Australasia's big data in biomedicine & healthcare conference

3 - 4 APRIL 2014 MELBOURNE

# BIG DATA.

## and healthcare analytics

**BIG INSIGHTS: HARNESSING THE POWER OF HEALTH DATA**

BIG QUESTIONS: BIG INSIGHTS

HEALTH OUTCOMES

DATA VISUALISATION

WORKFORCE

DATA LINKAGE & ANALYTICS

GENOMICS

PRIVACY

proceedings

#bd14

hisa

hisa

# Be part of a growing community of e-health experts: Join HISA today

Influence the agenda

Build professional networks

Be part of a national community

Career development & certification

Access to Australia's largest network of e-health experts and leaders

Exclusive member only benefits & resources

## SPONSORS

GOLD SPONSOR

Bupa

STATE HOST PARTNER

State Government Victoria AUSTRALIA

MELBOURNE VICTORIA AUSTRALIA

CONFERENCE PARTNER

CSIRO

## SUPPORTING ORGANISATIONS

ADMA

AHRDMA

Australian Library and Information Association

Australian National Consultative Committee on Health

DIGITAL HEALTH

NCOPP

THE NATIONAL INSTITUTE FOR HEALTH INNOVATION THE UNIVERSITY OF AUCKLAND

Australian Government National Health and Medical Research Council

S·A·R·R·A·H

VeHN

# CHAIR WELCOME

## conference chair

**Susan Walker**
Chair
Big Data 2014
Conference

I am delighted to welcome you to HISA's Big Data 2014 Conference! This year's conference builds upon the success of our inaugural Big Data meeting and Data Governance meetings prior to 2013.

Last year, Prof Fiona Stanley spoke about the fact that whilst yes, we have potential access to significant volumes of anonymous data she challenged us to consider: what are we going to do with it, and how can we break down the barriers to sharing this? Feedback from attendees echoed this and there was much positivity about our capabilities and the progress to date in navigating big data issues….but a challenge for us to consider…where does that leave us?

Feeling duly rallied, the strategic advisory committee reflected that well, if big data is the answer…what are the BIG QUESTIONS? We wanted to explore what insights could be gained from the use of big data and the extent to which we can create tangible gains for the health of Australians.

To this end, we have organised an exciting program of speakers from researchers to those responsible for delivery of care (both pointy ends), and I hope you see the call for action reflected in the themes of our conference.

A huge thanks to the strategic advisory committee for their generous donation of time and intellect and for the HISA team for their constant support.

Enjoy!

## STRATEGIC ADVISORY COMMITTEE

**Susan Walker**
Chair, Big Data 2014 Conference,
General Manager, Australian Centre for Health Innovation

**A/Prof Karin Verspoor**
Chair, Big Data 2014 Scientific Program,
Associate Professor, Department of Computing and Information Systems, University of Melbourne

**Dr Paul Cooper**
Chair, Big Data 2014 Industry/Clinical Case Study Program,
Health Industry Director, SMS Management and Technology

**Jon Buttery**
Team Leader, Modelling, Department of Health, Vic

**Nigel Chartres**
Advisor and Project Manager, HISA

**Madalene Crow**
Project Officer, St Vincent's Public Hospital

**Shane Downey**
Manager - Data Services, Mater Health Services

**Dr Sankalp Khanna**
Research Scientist,
The Australian E-Health Research Centre, CSIRO

**Dr Jia-Yee Lee**
Office of the Health Innovation and Reform Council

**Sanja Lujic**
Lecturer in Biostatistics,
Centre for Health Research,
University of Western Sydney

**Dr Louise Schaper**
Chief Executive Officer, HISA

**Dr Clive Morris**
Head, Strategic Policy Group,
The National Health and Medical Research Council

**Peter Williams**
Advisor, E-Health Policy and Engagement,
Department of Health, Vic

## scientific program chair

**A/Prof Karin Verspoor**
Chair
Big Data 2014
Scientific Program

In the year since the first HISA Big Data conference, the notion of "big data" seems to have exploded into the forefront of the global health informatics scene. Even this year's theme of "Big Questions, Big Insights" is echoed in the recent announcement from the US National Institutes of Health of a new program in "Big Data to Knowledge" (BD2K). Hospitals, insurers, biomedical researchers and even patients themselves are generating increasingly large quantities of data, thanks to improved health information technical infrastructure, advances in technologies such as high-throughput sequencing, and on-line communication. All of this data in turn demands new strategies for managing, organising, analysing, and interpreting it.

In this year's scientific program, we have an exciting representation of research that spans a broad spectrum of biomedical data types and makes technical contributions that address big data issues from storage and access to algorithms for large-scale processing, visualisation, and health outcome prediction. There are substantial parallels in the scientific program with the themes of the industry/clinical program, which suggests that the scientific community is paying attention to real-world health information requirements. I hope that the conference will provide a rewarding opportunity to develop further collaborations and deepen understanding of the challenges and opportunities of big biomedical data in the community as a whole.

# PROGRAM DAY 1 – THURSDAY 3 APRIL

| | | |
|---|---|---|
| 7.30am | Registration desk opens | |
| 9.00am | **Conference Official Opening**<br>Introduction and welcome<br>Susan Walker Chair, Big Data 2014 Conference | State 3<br>Chair: Susan Walker |
| 9.10am | Welcome to country<br>Ron Jones Wurundjeri Tribe Elder | |
| | **Industry insights** | |
| 9.15am | Big data: Big opportunities, big thinking. Insights from Google<br>Angelo Joseph Head of Sales Engineering, Google Enterprise | |
| 9.45am | The internet of everything: Building health and wellness into the fabric of the community<br>Dr Brendan Lovelock Health Practice Lead, Cisco | |
| 10.15am | Discussion | |
| 10.30am | Morning tea and exhibition | Lake |
| | **Workforce** | |
| 11.00am | Putting together a bioinformatics team: 2014 compared with 1997<br>Prof Terry Speed Head, Division of Bioinformatics, Walter and Eliza Hall Institute of Medical Research | State 3<br>Chair: Dr Louise Schaper |
| 11.20am | Engaging doctors with (big) data – why is it hard?<br>A/Prof Christine Jorm Associate Dean Professionalism, Sydney Medical School | |
| 11.40am | Workforce panel facilitated by Dr Louise Schaper CEO, HISA<br>Julie Brophy Manager Health Information Workforce Strategy, Health Workforce Branch, Department of Health, Vic; A/Prof Christine Jorm Associate Dean Professionalism, Sydney Medical School; Prof Terry Speed Head, Division of Bioinformatics, Walter and Eliza Hall Institute of Medical Research | |
| 12.45pm | Lunch and exhibition | Lake |
| 1.45 - 3.15pm | CONCURRENT 1<br>Scientific Stream 1<br>**Big data analytics** | State 3<br>Chair: Prof Svetha Venkatesh |
| 1.45pm | Text mining for lung cancer cases over large patient admission data<br>Dr Lawrence Cavedon RMIT University | |
| 2.03pm | Causality driven data integration for adverse drug reaction discovery<br>Dr Chen Wang CSIRO | |
| 2.21pm | Big insights into patient flow<br>Dr Sarah Dods CSIRO and Dr Sankalp Khanna The Australian E-Health Research Centre, CSIRO | |
| 2.39pm | Enhancing diagnostics for invasive Aspergillosis using machine learning<br>Simone Romano University of Melbourne | |
| 2.57pm | Modeling of time series health data using Dynamic Bayesian Networks: An application to predictions of patient outcomes after multiple surgeries<br>Xiongcai (Peter) Cai Centre of Health Informatics, The University of New South Wales | |
| 1.45 - 3.15pm | CONCURRENT 2<br>Industry/Clinical Case Study Stream 1<br>**From big data to big insights** | State 1&2<br>Chair: Dr Paul Cooper |
| 1.45pm | Benefits you can realise immediately from big data, with lessons learned from the corporate and healthcare sectors<br>Greg Taylor The NTF Group | |
| 2.03pm | Big data from small sites: Creating a primary care data warehouse<br>Jason Ferriggi Inner East Melbourne Medicare Local | |
| 2.21pm | Driving patient benefits in the areas of informatics, genomics, and biological system modelling through big health<br>Lindsay Kiley InterSystems | |
| 2.39pm | Health: Big data, big opportunity<br>David Cross Bupa Health Dialog | |
| 2.57pm | The impact of big data on healthcare delivery<br>Paul Colmer CSC | |
| 3.15pm | Afternoon tea and exhibition | Lake |
| | **Privacy** | |
| 3.45pm | Debate Proposition<br>To realise and harvest the potential for big data to improve healthcare delivery and outcomes we need open data<br><br>Facilitator<br>Alison Choy Flannigan Partner, Health, Aged Care and Life Sciences, Holman Webb Lawyers<br><br>Debaters<br>Dr Tim Churches SURE Epidemiologist and Technical Advisor, Sax Institute; Dr Peter Croll CEO, Peter Croll and Associates; Jessica Dean President, Australian Medical Students' Association; Emma Hossack President, International Association of Privacy Professionals; A/Prof Trish Williams School of Computer and Security Science, Edith Cowan University; Prof Ingrid Winship Executive Director of Research, Melbourne Health | State 3<br>Chair: Jon Buttery |
| 5.00pm | Launch of HISA Privacy Guideline | |
| 5.10pm | Day one concludes | |
| 5.10-6.10pm | Networking Reception | Lake |

