (EHRs), consumer health materials, and scientific journal articles display many syntactic and semantic aspects that are not taken into account by existing readability measurements.

Previous research in health information readability has focused on consumer health materials, research articles and EHRs, but not assessment of readability issues in EDSs. In light of known patient comprehension difficulties (Engel, Heisler et al. 2008), text analysis of EDS contents may provide additional valuable information about consumer

readability issues. The objective of this study, therefore, is to assess the factors that affect the patient’s comprehension of EDS contents. This research is an important step in directing future efforts to identify and intervene on readability issues of EDSs to overcome patient comprehension deficits.

3 Materials

We collected 200 de-identified randomly selected hard copies of EDSs from the clinical data repository of a metropolitan District Health Board managing two public hospitals: North Shore Hospital and Waitakere Hospital (Auckland, New Zealand) with yearly presentations of around 43,000 and 24,000 patients, respectively. The sample data was collected from a total of 62,674 EDSs generated during the period of June 2007 to July 2008. We retrieved 50 EDSs each from Emergency, Medicine, Surgery and Older Adult Health Services departments as hardcopy printouts (as it proved most expedient to perform de-identification by literally cutting identifying information from the page!). The sample data, once transferred to the research cite, was then transformed back to electronic format for automatic analysis. For this purpose we scanned the EDSs using OCR (optical character recognition) software included with the Hewlett-Packard ScanJet 2200c scanner. The scanned EDSs were then checked manually by the first author and errors in the scanned output corrected to reconcile to the hardcopy printout (most errors were in the lists of Medications and laboratory investigation results, with narrative text observed to scan with near-perfect accuracy).

The EDSs contains sections including diagnoses, admission reason, clinical management, discharge medications, follow up, procedures, allergies and adverse reactions and relevant laboratory results, as well as an advice to patient section. All sections have a combination of complete sentences and bulleted text except for diagnoses and discharge medications, which have bulleted text only. For purposes of readability analysis, we consider the combined text of six key sections of the EDS documents, as per Table 1. These sections represent the most important information that the consumers might be expected to understand.

4 Methods

In this section we describe in detail the three measures we used to assess the text difficulty and our natural language processing techniques adapted for measuring the text characteristics of EDS contents. We first present our strategy of measuring readability scores. We then describe our process to extract abbreviations from text. Finally, we present the

method of identifying difficult words or phrases in the text.

4.1 Readability Score

The readability score for all the key sections in our corpus was measured with the use of the Flesch–Kincaid readability scale (grade-level range, 0 to 12) using Microsoft Word (Zakaluk and Samuels 1988). The intuition behind Flesch–Kincaid formula is that it is the most common metric for readability evaluation that represents the minimum school grade the reader should have completed to understand a document.

The formula is based on counts such as the number of syllables per word and the number of words per sentence. In addition, we also calculated total number of characters, words and sentences to measure the syntactic features of the text. We also calculated the average word length (i.e. number of characters per word) and the average sentence length (i.e., number of words per sentence). All bulleted texts were considered to be a sentence. Distribution of word count of advice to patient and clinical management section was also calculated.

4.2 Abbreviations

We used three methods to extract abbreviations from EDS text. First we extracted the list of abbreviations defined in SNOMED CT 2008 version. The SNOMED CT which contains 311,000 unique concepts and their synonyms is a comprehensive source for medical terms, including clinical abbreviations. The terminology file in SNOMED CT “sct_descriptions” contains information associated with clinical terms, such as the unique concept identifier and source vocabulary. To extract abbreviations from source vocabulary we used the method reported by Liu (Liu, Lussier et al. 2001) of using a space-delimited hyphen in terms as a marker of an abbreviation and its expansion. For example, the abbreviation “DM” is extracted from the SNOMED CT term DM – Diabetes mellitus.

Second, as we are processing DS data collected from WDHB, we also used a list of locally approved abbreviation (supplied by the District Health Board) and integrated them into Abbreviation Lexicon.

For abbreviation detection, we used GATE - General Architecture of Text Engineering (Cunningham, Maynard et al. July 2002). GATE is an open source well-established suite for developing natural language processing application. Some of the basic modules of GATE are Tokenizer, Gazetteer and JAPE (Cunningham, Maynard et al. 2000). The Tokenizer splits texts into simple tokens, the Gazetteer matches phrases to named entities in a given list, and distinguishes lowercases and

uppercases. The JAPE grammars can be used to code a pattern-matching algorithm.