# PROGRAM DAY 2 – FRIDAY 4 APRIL

| | | |
|---|---|---|
| 7.30am | Registration desk opens | |
| 9.00am | Payer perspective (Insurer): Big data analytics to understand and influence human behaviour<br>Dr Bern Shen Chief Medical Officer, HealthCrowd (USA) | State 3<br>Chair: Dr Paul Cooper |
| 9.30am | Transforming information capture and decision making<br>Dr David Hansen CEO, The Australian E-Health Research Centre, CSIRO | |
| 10.00am | Health Hack 101: An introduction to this brave new world<br>Dr Maia Sauren ThoughtWorks Australia | |
| 10.10am | Discussion | |
| 10.25am | Morning tea and exhibition | Lake |
| 10.55 - 12.25pm | CONCURRENT 3<br>Scientific Stream 2<br>**Clinical genomics** | State 3<br>Chair: A/Prof Karin Verspoor |
| 10.55am | Identification of novel therapeutics for complex diseases from genome-wide association data<br>Mani Grover Deakin University | |
| 11.13am | GPU enables search for 2-way and 3-way interactions in GWAS<br>Dr Adam Kowalczyk University of Melbourne | |
| 11.31am | Personalised cloud-computed genomics at health-system-relevant scale<br>Dr Denis Bauer CSIRO | |
| 11.49am | Resolving ambiguity in genome assembly using high performance computing<br>Mahtab Mirmomeni IBM Research Australia | |
| 12.07pm | Implementing a clinical genomics infrastructure to sequence 18,000 human genomes per year<br>Dr Liviu Constantinescu Garvan Institute of Medical Research | |
| 10.55 - 12.25pm | CONCURRENT 4<br>Industry/Clinical Case Study Stream 2<br>**Harnessing the big data** | State 1&2<br>Chair: Dr Sankalp Khanna |
| 10.55am | Bedside data capture - combining patient management and research<br>Kate Birch University of Melbourne | |
| 11.13am | Clinical data hub for patient cohort selection<br>Dr Suman Sedai IBM Research Australia | |
| 11.31am | Clinical language processing for big data - trials and tribulations<br>Prof Jon Patrick Health Language Analytics | |
| 11.49am | Data driven rapid insights into population health status and needs<br>Dr Peter Del Fante Healthfirst Network | |
| 12.07pm | Data linkage and visualisation: Improving data quality for audit and research<br>Dr Leon Heffer BioGrid Australia | |
| 12.25pm | Lunch and exhibition | Lake |
| 1.25-2.55pm | CONCURRENT 5<br>Scientific Stream 3<br>**Knowledge from data** | State 3<br>Chair: Prof James Bailey |
| 1.25pm | Spatio-temporal visualisation of disease incidence and respective intervention strategies<br>Dr Priscilla Rogers IBM Research Australia | |
| 1.43pm | HealthMap: A visual platform for patient suicide risk review<br>Prof Svetha Venkatesh Deakin University | |
| 2.01pm | Health informatics visualisation engine: HIVE<br>Norm Good CSIRO | |
| 2.19pm | Classification of data and activities in self-quantification systems<br>Manal Almalki University of Melbourne | |
| 2.37pm | Parent perspectives on the secondary use of birth cohort data<br>Dr Kiran Pohar Manhas University of Calgary (CANADA) | |
| 1.25-2.55pm | CONCURRENT 6<br>Workshop<br>**The revolution in little data** | State 1&2<br>Chair: Dr Heather Leslie |
| | What is 'little data' – an introduction to clinical data models and archetypes<br><br>International approaches to standardisation and collaboration around data models, including CIMI and openEHR<br><br>Clinicians and non-technical domain experts lead this approach<br><br>Getting little data 'right' supports better quality big data<br><br>Dr Heather Leslie and Dr Hugh Leslie Ocean Informatics | |
| 2.55pm | Afternoon tea and exhibition | Lake |
| | **Implementation insights** | |
| 3.25pm | Executive perspective<br>A/Prof Andrew Way CEO, Alfred Health | State 3<br>Chair:Susan Walker |
| 3.55pm | Research perspective<br>Dr Mirana Ramialison Group Leader in Systems Developmental Biology, Australian Regenerative Medicine Institute | |
| 4.25pm | Thank you and closing remarks<br>Susan Walker Chair, Big Data 2014 Conference | |
| 4.40pm | Conference concludes | |

# KEYNOTE SPEAKERS


**Dr David Hansen**

Chief Executive Officer
The Australian E-Health
Research Centre, CSIRO

## TRANSFORMING DATA

Friday 4 April 9.30am

*Transforming information capture and decision making*

David Hansen is CEO of the Australian E-Health Research Centre, a joint venture between CSIRO and the Queensland Government. David leads a research portfolio developing information and communication technologies for the healthcare system. These include projects for resource planning, biomedical imaging, mobile and telehealth and technologies that will underpin the e-health architecture in Australia. Prior to joining CSIRO, David wo`rked for LION bioscience Ltd in the UK, developing genomic data and tool integration software that was used to publish the first human genome and is now used at over 200 pharmaceutical and biotechnology companies and research institutes worldwide.

**A/Prof Christine Jorm**

Associate Dean (Professionalism)
Sydney Medical School

**Understanding how doctors think and feel - including what makes them unreceptive to data - means new organisational experiences can be designed to change behaviour and thus improve patient care.**

## WORKFORCE

Thursday 3 April 11.20am

*Engaging doctors with (big) data – why is it hard?*

Christine is an anaesthetist with doctorates in neuropharmacology and sociology. Her book 'Reconstructing Medical Practice - Engagement, Professionalism and Critical Relationships in Health Care' examines why doctors are limited in their ability to admit to error or engage with the system. Barriers include the uncertain nature of their work and care for individual patients. Regulation is a limited approach to ensuring good care but rebuilding organisational engagement - with relationships underpinned by data is possible.

**Angelo Joseph**

Head of Sales Engineering
Google Enterprise

**Google has been investing heavily solving issues around massive scale and big data for several years. Hear an industry viewpoint on examples and observations in this growing area of focus.**

## INDUSTRY INSIGHTS

Thursday 3 April 9.15am

*Big data: Big opportunities, big thinking. Insights from Google*

Angelo holds a bachelor's degree in electrical engineering with a telecommunications major from the University of Technology, Sydney. Angelo leads pre-sales engineering for Google's enterprise solutions for business, government and education in Australia and New Zealand. Angelo brings more than two decades of experience in the IT industry and has been on the forefront of adoption of new IT such as Web, Java, Portals Thin Clients, Identity Management and Cloud. Before Google, Angelo held leadership positions within pre-sales divisions at IBM, Sun Microsystems and Oracle.

# KEYNOTE SPEAKERS


**Dr Brendan Lovelock**

Health Practice Lead
Cisco

## INDUSTRY INSIGHTS

Thursday 3 April 9.45am

*The internet of everything: Building health and wellness into the fabric of the community*

Brendan Lovelock is the Health Industry Practice Lead for Cisco Australia. In this role, Brendan is responsible for developing transformative information technology solutions and services facilitating the delivery of safe, affordable and accessible healthcare. The focus is to drive quality and performance through improved utilisation of healthcare resources across the whole care provider ecosystem. Brendan has an extensive background in business management and technology commercialisation, having held senior executive positions with Telstra and Eastman Kodak, both in Australia and internationally. He has also managed a number of smaller Australian software and consulting organisations, delivering reporting and business management systems into the telecommunications and technology services markets. Brendan has chaired numerous conferences and symposia on healthcare systems design and has published on that subject. He currently leads the Digital Hospital Design group at the Health Informatics Society of Australia and The Foundry, an industry collaborative to bring technology developers and solution designers together with clinicians and healthcare managers.


**Dr Mirana Ramialison**

Group Leader in Systems
Developmental Biology
Australian Regenerative
Medicine Institute

**"Research in understanding individual diseases has been fast-forwarded by -omics technologies, bringing personalised medicine closer."**

## IMPLEMENTATION INSIGHTS

Friday 4 April 3.55pm

*Research perspective*

Dr Ramialison is an NHMRC/Heart Foundation Career Development Fellow. After her PhD at the European Molecular Biology Laboratory in Germany, she conducted her post-doctoral research at the Victor Chang Cardiac Research Institute in Sydney. She is now a Faculty member of the Australian Regenerative Medicine Institute, where she leads her laboratory researching on heart development, evolution and disease using bioinformatics.


**Dr Maia Sauren**

ThoughtWorks/
Open Knowledge Melbourne

**"What can you do in one weekend? Quite a lot, it turns out, if you're willing to collaborate with strangers."**

Friday 4 April 10.00am

*Health Hack 101: An introduction to this brave new world*

Dr Maia Sauren is a biomedical engineering researcher turned software consultant. Maia moonlights for the Open Knowledge Foundation, a not for profit that aims to help organisations make data and information available and open.

# KEYNOTE SPEAKERS


**Dr Bern Shen**

Chief Medical Officer
HealthCrowd (USA)

**Health insurers need to understand risk, which in turn requires accurate information. Mobile health, while no panacea, can provide new insights into the daily health behaviours and choices of patients, potentially benefitting not only payers but also clinicians and patients themselves.**

**IMPLEMENTATION INSIGHTS**

Friday 4 April 9.00am

*Payer perspective (Insurer):*
*Big data analytics to understand and influence human behaviour*

Bern Shen is Chief Medical Officer and co-founder of HealthCrowd. Bern practiced emergency medicine for 15 years and has 15 years of health tech experience in large companies, startups, and as an angel investor. He holds an A.B. from Harvard and an MD, MPhil. from Yale.


**Prof Terry Speed**

Head, Division of Bioinformatics
Walter and Eliza Hall Institute
of Medical Research

**Big Data is data. That's what statisticians have been dealing with for centuries.**

**WORKFORCE**

Thursday 3 April 11.00am

*Putting together a bioinformatics team: 2014 compared with 1997*

Terry Speed's research interests lie in the application of statistics to genetics and genomics, and to related fields such as proteomics, metabolomics and epigenomics. He works at the Walter and Eliza Hall Institute of Medical Research in Melbourne, and is an active emeritus professor in Statistics at the University of California at Berkeley.


**A/Prof Andrew Way**

Chief Executive Officer
Alfred Health

**Big data is nothing new in health, the idea of big data is what's new and exciting.**

**IMPLEMENTATION INSIGHTS**

Friday 4 April 3.25pm

*Executive perspective*

A/Prof Andrew Way commenced as CEO of Alfred Health, Melbourne on 1 July 2009 having had an extensive career in the English NHS. Andrew's main interests are quality and safety, patient access to healthcare, embedding translational and clinical research, and sound financial clinical services. Andrew has been a major driver behind the development of the Monash Partners Academic Health Science Centre, Victoria's first AHSC. Andrew has been appointed to several Ministerial and other advisory committees and is a Director on a number of Boards.

**Julie Brophy**

Manager Health Information Workforce Strategy, Health Workforce Branch, Department of Health, Vic

> **The health information agenda needs a stronger clinician interface to deliver the change we need.**

**WORKFORCE**

Thursday 3 April 11.40am
*Workforce panel*

Julie currently holds the position of Manager, Health Information Workforce Strategy, with the Department of Health Victoria in which role she is responsible for the Victorian strategies to meet the current and future demand for an appropriately skilled and qualified Health Information Workforce. Prior to this she has held other positions in the Victorian Department of Health, including Principal Advisor, Costing Policy and Analysis, which oversaw the development of standards for costing and reporting of cost data in Victoria and Manager Funding Systems and Costing, which was responsible for the development of the Victorian funding models and the cost data collection. She has also held positions with the QLD Department of Health assisting with the implementation of their Casemix Funding Model, and several positions in Victorian health services in roles including managing Decision Support Units, Business Intelligence, Clinical Costing, Strategic Planning, and as a Health Information Manager. Julie originally graduated as a Health Information Manager, but has since also completed post graduate qualifications in Information Technology and Health Statistics.

**Paul Madden**

Deputy Secretary and Chief Information and Knowledge Officer Department of Health

**WORKFORCE**

Thursday 3 April 11.40am
*Workforce panel*

Mr Paul Madden was appointed to the position of Deputy Secretary and Chief Information and Knowledge Officer in 2010. His role includes the development and implementation of visions, strategies and plans for information, knowledge, technology, and performance management. Paul also provides strategic guidance and advice in relation to technical aspects for the various e-health initiatives such as the Personally Controlled E-Health Record (PCEHR), telehealth, ePrescribing and health system performance reporting. Paul has also managed projects to implement the enterprise information management strategy, data governance and enterprise IT governance and approval. Mr Madden is a member of the Departmental Executive Committee and is also the chair of the Departmental Information, Knowledge and Technology Committee which provides advice and makes recommendations to the Executive Committee on information, knowledge and technology strategies and plans. He also Chairs the Data Governance Council which provides advice and assists with the implementation of consistent information management policies and approaches.

# DEBATERS

**Dr Tim Churches**

SURE Epidemiologist and
Technical Advisor
Sax Institute

**The [big data] curse of [high] dimensionality applies as much to personal privacy and confidentiality as it does to database management, data mining and statistical models.**

**PRIVACY**

Thursday 3 April 3.45pm

Tim has designed and implemented several population health and research information systems, including the ANZICS intensive care outcome monitoring system, a population health data warehouse for NSW Health, a near real-time ED surveillance system, a communicable disease outbreak control system, a probabilistic record linkage engine, and SURE, a highly secure remote-access analysis facility for health researchers. He has authored papers on privacy-preserving methods for record linkage and guidelines for personal information disclosure control in research output.


**Dr Peter Croll**

Chief Executive Officer
Peter Croll and Associates

**Privacy with big data - not an obstacle but a balance for better health outcomes**

**PRIVACY**

Thursday 3 April 3.45pm

Dr Peter Croll is a leader in health informatics who founded PeterCroll.com, a consultancy specialising in trustworthy ICT solutions. In Australia he has held four professorships at universities, been appointed as National Fellow for CSIRO, a past Vice President for HISA Ltd. and currently chairs the IMIA working party on Health Information Privacy and Security.


**Jessica Dean**

President
Australian Medical
Students' Association

**Medical students in Australia provide opportunity for observation and cultural change within the medical profession.**

**PRIVACY**

Thursday 3 April 3.45pm

Jessica Dean is the President of the Australian Medical Students' Association, which is the peak representative organisation for Australia's 17,500 medical students. Jessica is a 6th year Medicine/Law student at Monash University. She is currently completing an Honours Project at The Alfred Hospital in Bioethics.

**Emma Hossack**

President
International Association of
Privacy Professionals

**Privacy by design is embedded in the new Australian Privacy Principles. This approach could lead to open data and big data benefits.**

**PRIVACY**

Thursday 3 April 3.45pm

Emma stopped practising law to become CEO of Extensia in 2007. Extensia's shared electronic health record has been deployed across Australia giving Emma a deep understanding of how medical software can be deployed in a privacy enhancing manner. Emma is also CEO of EDOCX, another privacy compliant product, President of the International Association of Privacy Professionals, Vice president of the Medical Software Industry Association and on the Board of CIRCA


**A/Prof Trish Williams**

School of Computer
and Security Science
Edith Cowan University

**PRIVACY**

Thursday 3 April 3.45pm

Associate Professor Trish Williams is the E-Health Research Group Leader in the School of Computer and Security Science, Edith Cowan University, WA. Trish is internationally recognised for her medical information security expertise. She has over 28 years' experience in healthcare computing with 15 years industry experience in general practice and pharmacy computing before joining academia in 2001. Trish is the primary author of The Royal Australian College of General Practitioners Computer and Information Security Standards, advises the General Practice Data Governance Council, and is a member of the E-Health Industry Clinical Safety and Security Committee. Trish is Chair of HL7 Australia, International Co-Chair of HL7 Security, expert of numerous health informatics ISO standards and has over 70 medical information security publications.


**Prof Ingrid Winship**

Executive Director of Research
Melbourne Health

**PRIVACY**

Thursday 3 April 3.45pm

Professor Ingrid Winship is the Executive Director of Research for Melbourne Health and the Chair of Adult Clinical Genetics at The University of Melbourne. A medical graduate of the University of Cape Town, she completed postgraduate training in genetics and dermatology. In 1994, Professor Winship took up an academic position at the University of Auckland and later became Professor of Clinical Genetics and Associate Dean for Research in the Faculty of Medicine and Health Sciences. Professor Winship is a member (immediate past Chair) of the Bio21 Victorian Hospital Director's Forum and a member of the Bio21 Scientific Advisory Council. She is a member of the Victorian Cancer Agency, NHMRC Human Genetic Advisory Committee and the Victorian Life Sciences Computation Initiative Steering Committee. Professor Winship serves on the Board of the Walter and Eliza Hall Institute and the Peter Doherty Institute Council.