The key sections (Diagnoses, Clinical Management, Medications, Follow Up, Advice to Patient and Allergies and Adverse Reactions) in each document in the corpus were processed by GATE to calculate number of abbreviations. Figure 1 illustrated the complete named entities extraction process in EDS text. The GATE operational model of abbreviation recognition is shown in Figure 1(a). The text splits into words by the GATE Tokenizer. We considered a word to be any token that does not contain punctuation symbols. For the use of abbreviation analysis, we used GATE Gazetteer. We built new Gazetteer list for our Abbreviation Lexicon, which consist of SNOMED CT and WDHB approved abbreviations. Finally, a JAPE grammar was written to extract all ‘unknown abbreviations’ in the text. The definition of ‘unknown abbreviation’ is an all-capital-letters word having length of one to seven characters and not falling into either of SNOMED CT or approved abbreviations.

For analyzing abbreviations in Clinical Management and Advice to Patient sections, a JAPE grammar was written to extract abbreviations in the Noun Phrases of these sections. These sections provide information about in-hospital and post-discharge care to the patient in free text form. Our hypothesis is that Noun Phrases comprise the most meaningful content in such text and have a higher percentage of abbreviations and difficult words, as compared to other parts of speech, making the entire content difficult to assimilate. Noun Phrase chunker in GATE is an implementation of the Ramshaw and Marcus transformational learning-based noun phrase chunker (Ramshaw and Marcus 1995). This module uses the set of rules and the lexicon to group Part of Speech tagged words, generated by GATE Part of Speech Tagger, into the noun phrases.

For the use of abbreviations in Noun Phrases, we extracted abbreviations used in the Noun Phrases of the Clinical Management and Advice to Patient sections. Figure 1(b) shows the GATE operational model of recognizing abbreviations in Noun Phrases. The text was grouped into words by the Noun Phrase chunker. To identify abbreviations in Noun Phrases we used the abbreviation Gazetteer list as described above. Finally a JAPE grammar is written to extract all components of noun phrase that recognized as abbreviations.

4.3 Open Access Consumer Health Vocabulary

Open Access Collaborative’s Consumer Health Vocabulary (Zeng and Tse 2006) (CHV) link