# FACILITATORS


**Dr Louise Schaper**

Chief Executive Officer
Health Informatics
Society of Australia

**HISA:**
**Preparing a 21st century healthcare workforce to deliver a 21st century healthcare system.**

**WORKFORCE**

Thursday 3 April 11.40am
Workforce panel

Louise is an innovator and a change agent who doesn't sit still and whose passion and enthusiasm for health informatics is shaping a new future for HISA. Her appointment as CEO came on the back of 10 years of experience in, and applied passion for, health informatics. With a background as an occupational therapist, Louise has a PhD on technology acceptance amongst healthcare professionals and is a graduate of the Stanford executive leadership program for non-profit leaders. Louise is a world leader in health informatics and is intimately connected to Australia's substantial health reform efforts, where e-health is a key enabler to achieving high quality, safe, sustainable and patient-centred care. Under Louise's leadership HISA is leading the discussion in big data in healthcare and hosts Australia's preeminent conference on big data and healthcare analytics.

**Alison Choy Flannigan**

Partner
Health Aged Care and Life Sciences
Holman Webb Lawyers

**PRIVACY**

Thursday 3 April 3.45pm
*Privacy Debate*

Alison Choy Flannigan leads the specialist Health, Aged Care and Life sciences team at Holman Webb. She has over 20 years of corporate and commercial experience. Alison was previously General Counsel of Ramsay Health Care Limited (one of Australia's largest private hospital operators and a top ASX listed company with operations in Australia and offshore) and was previously a partner of a major Australian National law firm. She has also been Company Secretary of Research Australia Limited and a member of a number of hospital committees, risk management committees and advisory boards. In each year between 2008-2014, she was nominated by her peers in Best Lawyers International: Australia as one of Australia's 'best lawyers' in the areas of health and aged care. Alison regularly advises healthcare providers on IT and privacy issues and advised NEHTA on privacy issues associated with the personally controlled electronic health records.

# Health Services

## Better accessibility, productivity and quality of care through digital innovation.

CSIRO's Health Services Theme is already the largest coordinated health services research activity in Australia.

Our research portfolio is focussed around working closely with clinical partners to improve access to health services via broadband and mobile communication platforms, to develop tools for operational and clinical productivity, optimise use of precious resources, and to improve patient safety and health outcomes.

We can help clinicians to do their job better with support tools for clinical decision-making. We can enable better access to services for remote communities. We can improve the way we care for those who suffer from chronic diseases. We can develop better ways to store, manage, share and use health and patient information. With new and intelligent tools that improve the way health services are delivered, we can provide a better quality of care for all Australians.

**e** colin.kelly@csiro.au **w** www.csiro.au/healthservices
**t** +61 2 9372 4525 **m** +61 477 371 391

chia.org.au

CERTIFIED HEALTH INFORMATICIAN AUSTRALASIA

**CHIA:**
**because building and delivering the future of healthcare requires a highly skilled, knowledgeable and experienced workforce.**

# The revolution in little data

Electronic health records (EHRs) have been around for over 30 years now, but it is still hard to share information between health software programs. One of the largely unsolved challenges is how to exchange and re-use clinical data. Development of computable, clinical models that express the detailed clinical data patterns, the 'little data', has been slowly evolving but is now starting to gain momentum, both here in Australia and overseas.

Clinicians and other domain experts are taking on a leadership role to ensure that each clinical model represents the health information they require to support direct patient care, exchange, aggregation and analysis and for clinical decision support. As more organisations use and share these clinical models the health information collected will be captured in a common, standardised pattern, and so sharing and reuse of the data becomes much simpler.

If we collaborate to ensure that the little data is fit for clinical purpose, then we can start to break down the silos of health information, support collection of high quality clinical data and provide a firm foundation for safe aggregation, analysis and re-use of health data. Getting the 'little data' right, will underpin and enhance the outcomes we seek through 'big data'.

LEARNING OBJECTIVES

- What is 'little data' – an introduction to clinical models and archetypes
- Overview of current clinical model activity in Australia and overseas
- Awareness of the critical role for clinicians and other domain experts in lead this approach
- Understanding the benefits of getting the 'little data' right
    - Electronic health records
    - Interoperability
    - High quality big data
    - Re-use of data – research, population health etc
    - Knowledge-based activities e.g. clinical decision support
- How to get involved

TARGET AUDIENCE

- No technical knowledge is assumed
- Clinicians, program leaders and non-technical attendees are particularly encouraged to attend

**Dr Heather Leslie**

Director of Clinical Modelling,
Ocean Informatics

Dr Heather Leslie is an experienced GP who made the transition to health informatics over 15 years ago. She is currently Director of Clinical Modelling at Ocean Informatics, and Clinical Program Lead at the openEHR Foundation. Heather has advised national e-health programs on clinical modelling in Australia and Europe, and conducted openEHR archetype training in Australia, New Zealand, Europe & USA.

**Dr Hugh Leslie**

CEO
Ocean Informatics

Dr Hugh Leslie is a practicing GP and health informatician. He is currently the CEO of Ocean Informatics and has been actively engaged in the international health standards space, including HL7. Bridging health and IT domains, Hugh has led a number of successful Australian projects that have used the direct application of clinical modelling technologies to solve real world problems.

# Health Informatics Society Australia

Improving Australian healthcare through the use of technology and information

## About HISA

We have a vested interest in growing the capacity and capability in health IT and are **passionate advocates for the e-health enabled transformaton of healthcare.**

We are a **not-for-profit, member organisation** with a broad and diverse stakeholder community of 1000+ active members and a database of over 13,000 committed participants in digital health, e-health and health informatics.

Australia's only community of individuals and organisations that share a common passion, expertise and leadership in digital healthcare.

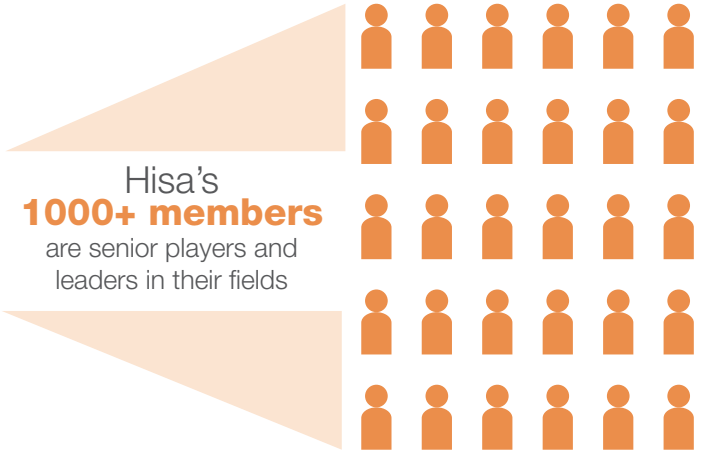
## HISA's Strengths

- Independence and integrity
- Knowledge of the inudustry
- Reach and networks throughout healthcare and health IT
- Understanding of the clinical ecosystem
- Event and content delivery and management
- Intelligence generation, analysis and distribution

We have **access to the best minds in the e-health nationally and globally.**

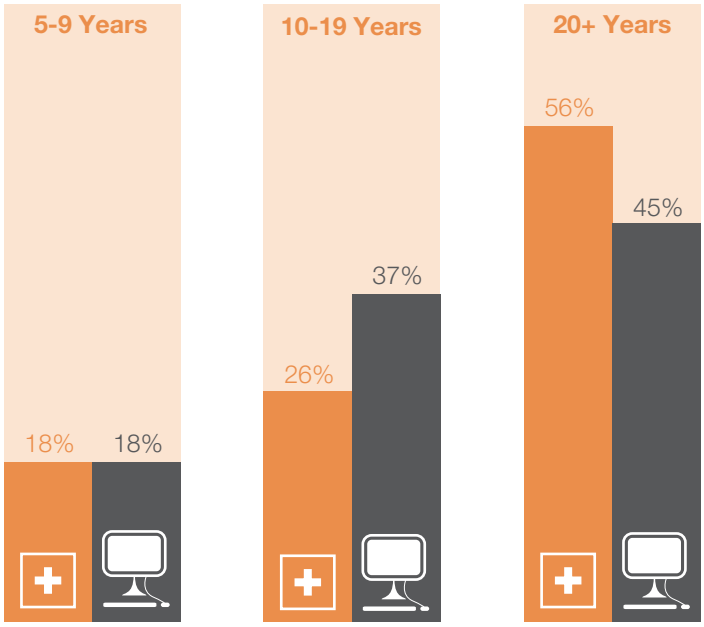
# About our members...