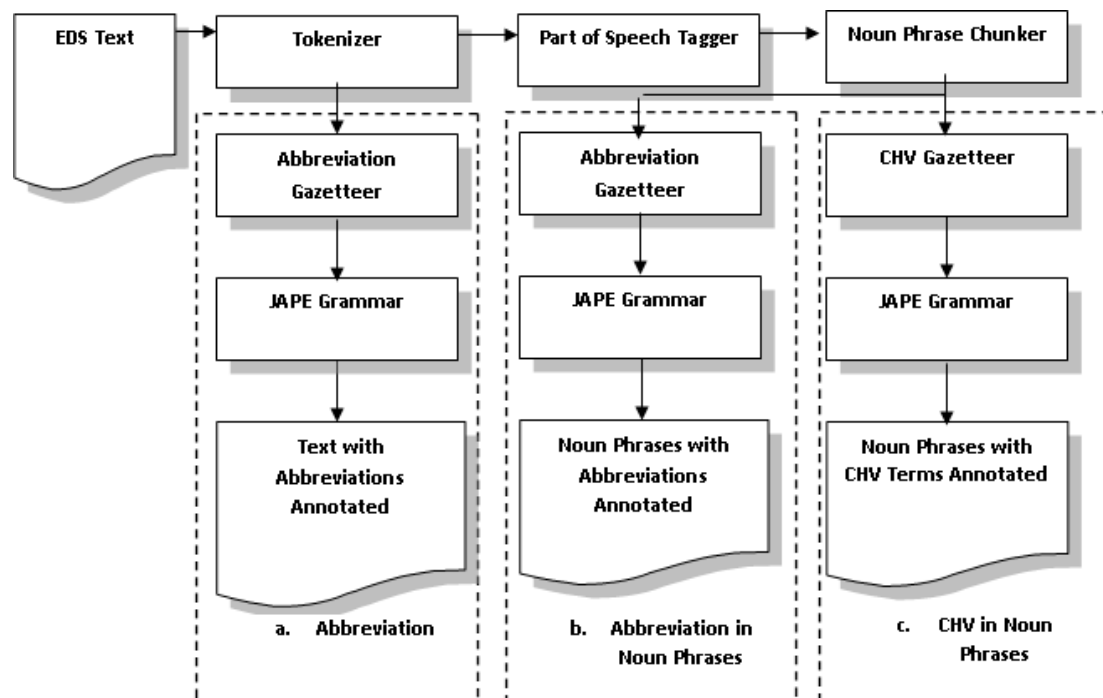


Figure 1. GATE Named Entities Extraction Process

technical terms or jargon used by health care professionals ("myocardial infarction") with consumer health specific words and phrases ("heart attack"). Each term in CHV has three associated familiarity scores: a *frequency-based term score* (calculated by a support-vector machine model based on term occurrence frequency in several health text corpora), a *context-based term score* (calculated based on term co-occurrence patterns in a health-specific query log data), and a *context-based concept score* (calculated on the basis of concept co-occurrences in medical literature and log data as well as semantic relations in medical vocabularies). The term scores reflect the string-level difficulty to estimate the likelihood the term will be recognized by an average consumer. The concept score estimates the concept-level difficulty for consumers. The scores range between 0 and 1, with a score of 0.8 to 1.0 representing "likely", 0.5 to 0.8 "somewhat likely" and below 0.5 "not likely" for a term to be familiar to a consumer (Keselman, Tse et al. 2007).

For term familiarity analysis, we extracted CHV terms used in the Noun Phrases of the Clinical Management and Advice to Patient sections. The GATE operational model of CHV term recognition is illustrated in Figure 1(c). The text was grouped into words by the Noun Phrase chunker. To identify CHV

terms in Noun Phrases we built a new Gazetteer list for CHV terms. Finally a JAPE grammar is written to extract all components of noun phrase that mapped to terms defined in the CHV.

To calculate the level of understanding of these free text sections by consumers, we use CHV term and concept familiarity scores to gauge the semantic complexity of the contents. Some extracted CHV terms did not have scores assigned (indicated as a -1); while analyzing scores of CHV terms in noun phrases, these missing score terms were excluded.

5 Results

5.1 Readability Scores

The text characteristics of the EDS sections are reported Table 1. On the level of text unit length, the total word count differs radically; the Clinical Management section has 8 times the word count of the Advice to Patient section. The Advice to Patient section uses longer sentences as compared to other sections. The shorter sentence length of the Diagnoses and Allergies and Adverse Reactions texts are largely due to incomplete sentences and use of abbreviations (see section 4.2).

On the readability grade level, we found that Flesch-Kincaid Grade Level scores vary in different

sections. Advice to Patient requires a grade level of above 6th grade while the Diagnoses and Follow-up requires a grade level that is above 12th grade.

The word count of the Advice to Patient section with respect to Clinical Management and the whole EDS is shown as scatter plots in Figures 2 and 3.

Linear regression was used to model length of Advice to Patient; although somewhat influenced by outlier data, the regression (shown as the square of the correlation coefficient - R^2) with a value of 0.0084 in Figure 2 and 0.034 in Figure 3 gives a clear indication of only a very weak association.

Section of Discharge Summary	Total Words	Words per Sentence	Characters per Word	Flesch-Kincaid Grade Level
Diagnoses	4079	4.3	5.9	12.3
Clinical Management	40090	8.9	4.9	8.9
Medications	9953	10.0	4.8	9.5
Follow Up	2148	10.4	6.4	12.3
Advice to Patient	4719	12.3	4.4	6.5
Allergies and Adverse Reactions	345	1.7	5.7	7.8

Table 1. Readability statistics of the EDS Corpus

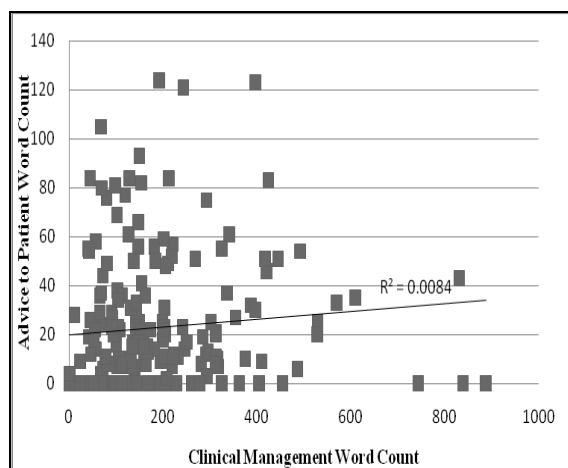


Figure 2. Scatter plot of word count of Advice to Patient and Clinical Management

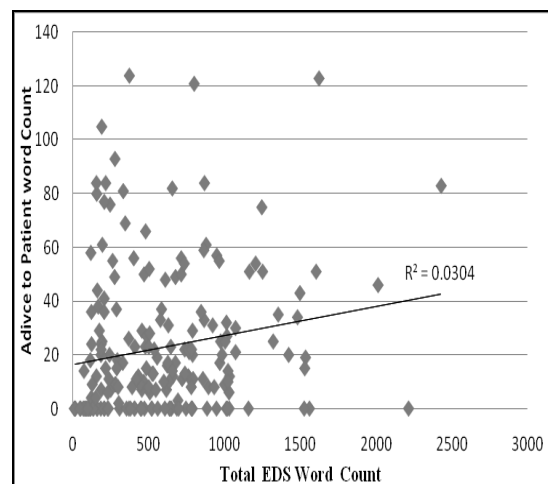


Figure 3. Scatter plot of word count of Advice to Patient and Total Word Count

5.2 Abbreviations

The percentage of abbreviations used in EDS sections are reported in Figure 4. The bar-graph shows that Allergies and Adverse Reactions have the highest percentage of abbreviations followed by Medications

and Diagnoses sections. In the Clinical Management almost 8% of words were abbreviations. In the Advice to Patient almost 4% of words were abbreviations (almost 50% lower).

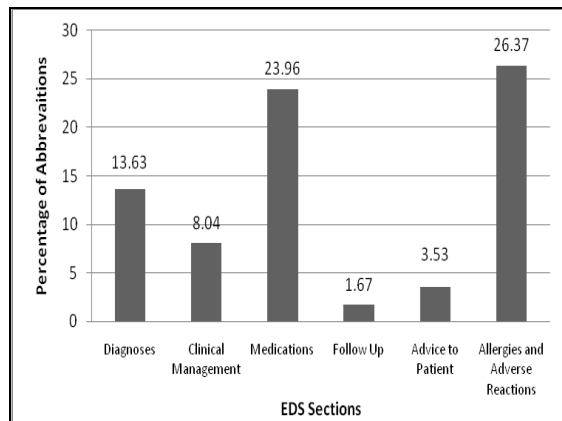


Figure 4. Abbreviations in EDS Sections

The distribution of number of abbreviations in Noun Phrases of the Advice to Patient section and Clinical Management is shown in as scatter plots in Figures 5 and 6 respectively. The R^2 for abbreviation count in Clinical Management section (Figure 5) has a value of 0.6681, which shows the strong association between the variables. While the R^2 value of 0.3607 in Figure 6 indicating a weak association of abbreviations count in Noun Phrases of Advice to Patient text.

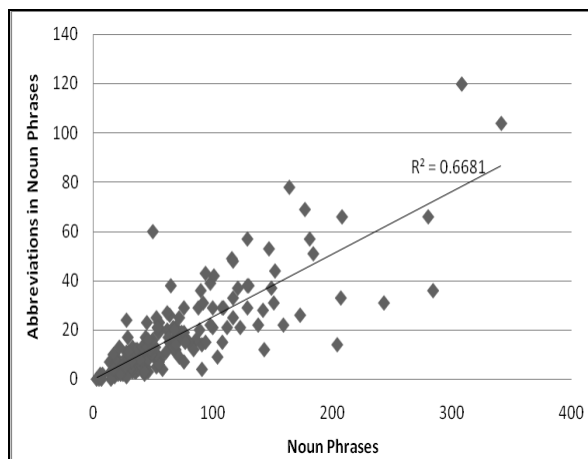


Figure 5. Distribution of Number of Abbreviations in Noun Phrases of Clinical Management Section

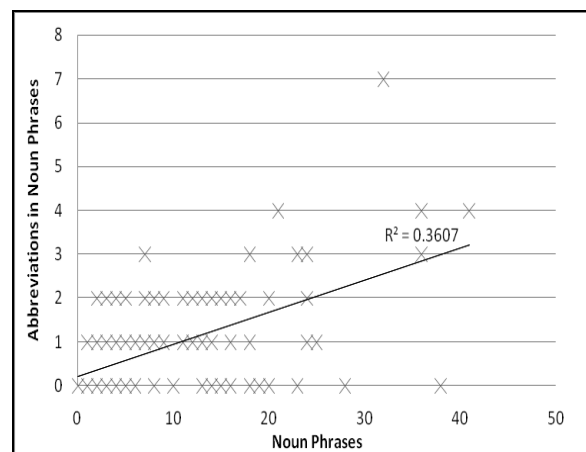


Figure 6. Distribution of Number of Abbreviations in Noun Phrases of Advice to Patient Section