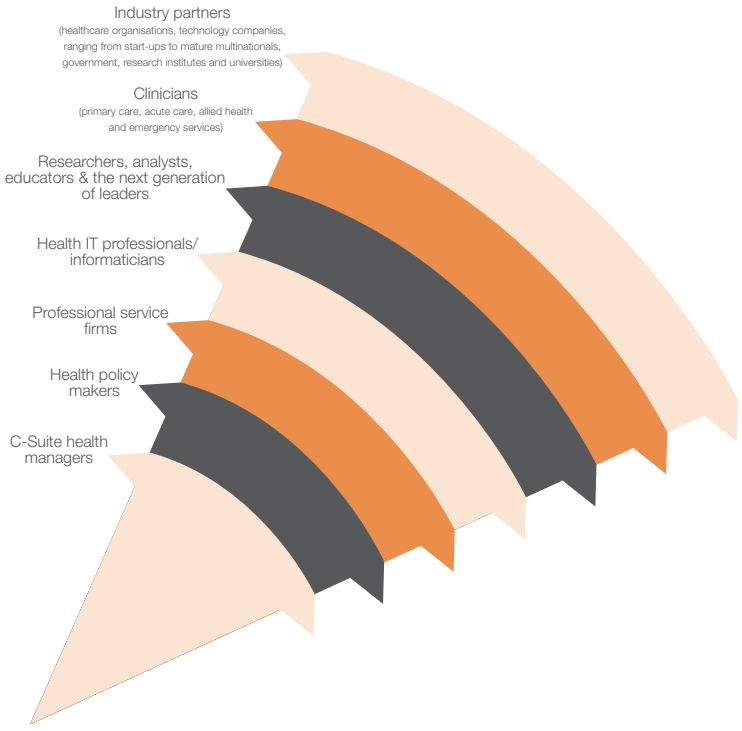
HISA brings together the brightest minds in the business and our events are renowned for their networking & learning opportunities as well as cutting-edge, exeptional and quality content.




HISA members are industry leaders. Together, our membership represents **thousands of years of combined expereince in health and health IT.**

Years of experience:




## HISA's Future

In recent years HISA's membership has grown substantially; **doubled members' equity and tripled our staff.**

Our strategic priorities into the future are:

**Vision**

**Workforce**
Preparing a 21st century healthcare workforce to deliver a 21st century healthcare system

**Leadership**
HISA will lead and catalyse the ongoing improvements in Australian healthcare that are ahceived through the use of IT

**Professional Advancement**
HISA leads the development, promotion & support of the profession of health informatics

**Influence & Advocacy**
HISA provides a national focus for e-health and health informatics and uses its influence to create opportunities for change, collaboration, growth, advocacy, representation and support for our community to drive transformative innovation in healthcare

**Learning, Education & Training**
Delivering learning and development programs that motivate and enable healthcare stakeholders to improve healthcare through the use of IT

**Marketing & Communications**
HISA communications are creative, targeted and regular, use a variety of mediums, reflects HISA's valuses and supports HISA's activities and strategic direction

**Community**
HISA develops a fosters communities of influence

**Organisational Capability**
To build a sustainable, well governed, socially responsible and innovative organisation that is able to deliver HISA's vision


**Manal Almalki**

PhD Candidate
University of Melbourne

Manal Almalki is a university computer science lecturer currently on a Saudi Arabian Government scholarship at the Health and Biomedical Informatics Unit at the University of Melbourne. She is undertaking PhD studies in the field of personal informatics for self-quantification.

# Classification of data and activities in self-quantification systems

Manal Almalki, Guillermo Lopez-Campos, Kathleen Gray, Fernando Martin-Sanchez

Health and Biomedical Informatics Research Unit, University of Melbourne

## SUMMARY

Self-quantification may be seen as an emerging paradigm for health data. In recent years the general public has become more health-conscious, due in part to the self-tracking and quantification technologies that enable the non-expert to easily capture and share significant health-related information on a daily basis (Mehta, 2011). This self-tracking of personal health and fitness data has the potential to introduce new research methods in citizen science, and in formal research into personalised medicine and healthcare (Swan, 2009). Such methods capture data in real tasks, natural settings, and in situ, as well as facilitate the measurement of some health and life aspects longitudinally, with an aim of generating healthcare-related hypotheses. However, this field lacks a systematic approach to classifying these data, and making sense of these observational measurements. This paper reports on our data classification model, and how it can be used in data collection, data analysis, data curation and data exchange.

## INTRODUCTION

Self-quantification contributes significantly to the health big data phenomenon. Self-quantification is the use of multiple self-tracking devices by individuals and populations, and it may generate and aggregate physiological, environmental and genetic data on a grand scale. 69% of U.S. adults keep track of at least one health aspect such as weight, diet, exercise routine, or symptom (Fox, & Duggan, 2013). Smartphone-based fitness and mHealth (mobile health) devices users may globally approach 100 million by 2018, up from 15 million in 2013 (Juniper Research, 2013). Thus, self-quantification can generate data that are big in themselves. Furthermore, people with more serious health concerns are more likely to track multiple health aspects, which consequently could produce huger volumes and a broad range of data types.

Some self-trackers are concerned with helping themselves, and they tend to test random ideas which are not medically proven to be associated, however others are interested to share and compare their data. The intersection between self-quantification and big data poses major challenges in making sense of these data in shared settings, such as support groups, or health research. One challenge is providing a unified language for the measurements that are being made. Over the last few years, we can find much work being done on data classification from the description of health-related states (such as in WHO-ICF), the prescription of mobile health apps (e.g. Happtique), or the function of the health apps (e.g. European Directory of Health Apps). However, we have not seen a data classification that is designed to support aggregation of data generated from personal self-quantification. As yet the field of self-quantification lacks a formal architecture for data and measurements, which could contribute to new discoveries and improved health outcomes.

## DESCRIPTION

We propose a classification model called Classification of Data and Activities in Self-Quantification Systems (CDA-SQS), see Figure 1. This model is adapted from the International Classification of Functioning, Disability and Health (ICF) that has been developed by the World Health Organization (WHO).

Our data classification model is designed with consideration to the following general principles:

- Health and wellness as the basic organising concept.
- Fit within a comprehensive framework for describing self-tracking practices (e.g. tools and technologies, data and measurements, time and location, etc.).
- Reference to pre-existing classification systems developed to account for conventional and unconventional observations of potential influences on a health condition.

The proposed classification model consists of three domains (Figure 1). Each domain has several categories as follows:

1. Body structures and functions domain which includes: mental functions, sensory functions, sensation of pain, voice and speech functions, cardiovascular system, haematological system, immunological system, respiratory system, digestive system, metabolic system, endocrine system, genitourinary functions, reproductive functions, skeletal system, muscular system, nervous system, skin, hair, nails, genome (DNA, RNA and genes), and microbes categories.
2. Body actions and activities domain which includes: learning and applying knowledge, communication, mobility, self-care, domestic life, interpersonal interactions, education, work and employment, economic life, recreation and leisure, and religion and spirituality categories.
3. Around body domain which includes: relationships and attitudes, products or substances for personal consumption, products and technology for use, and natural environment and human-made changes to environment categories.

This classification model describes these domains as interactive and dynamic rather than linear or static. It is applicable to all people, whatever their health condition. It is also relevant to all self-tracking and quantification practice and technologies identified in the authors' prior review (Almalki, Martin-Sanchez, & Gray, 2013).

Our data classification model can be used for describing the vast array of measurements generated in self-tracking. If we think of self-quantification as a way of investigating factors which affect health and fitness, we can see that we need to describe three main components as illustrated in Figure 2. The component number one provides the investigation questions or hypothesises. The second component sets the main attributes of a particular study, the study's sample, the assays, and describes the instruments used in the study. Such instruments are classified into two categories: primary and secondary self-quantification systems (SQS). This SQS taxonomy is explained in detail in Almalki, Martin-Sanchez, and Gary (2013). Also, the second component explains the measurements – this is where our model provides a way to classify such data and their types. The third component is the data generated from the investigation.

## CONCLUSION

Self-quantification produces big data, and has the potential to advance healthcare knowledge. However, it lacks a formal architecture for describing the data that are generated. Our CDA-SQS model for classifying such data overcomes this problem and enables more systematic research in this field.


Figure 1. The proposed CDA -SQS model

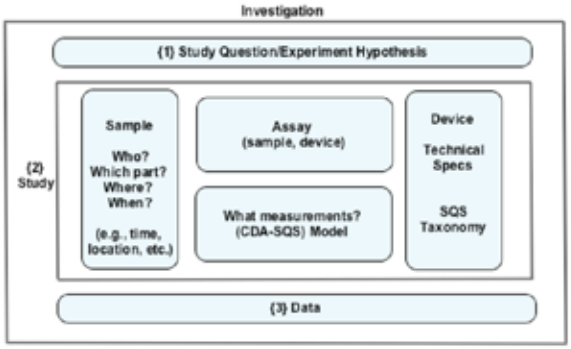



Figure 2. Self-quantification investigation components. {Numbers are used for illustration only}.