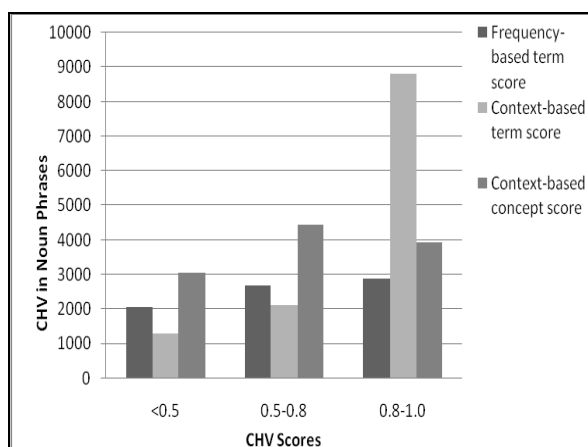


Figure 7. CHV Scores in Clinical Management

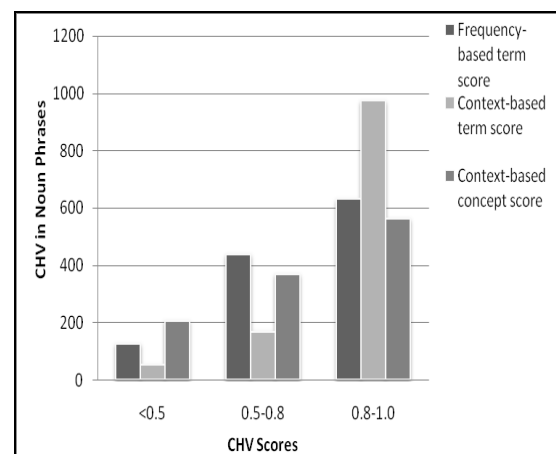


Figure 8. CHV Scores in Advice to Patient

5.3 Consumer Health Vocabulary Scores

Three CHV scores in Noun Phrases of Clinical Management and Advice to Patient sections are reported in Figure 7 and 8 respectively. Greater number of Noun Phrases in Clinical Management has lower scores than Advice to Patient in all three CHV measures. The percentage of CHV terms in Clinical Management having frequency-based, context-based and concept-based terms scores < 0.5 are 27%, 10%, and 27%, respectively. In contrast, the Advice to Patient percentage of CHV terms having frequency-based, context-based and concept-based terms scores < 0.5 are 10%, 4%, and 18%, respectively.

6 Discussion:

Although discharge summaries are primarily written for health professionals, health consumers access their contents for self-care; but comprehension of discharge instructions and readability for the consumer audience has been recognized as problematic. Our analyses indicate that the Advice to Patient section of the EDS, while written at an approachable grade level and low abbreviation density, is a very brief component of the total document. More concerning, the length of the section is largely uncorrelated to the length of the Clinical Management section or the total length of the EDS. The results suggest that Advice to Patient does not provide complete information about the condition, treatment plan and medication side effects to patients but, rather, readable 'stock phrases.' For example, in 31 of the EDSs the only advice for the patient was a single sentence advising them to see their GP for any medical concern. Therefore, for optimal post-discharge care, patients will in fact have to look outside of the Advice to Patient section in the present EDS documents to find all the information they might need.

The Clinical Management section, which provides a summary of diagnosis and treatment plan, unsurprisingly, exhibits characteristics making its content difficult to understand by a lay person, which aligns with similar findings based on characteristics of EHR reports (Zeng-Treitler, Kim et al. 2007). The Clinical Management section has more than twice the abbreviation density of Advice to Patients and a much higher frequency of terms that are unlikely to be understood by the consumer based on the Open Access Consumer Health Vocabulary (Zeng and Tse 2006).

It is assumed that after discharge from hospital patients will be transferred to the care of a general practitioner. Thus, patient understanding of the entire Discharge Summary contents is probably not the most important factor in the discharge care plan. However, to improve the post-discharge self care it is an integral component.

Our results indicate a need to improve the readability of EDS documents for patients. In the first instance, there should be improved emphasis and training of hospital staff with respect to the importance of addressing the consumer audience in the EDS. However, we also see an opportunity for computer-based support to play an important role in improving readability for the health consumer.

A first opportunity comes with electronic decision support at the time of EDS authoring. The authoring environment could provide interactive feedback on CHV scores of terms used, unknown or low readability abbreviations, and correlation of length of Advice to Patient to other key sections of the document. Secondly, we live in an increasingly electronic consumer environment where many health consumers (e.g., as is the case presently for Kaiser Permanente patients) access a wide range of provider information and services online (Zhou, Garrido et al. 2007). If the EDS is presented online, then there is potential to provide a degree of support for less readable terms through hypertext, both with links to consumer health resources and through the ability to direct a question to the primary care physician, the EDS author or another source of support.

A key limitation of the present analysis is that it is purely a computational one utilizing statistical and automated term matching characteristics of the data. A further analysis of a sample from the present corpus is under way involving the professional judgment of physicians and medical records staff as feedback on the suitability and completeness of the EDS sections. Consulting the consumers themselves is a further important step for this research program. Another limitation is the use of a single District Health Board as the source for our sample, in that the specific policies and software infrastructure of this jurisdiction limit the ability to generalize the findings.

7 Conclusion

This study analyzed the characteristics of a corpus of Electronic Discharge Summaries (EDSs) with an eye to the readability for health consumers. We examined content and readability of its sections in terms of text length and grade level complexity, use of abbreviations and noun phrase complexity based on the Open Access Consumer Health Vocabulary. We find that the Advice to Patient section has acceptable readability but is brief, and does not tend to lengthen in proportion to the Clinical Management section of the EDS, suggesting that consumers will need to look outside of the Advice to Patient section for information they need.

The Clinical Management section, while acceptable by traditional readability measures such as Flesch-Kincaid Grade Level, has a higher density of abbreviations than Advice to Patient and considerable density of noun phrases that are unlikely to be understood by consumers. If patients are intended to be a primary audience of the EDS, then efforts should be made to improve readability for ordinary health consumers. Such efforts should include improved training of staff to focus on the needs of patients as an audience, but should also be leveraging the potential of software to aid readability both at the EDS authoring and reading stages.

8 Acknowledgments

This work was supported by a Higher Education Commission, Pakistan scholarship. The research protocol was approved by the University of Auckland Human Participants Ethics Committee under protocol number

2008/221 and by the Waitemata District Health Board Knowledge Centre.

9 References

- Barretto, S., S. Chu, et al. (2006). National Discharge Summary: Data Content Specifications Version 1.0. http://www.nehta.gov.au/component/docman/doc_download/175-national-discharge-summary-data-content-specifications-v10. Accessed 13 Aug 2009.
- Clarke, C., S. Friedman, et al. (Jan 2005). Emergency department discharge instructions comprehension and compliance study. *Canadian Journal of Emergency Medicine* **7**(1): 7.
- Cunningham, H., D. Maynard, et al. (July 2002). GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. *40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, Philadelphia.
- Cunningham, H., D. Maynard, et al. (2000). JAPE: a Java Annotation Patterns Engine (Second Edition). Technical report CS--00--10, University of Sheffield, Department of Computer Science.
- Elhadad, N. (2006). Comprehending technical texts: predicting and defining unfamiliar terms. *American Medical Informatics Association (AMIA) Annual Symposium Proceedings*: 239-43.
- Engel KG, Heisler M, et al. (2008 Jul). Patient Comprehension of Emergency Department Care and Instructions: Are Patients Aware of When They Do Not Understand? *Annals of Emergency Medicine* (10).
- Enguidanos, E. R. and P. Rosen (1997). Language as a factor affecting follow-up compliance from the emergency department. *Journal of Emergency Medicine* **15**(1): 9-12.
- Heng, K. W. J., K. Y. Tham, et al. (2007). Recall of discharge advice given to patients with minor head injury presenting to a Singapore emergency department. *Singapore Medical Journal* **48**(12): 1107-10.
- Keselman, A., T. Tse, et al. (2007). Assessing consumer health vocabulary familiarity: an exploratory study. *Journal of Medical Internet Research* **9**(1): e5.
- Kim, H., S. Goryachev, et al. (2007). Beyond surface characteristics: a new health text-specific readability measurement. *American Medical Informatics Association Annual Symposium Proceedings* 418-22.
- Leroy, G., E. Eryilmaz, et al. (2006). Health information text characteristics. *American Medical Informatics Association Annual Symposium*, Washington DC.
- Leroy, G., S. Helmreich, et al. (2008). Evaluating online health information: beyond readability formulas. *American Medical Informatics Association Annual Symposium Proceedings* 394-8.
- Liu, H., Y. A. Lussier, et al. (2001). A study of abbreviations in the UMLS. *American Medical Informatics Association Annual Symposium Proceeding*. 393-7.
- Makaryus, A. N. and E. A. Friedman (2005). Patients' understanding of their treatment plans and diagnosis at discharge. *Mayo Clinic Proceedings* **80**(8): 991-4.
- Maloney, L. R. and M. E. Weiss (2008). Patients' perceptions of hospital discharge informational content. *Clinical Nursing Research* **17**(3): 200-19.
- Ramshaw, L. and M. Marcus (1995). Text Chunking using Transformation-Based Learning. *Third Workshop on Very Large Corpora*.
- Rosemblat, G., R. Logan, et al. (2006). Text Features and Readability: Expert Evaluation of Consumer Health Text. *Mednet 2006: 11th World Congress on Internet in Medicine the Society for Internet in Medicine*, Toronto, Canada.
- Walraven, C. (1999). What Is Necessary for High-Quality Discharge Summaries? *American Journal of Medical Quality* **14**(4): 10.
- Walsh, K. E. and J. H. Gurwitz (2008). Medical abbreviations: writing little and communicating less.[comment]. *Archives of Disease in Childhood* **93**(10): 816-7.
- Zakaluk, B. and S. Samuels (1988). Readability: Its Past, Present, and Future. *International Reading Association*.
- Zeng-Treitler, Q., S. Goryachev, et al. (2007). Making texts in electronic health records comprehensible to consumers: a prototype translator. *American Medical Informatics Association Annual Symposium Proceeding* 846-50.
- Zeng-Treitler, Q., H. Kim, et al. (2007). Text characteristics of clinical reports and their implications for the readability of personal health records. *Studies in Health Technology & Informatics* **129**(Pt 2): 1117-21.
- Zeng, Q. and T. Tse (2006). Exploring and developing consumer health vocabularies. *Journal of American Medical Informatics Association*.
- Zhou, Y. Y., T. Garrido, et al. (2007). Patient access to an electronic health record with secure messaging: impact on primary care utilization. *American Journal of Managed Care* **13**(7): 418-24.