## SELECT BIBLIOGRAPHY


1. Almalki, M, Martin-Sanchez, F & Gray, K 2013, Self-Quantification: The Informatics of Personal Data Management for Health and Fitness, Institute for a Broadband-Enabled Society (IBES), The University of Melbourne, Health and Biomedical Informatics Centre, University of Melbourne, 9780734048318, <http://www.broadband.unimelb.edu.au/resources/white-paper/2013/Self-Quantification.pdf>.
2. Almalki, M, Martin-Sanchez, F & Gray, K 2013. The Use of Self-Quantification Systems: Big Data Prospects and Challenges. Proceedings of HISA BIG DATA 2013 conference.
3. Fox, S & Duggan, M 2013, Tracking for Health, Pew Research Center, <http://www.pewinternet.org/Reports/2013/Tracking-for-Health.aspx>.
4. Happtique 2011, Mobile health has taken the world by storm, viewed 6 Feb 2013, <http://www.happtique.com/wp-content/uploads/HAPP_booklet010212hi.pdf>.
5. Juniper Research 2013, Mobile Health & Fitness: Monitoring, App-enabled Devices & Cost Savings 2013-2018, <http://www.juniperresearch.com/reports/mobile_health_fitness>.
6. Madelin, R 2012, The European Directory of Health Apps, PatientView, England, <http://www.patient-view.com/uploads/6/5/7/9/6579846/pv_appdirectory_final_web_300812.pdf>.
7. Mehta, R 2011, 'The Self-Quantification Movement–Implications For Health Care Professionals', SelfCare Journal, vol. 2, no. 3, pp. 87-92.
8. Swan, M 2009, 'Emerging patient-driven health care models: an examination of health social networks, consumer personalized medicine and quantified self-tracking', International journal of environmental research and public health, vol. 6, no. 2, pp. 492-525.
9. World Health Organization (WHO) 2002, Towards a Common Language for Functioning, Disability and Health CF, Geneva, <www.who.int/classifications/icf/training/icfbeginnersguide.pdf>.

**Dr Denis Bauer**

Research Scientist
CSIRO

denis.bauer@csiro.au

Dr Bauer is interested in high-performance-compute-systems for integrating large data-volumes to inform strategic interventions for human health. She has a PhD in Bioinformatics and Post-Docs in machine-learning and genetics, published in Nature Genetics and Genome Research, was invited speaker at Bio-IT World Asia, and attracted more than AU$360,000 in funding.

# Personalised cloud-computed genomics at health-system-relevant scale

Denis C. Bauer [a,b], Piotr Szul[c], Fabian A. Buske[d]

[a] Preventative Health Flagship, CSIRO, North Ryde, NSW, 2113, Australia
[b] Computational Informatics, CSIRO, North Ryde, NSW, Australia, 2113, Australia
[c] Computational Informatics, CSIRO, Marsfield, NSW, Australia, 2122, Australia
[d] Cancer Epigenetics Program, Cancer Research Division, Kinghorn Cancer Centre, Garvan Institute of Medical Research, Sydney, 2010, NSW, Australia

## SUMMARY

Genomic information is increasingly incorporated into medical practice for diagnosis and personalised treatment. However, processing genomic information at a scale relevant for the health-system remains challenging due to computational requirements as well as high demands on data reproducibility and data provenance. Here, we present Next Generation Sequencing Analysis for Enterprises (NGSANE), a Linux-based, High Performance Computing (HPC) framework for production informatics, tailored to the demands and fast pace of personalised medicine, which is available as on-demand virtual cluster in Amazon's Elastic cloud.

## INTRODUCTION

Unprecedented computational capabilities and high-throughput data collection methods promise a new era of personalised, evidence-based healthcare, utilising individual genetic or genomic testing to tailor health management as demonstrated by recent successes in rare genetic disorders[1,2] or stratified cancer treatments[3]. An analysis can take up to 4633 CPU hours per sample to process whole exome sequencing data and produce fully annotated genomic variants (see Figure 1A, CPU-single-threaded). The time, especially in the mapping stage, can be substantially reduced (7 fold) by utilising multithreading on High Performance Computing (HPC) clusters, where parallelisation between and within sample analysis can be easily implemented (see Figure 1A, CPU-multi-threaded).

To achieve minimal time delay between analysis tasks (i.e. mapping, recalibration, variant call, annotation) workflows are commonly automated by means of software 'pipelines'. While high demands are posed on data provenance and reproducibility of these pipelines, individual analysis components depreciate rapidly due to evolving technology and analysis methods, often rendering entire versions of production informatics pipelines obsolete.

Furthermore, the necessary parallelisation requires a large investment associated with compute hardware and IT personnel, which is a barrier to entry for small laboratories and difficult to maintain at peak times for larger institutes. This hampers the creation of time-reliable production informatics environments for clinical genomics. Commercial cloud computing frameworks, like Amazon Web Services (AWS) provide an economical alternative to in-house compute clusters as they allow outsourcing of computation to third-party providers, while retaining the software and compute flexibility.

To cater for this resource-hungry, fast pace yet sensitive environment of personalised medicine, we developed NGSANE, a Linux-based, HPC-enabled framework that minimises overhead for set up and processing of new projects yet maintains full flexibility of custom scripting and data provenance when processing raw sequencing data either on a local cluster or Amazon's Elastic Compute Cloud (EC2).

## DESCRIPTION

Unlike currently available tools like Galaxy[4], BPipe[5], SeqWare[6] or Atlas2[7], NGSANE constructs pipelines based on Linux bash commands, which enables the use of hot swappable, modular components as opposed to the more rigid program-call wrapping by higher level languages or web-based services.

NGSANE separates project specific files from reference data, scripts, and software suites that are common to multiple projects. Access to confidential data is transparently handled via the underlying Linux permission system. A project specific configuration file defining the compute environment as well as the analysis tasks to perform facilitates the transaction between projects and framework. A full audit trail is generated recording performed tasks, utilised reference data, timestamps, software versions as well as HPC log files, including any errors.

Individual task blocks (e.g. read mapping) are packaged into bash script modules, which can be executed locally or on data subsets to test module code, submission parameters and compute environment in stages thereby mitigating the lack of debug-support from higher level languages/submission frameworks. During production, NGSANE automatically submits separate module calls for each individual data set to the HPC queue. This allows different existing modules, parameter settings, or software versions to be executed by changes to the project specific configuration file rather than the software code (hot swapping).

NGSANE gracefully recovers from unsuccessfully executed jobs be it due to failed commands, missing or incorrect input or under-resourced HPC jobs by enabling a clean restart from the most recent successfully executed checkpoint. Workflows can be fully automated by utilising NGSANE's control over HPC queuing systems and by leveraging the customisable interfaces between modules when submitting multiple dependent stages at once.

NGSANE supports the generation of a high-level summary (Project Card) to enable informed decisions about the experimental success. This interactive HTML report provides an access point for new lab members or collaborators, as well as a gold standard that can be used for testing purposes in a continuous integration server framework.

NGSANE is available as an Amazon Machine Image (AMI), which can be deployed to Amazon's EC2 by using, for example, MIT's StarCluster framework (http://star.mit.edu/cluster/) to launch a virtual cluster on demand (see Figure 1B). Other than regular on-demand instances, whose availability is guaranteed at a fixed price, StarCluster also offers command line-based sourcing of Spot Instances, where prices are based on current supply and demand. While Spot Instances can be acquired at a substantially lower price, their availability is not guaranteed. Hence NGSANE's checkpoint recovery is critical in such an unstable, competitive environment. Finally, NGSANE's HPC job partitioning and submission structure is independent from the program calls, therefore allowing new technologies (e.g. Hadoop) to be incorporated.

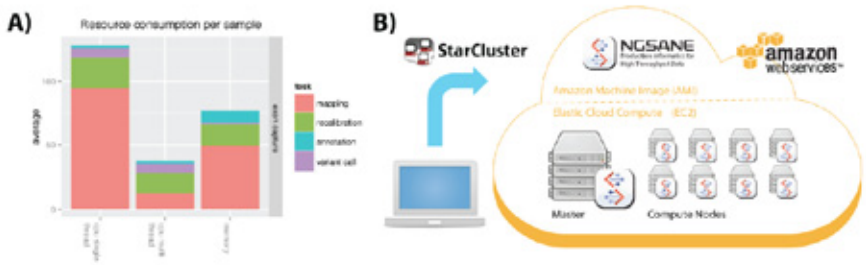



Figure 1. A) Resource consumption of the four steps involved in exon capture genomic data analysis. The average per sample is plotted in hours and gigabytes for CPU usage (single and multithreaded) and RAM memory usage, respectively. B) Schematic for a nine-node on-demand cluster with the NGSANE AMI deployed on every node on the EC2 service as launched by StarCluster.

## CONCLUSION

NGSANE is a flexible HPC framework for NGS data analysis that is specifically tailored to the demands and issues of personalised genomics. NGSANE is implemented in bash and publicly available under BSD (3-Clause) licence via GitHub at https://github.com/BauerLab/ngsane. Currently implemented workflows include those for adapter trimming, read mapping, peak calling, motif discovery, transcript assembly, variant calling and chromatin conformation analysis.