Author Index

- Adnan, Mehnaz, 77
Amiel, H., 39
Ansell, Peter, 69

BaniMustafa, A., 39
Bertoli, M., 39
Bjering, Heidi, 29

Caelli, William, 7
Connor, Jason, 45

Dooris, Mark, 45

Foo, Jin Hong, 7

Gallagher, Marcus, 45
Ganascia, J. G., 39

Hansen, David, iii
Hogan, James, 69
Hu, Hongxiang, 17
Huda, S., 39

Kisely, Steve, 3

Lévy, P., 39
Lee, Min Hui, 7
Li, Min, 53
Li, Weihao, 7
Liu, Vicky, 7
Luo, Wei, 45

Ma, L., 39
Maeder, Anthony, iii
May, Lauren, 7

McGregor, Carolyn, 29
Molla, Diego, 61
Mortimer, Lachlan, 45
Muecke, N., 39

Ng, Zi Hao, 7

O’Kane, Di, 45
Ofoghi, B., 39
Orr, Martin, 77

Patrick, Jon, 53
Philippe, C., 39

Roberts, Col, 45
Roe, Paul, 69

Saleem, M., 39
Smith, Jason, 7
Stiller, Anthony D., 23
Stranieri, A., 39
Sukhorukova, N., 39

Tutos, Andreea, 61

Ugon, A., 39
Ugon, J., 39

Vamplew, P., 39

Warren, Jim, 77
Wiles, Janet, 45

Yan, Jun, 17
Yu, Ping, 17

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 84 - Artificial Intelligence and Data Mining 2007

Edited by Kok-Leong Ong, Deakin University, Australia, Wenyuan Li, University of Texas at Dallas, USA and Junbin Gao, Charles Sturt University, Australia. December, 2007. 978-1-920682-65-1.

Contains the proceedings of the 2nd International Workshop on Integrating AI and Data Mining (AIDM 2007), Gold Coast, Australia. December 2007.

Volume 85 - Advances in Ontologies 2007

Edited by Thomas Meyer, Meraka Institute, South Africa and Abhaya Nayak, Macquarie University, Australia. December, 2007. 978-1-920682-66-8.

Contains the proceedings of the 3rd Australasian Ontology Workshop (AOW 2007), Gold Coast, Queensland, Australia.

Volume 86 - Safety Critical Systems and Software 2007

Edited by Tony Cant, Defence Science and Technology Organisation, Australia. December, 2007. 978-1-920682-67-5.

Contains the proceedings of the 12th Australian Conference on Safety Critical Systems and Software, August 2007, Adelaide, Australia.

Volume 87 - Data Mining and Analytics 2008

Edited by John F. Roddick, Jiuyong Li, Peter Christen and Paul Kennedy. November, 2008. 978-1-920682-68-2.

Contains the proceedings of the 7th Australasian Data Mining Conference (AusDM 2008), Adelaide, Australia. December 2008.

Volume 88 - Koli Calling 2007

Edited by Raymond Lister University of Technology, Sydney and Simon University of Newcastle. November, 2007. 978-1-920682-69-9.

Contains the proceedings of the 7th Baltic Sea Conference on Computing Education Research.

Volume 89 - Australian Video

Edited by Heng Tao Shen and Michael Frater. October, 2008. 978-1-920682-70-5.

Contains the proceedings of the 1st Australian Video Conference.

Volume 90 - Advances in Ontologies

Edited by Thomas Meyer, Meraka Institute, South Africa and Mehmet Orgun, Macquarie University, Australia. September, 2008. 978-1-920682-71-2.

Contains the proceedings of the Knowledge Representation Ontology Workshop (KROW 2008), Sydney, September 2008.

Volume 91 - Computer Science 2009

Edited by Bernard Mans Macquarie University. January, 2009. 978-1-920682-72-9.

Contains the proceedings of the Thirty-Second Australasian Computer Science Conference (ACSC2009), Wellington, New Zealand, January 2009.

Volume 92 - Database Technologies 2009

Edited by Xuemin Lin, University of New South Wales and Athman Bouguettaya, CSIRO. January, 2009. 978-1-920682-73-6.

Contains the proceedings of the Twentieth Australasian Database Conference (ADC2009), Wellington, New Zealand, January 2009.

Volume 93 - User Interfaces 2009

Edited by Paul Calder Flinders University and Gerald Weber University of Auckland. January, 2009. 978-1-920682-74-3.

Contains the proceedings of the Tenth Australasian User Interface Conference (AUIC2009), Wellington, New Zealand, January 2009.

Volume 94 - Theory of Computing 2009

Edited by Prabhhu Manyem, University of Ballarat and Rod Downey, Victoria University of Wellington. January, 2009. 978-1-920682-75-0.

Contains the proceedings of the Fifteenth Computing: The Australasian Theory Symposium (CATS2009), Wellington, New Zealand, January 2009.

Volume 95 - Computing Education 2009

Edited by Margaret Hamilton, RMIT University and Tony Clear, Auckland University of Technology. January, 2009. 978-1-920682-76-7.

Contains the proceedings of the Eleventh Australasian Computing Education Conference (ACE2009), Wellington, New Zealand, January 2009.

Volume 96 - Conceptual Modelling 2009

Edited by Markus Kirchberg, Institute for Infocomm Research, A*STAR, Singapore and Sebastian Link, Victoria University of Wellington, New Zealand. January, 2009. 978-1-920682-77-4.

Contains the proceedings of the Fifth Asia-Pacific Conference on Conceptual Modelling (APCCM2008), Wollongong, NSW, Australia, January 2008.

Volume 97 - Health Data and Knowledge Management 2009

Edited by James R. Warren, University of Auckland. January, 2009. 978-1-920682-78-1.

Contains the proceedings of the Third Australasian Workshop on Health Data and Knowledge Management (HDKM 2009), Wellington, New Zealand, January 2009.

Volume 98 - Information Security 2009

Edited by Ljiljana Brankovic, University of Newcastle and Willy Susilo, University of Wollongong. January, 2009. 978-1-920682-79-8.

Contains the proceedings of the Australasian Information Security Conference (AISC 2009), Wellington, New Zealand, January 2009.

Volume 99 - Grid Computing and e-Research 2009

Edited by Paul Roe and Wayne Kelly, QUT. January, 2009. 978-1-920682-80-4.

Contains the proceedings of the Australasian Workshop on Grid Computing and e-Research (AusGrid 2009), Wellington, New Zealand, January 2009.

Volume 100 - Safety Critical Systems and Software 2007

Edited by Tony Cant, Defence Science and Technology Organisation, Australia. December, 2008. 978-1-920682-81-1.

Contains the proceedings of the 13th Australian Conference on Safety Critical Systems and Software, Canberra Australia.

Volume 101 - Data Mining and Analytics 2009

Edited by Paul J. Kennedy, University of Technology, Sydney, Kok-Leong Ong, Deakin University and Peter Christen, The Australian National University. November, 2009. 978-1-920682-82-8.

Contains the proceedings of the 8th Australasian Data Mining Conference (AusDM 2009), Melbourne Australia.