NGSANE is available for local cluster installation or as an AMI to be deployed as an on-demand cluster on Amazon's EC2. This facilitates production-scale processing of large sample numbers and enables research at population scale to produce insights into individual disease risk and stratify treatment for common diseases with impact on the health system.

## REFERENCES


1. Bainbridge, M.N., et al., Whole-genome sequencing for optimized patient management. Sci Transl Med, 2011. 3(87): p. 87re3-87re3.
2. Talkowski, M.E., et al., Clinical diagnosis by whole-genome sequencing of a prenatal sample. N Engl J Med, 2012. 367(23): p. 2226-32.
3. Pellatt, A.J., et al., Genetic and lifestyle influence on telomere length and subsequent risk of colon cancer in a case control study. Int J Mol Epidemiol Genet, 2012. 3(3): p. 184-194.
4. Goecks, J., A. Nekrutenko, and J. Taylor, Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol, 2010. 11(8): p. R86.
5. Sadedin, S.P., B. Pope, and A. Oshlack, Bpipe: a tool for running and managing bioinformatics pipelines. Bioinformatics, 2012. 28(11): p. 1525-6.
6. O'Connor, B.D., B. Merriman, and S.F. Nelson, SeqWare Query Engine: storing and searching sequence data in the cloud. BMC Bioinformatics, 2010. 11 Suppl 12: p. S2.
7. Evani, U.S., et al., Atlas2 Cloud: a framework for personal genome analysis in the cloud. BMC Genomics, 2012. 13 Suppl 6: p. S19.

**Dr Xiongcai (Peter) Cai**

Research Fellow
Centre for Health Informatics,
The University of New South Wales

x.cai@unsw.edu.au

Dr Xiongcai (Peter) Cai is a researcher with a general background in AI with expertise in the fields of machine learning, data mining, social network analysis and health informatics. He has been spending the past decade researching and developing models to better understand behaviours and patterns that relate real world human activities.

# Modeling of time series health data using Dynamic Bayesian Networks: An application to predictions of patient outcomes after multiple surgeries

Xiongcai Cai[a], Oscar Perez-Concha[a], Fernando Martin-Sanchez[b], Blanca Gallego[a]

[a] Centre for Health Informatics, Australian Institute of Health Innovation, University of New South Wales, NSW 2052
[b] Health and Biomedical Informatics Centre, University of Melbourne, Victoria 3010

## SUMMARY

**Objective:** To develop dynamic predictive models for real-time outcome predictions of hospitalised patients.

**Design:** Dynamic Bayesian networks (DBNs) were built to model patient outcomes that dynamically depend on patient's clinical profiles, temporal patterns of ward transfers and surgery data. These models were applied to predict remaining days of hospitalisation (RDH) for patients undergoing multiple surgeries and their performance compared against a static model based on Bayesian networks (BNs).

**Datasets:** Hospital data from a Sydney metropolitan hospital.

**Results:** The basic model uses static information at time of prediction. The DBN model uses static and temporal information extracted from a series of surgeries; DBNs show a significant improvement in patient outcome predictions with respect to the static model.

**Conclusion:** Time series health data can be dynamically modelled by DBNs to improve predictions of outcomes for patients undergoing multiple surgeries.

## INTRODUCTION

Healthcare systems are under increasing pressure to identify strategies to improve the current patterns of care. Although unstructured data analytics have been widely reported in big data, the importance of time series has not been fully explored yet. Real-time prediction of patient outcomes could benefit greatly from big data in health and time series analysis. In particular, prediction of RDH[1] is an important indicator to assess healthcare delivery and hospital management. Unexpectedly long length of stay may negatively impact patients and hospitals in a variety of ways, such as higher costs and increased exposure to adverse events. Current methods do not allow real-time automated stratification of risk. Rapidly identifying those patients at highest risk of extended RDH has a great potential to improve the quality of care, reduce avoidable harm and costs. DBNs[2] are specially suitable to tackle this problem, since they are probabilistic graphical models that allow temporal order, which can better capture the dynamical nature of the healthcare delivery processes: prognosis, treatment selection, surgery and recovery.

In this paper, we aim to develop a DBNs-based prediction model to investigate the roles of time series data for the prediction of RDH for patients undergoing multiple surgeries.

## DESCRIPTION

Medical records of patients who underwent consecutive surgeries at a Sydney metropolitan hospital between 2008-2012 were analysed. There are 5733 records in the dataset. Each admission is characterised by a set of attributes, which include: patient's characteristics (such as age), surgery information (main procedure, number of procedures, length of surgery) and ward type. These attributes, together with days already in hospital, constitute the inputs to the static BN. The outcome to be predicted is RDH. BNs[3,4] are static probabilistic graphical models that consists of nodes and arcs forming a directed acyclic graph, where nodes present domain variables (predictors), whereas arcs represent conditional probabilistic relationships among variables. DBNs[2] represent the evolution of a system over time, allowing a fixed structure network to present variables at multiple time points (slices), containing temporal dependencies between slices.

We learned both structures and model parameters: 1) In order to construct the static BNs (Figure 1), we learned intra-slice structures and reinforced them with domain knowledge. These BNs will be the baselines to compare with the DBNs; at the point of prediction, the BN will contain the information of the current

surgery, whereas the DBN will contain the temporal information of the consecutive surgeries. 2) We then fixed the intra-slice structure and learned the inter-slice dependencies. We computed the conditional probabilistic dependencies between any two-time slices by creating successive two-slice sequences and by learning the DBNs (Figure 2). For testing, we unrolled the DBN to the length of test sequences (Figure 3) and input test sequences as evidences to infer RDH of patients.

In our experiments, we performed 5-fold cross validation in both the learned BNs and DBNs. RDH is discretised into 12 bins. Compared to BNs, DBNs achieved a significant improvement in the prediction of RDH. Specifically, DBNs achieved 72.4% prediction accuracy, whereas BNs 25.8%. This implies a 180% improvement, which might be due to the ability of DBNs to dynamically update the model using temporal information from time series data. It requires about 30 minutes to construct the DBNs with 64-bit Windows 7 Enterprise, 2 cores of Intel® i7-3840QM CPU @ 2.80GHz and 8GB RAM.

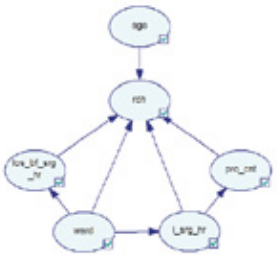


Figure 1. Static BN.

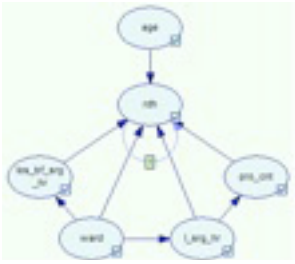


Figure 2. DBN.

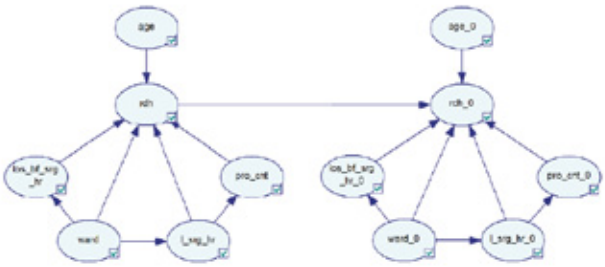


Figure 3. Unrolled DBN.

## CONCLUSION

We developed predictive DBNs models for predicting RDH of surgical patients using patient's trajectories and time series surgical information. Our experiments showed that DBNs significantly outperform BNs in RDH prediction after multiple surgeries. In the future, we plan to further apply DBNs in large-scale big data frameworks for health informatics.

**Funding:** This work was funded by National Health and Medical Research Council (NHMRC) Project grant 1045548 and Program Grant 568612.

**Ethics approval:** Ethics approval was obtained from the NSW Population and Health Services Research Ethics Committee and the NSW Human Research Ethics Committee. The corresponding author was responsible for the data analysis after the extraction. Its contents are the responsibility of the authors and do not reflect the views of NHMRC.

## REFERENCES

1. V. Liu, P. Kipnis, M. K. Gould, and G. J. Escobar, "Length of stay predictions: improvements through the use of automated laboratory and comorbidity variables," Medical care, vol. 48, no. 8, pp. 739-744, 2010.
2. M. KP., Dynamic Bayesian networks: representation inference and learning, UC Berkeley, 2002.
3. L. PJ, v. d. G. LC, and A.-H. A., "Bayesian networks in biomedicine and health-care," Artificial Intelligence in Medicine, vol. 30, pp. 201-214, 2004.
4. J. Pearl, Causality: Models, Reasoning, and Inference: Cambridge University Press, 2000.

Dr Lawrence Cavedon

Senior Lecturer
RMIT University

lawrence.cavedon@rmit.edu.au

Dr Lawrence Cavedon is a Senior Lecturer in the School of Computer Science and IT at RMIT University, and until recently a Senior Researcher at NICTA's Victorian Research Laboratory, where he was a member of the Biomedical Informatics team. Lawrence's current research includes text mining for biomedical applications, spoken dialogue management, and other topics in Artificial Intelligence.

# Text mining for lung cancer cases over large patient admission data

David Martinez[a,e], Lawrence Cavedon[a,b,e], Zaf Alam[c], Christopher Bain[c,d], Karin Verspoor[a,e]

[a] The University of Melbourne
[b] RMIT University
[c] Alfred Health
[d] Monash University
[e] NICTA VRL

## SUMMARY

We describe a text mining system running over a large clinical repository for the detection of lung cancer admissions, and evaluate its performance over different scenarios. Our results show that a Machine Learning classifier is able to obtain significant gains over a keyword-matching approach, and also that combining patient metadata with the textual content further improves performance.

## INTRODUCTION

The increasing availability of linked electronic patient data creates opportunities for analysis, prediction, and automation of tasks. A challenge is that much of this data remains in text format, requiring the use of Natural Language Processing (NLP) techniques to extract actionable information. Text classification according to disease is a crucial technique for retrieving specific cases or creating patient cohorts, for enabling analytics and detection of patterns of disease occurrence, or supporting resource-planning a hospital system. It can also be a prelude to automatic ICD-coding, providing support for an extremely time-consuming manual process.

We describe initial work using data from an Informatics Platform developed at Alfred Health in Melbourne. We investigate the task of automatically assigning the ICD-10 code corresponding to lung cancer (C34, Malignant neoplasm of bronchus and lung) to a patient admission record, via application of a sophisticated text classifier using Machine Learning (ML), over two years of radiology reports from a hospital (756,520 text reports, along with associated metadata) for training and evaluation. We use manually assigned ICD codes to rigorously evaluate performance on different scenarios, using both cross-validation and time-series views of the dataset.

## METHOD

The dataset for this study was extracted from the Alfred Health Informatics Platform (called REASON); it consists of all radiology reports for financial years 2011-2012 and 2012-2013. Each report is assigned an admission identifier, which is in turn linked to patient metadata, including demographics, reason for admission, etc. The metadata includes the ICD-10 codes assigned to the admission, which are used as ground truth to build a gold standard. We define the task as a binary classification problem: determine whether each admission in the test set is associated to the ICD-10 code for lung cancer: C34, Malignant neoplasm of bronchus and lung. An admission is represented by radiology scans linked to it, along with associated metadata.

Classification of lung cancer is a challenging task for automatic systems for two reasons: (i) manually-crafted keywords and phrases produce large numbers of false negatives, and also several false positives; and (ii) for our dataset only 0.8% of the admissions were positive for lung cancer: the highly-skewed nature of the data poses a specific challenge to automated ML approaches, which generally perform better over balanced class distributions.

A classifier was developed using a classical supervised learning framework. For feature representation we combined characteristics obtained from the text, along with the metadata linked to each admission, leaving out any ICD-codes since those are the target for predictions. Text in the reports was processed using the MetaMap tool[1] from the US National Library of Medicine: this identifies phrases and the polarity (negative or positive) of each, using the integrated module NegEx. We created a feature vector combining phrases obtained from MetaMap, the Bag-of-Words (BOW) representation of the text, and the metadata fields. We used the Weka Toolkit[2] implementation of the Support Vector Machine algorithm, since this has performed robustly in our previous work (e.g.[3]). We also tested the effect of applying a greedy correlation-based feature subset selection filter[4].

## RESULTS

We constructed a baseline system using a simple term/phrase-matching approach, using the following (manually constructed) list of terms: “lung cancer”, “lung malignancy”, “lung malignant”, “lung neoplasm”, “lung tumour”, and “lung carcinoma”. The performance of this approach is shown at the bottom of Table 1, using the standard metrics of precision (i.e., positive predictive value), recall (i.e., sensitivity), and F-score (the harmonic mean of them). Precision in particular is low, indicating that many identified phrases were negated or neutral with respect to lung cancer. Recall is higher, but the baseline still fails to identify over one quarter of relevant admissions.

We applied the ML approach outlined above. We report here the results of the basic pipeline without use of feature selection: applying feature selection actually reduced performance, possibly because of the low proportion of positive instances in our dataset. Cross-validation was applied using random stratified 10-fold cross-validation. The results of this experiment are shown in the top two rows of Table 1 for two settings: (i) full feature set (including the metadata described above), and (ii) textual features only. There is clear improvement over the baseline in both cases, particularly in precision. The use of metadata contributes to higher performance, which illustrates the importance of linking different sources of data.

| CLASSIFIER | PRECISION | RECALL | F-SCORE |
|---|---|---|---|
| Full feature set (including metadata) | 0.871 (0.047) | 0.820 (0.057) | 0.843 (0.041) |
| Textual features only | 0.855 (0.048) | 0.800 (0.052) | 0.825 (0.034) |
| Baseline | 0.643 | 0.742 | 0.689 |

Table 1. Results table for the different evaluations. Standard deviation is shown between parentheses.

As a final experiment, we split the data into 3-month periods and performed two tests: (i) Test over each period using all previous history as training; and (ii) Test over each period using only the previous 3-month block as training. The results of this evaluation (using the full feature set) are shown in Figure 1, along with the keyword-matching baseline. We can see that, once we have accumulated enough training, using full history produces higher F-score than using only the previous quarter. However performance reaches a peak and then decreases over the final quarter, suggesting the possibility of changes in reporting that the model does not capture; further analysis is required to build a robust system.

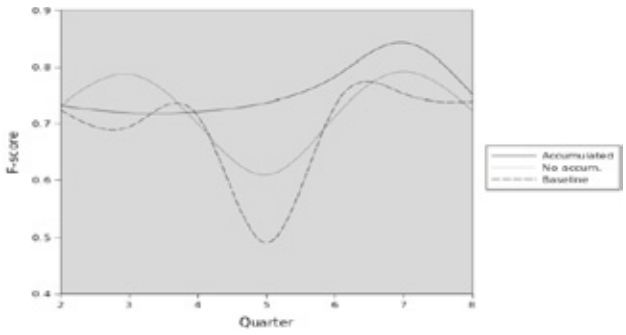



Figure 1. Time-series performance over the different classifiers

## CONCLUSION

Our analysis shows promising results for automatically identifying cases of lung cancer from radiology reports, with results clearly superior to a simple keyword-matching baseline. The experiments also highlight that the model does not always improve with more data, and error analysis is required to interpret the drop in performance for the last 3-month subset of our dataset. While the techniques themselves are fairly standard, an interesting finding is the performance improvement when using metadata on top of the textual features, illustrating the importance of relying on different data sources in building more informed systems. In future work, we plan to integrate other types of clinical information in textual form, such as pathology reports, and evaluate using other disease codes.

[1] Martinez, Cavedon and Verspoor are no longer affiliated with NICTA. NICTA is funded by the Australian Government through the Dept. of Communications and the Australian Research Council through the ICT Centre of Excellence Program.

[2] International Classification of Diseases: http://www.who.int/classifications/icd/en/

## REFERENCES

1. A. R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. AMIA Annual Symposium Proceedings, Washington DC, 2001: 17—21.
2. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten. The WEKA Data Mining Software: An Update. SIGKDD Explorations, 2009, Volume 11, Issue 1.
3. D. Martinez, H. Suominen, M. Ananda-Rajah, L. Cavedon, Biosurveillance for Invasive Fungal Infections via text mining, CLEF Wshop on Cross-Language Eval of Methods, Applications, Resources for eHealth Document Analysis, Rome 2012.
4. M. Hall. Correlation-based Feature Subset Selection for Machine Learning. PhD thesis, Dept. Comp. Sci., U. Waikato, 1999.

# Implementing a clinical genomics infrastructure to sequence 18,000 human genomes per year

Liviu Constantinescu[a], Mark Cowley[a,b], Kevin Ying[a], Peter Budd[a], Derrick Lin[a], Warren Kaplan[a,b], Marcel Dinger[a,b]

[a]Kinghorn Centre for Clinical Genomics, Garvan Institute of Medical Research, 384 Victoria St, Darlinghurst, NSW 2010, Australia
[b]St Vincent's Clinical School, Faculty of Medicine, University of New South Wales, Darlinghurst, NSW 2010, Australia

## SUMMARY

Clinical genomics is a rapidly evolving field focused on the use of genome sequencing information to guide patient diagnosis and treatment. Whole genome sequencing has been dubbed "the test to replace all genetic tests", since one sequencing run can identify all genetic variants present in a patient's genome. Implementing clinical-grade, whole genome sequencing across large patient cohorts represents a substantial big data challenge. We will present our "Sabretooth" plan for scaling operations in our centre from an estimated 800 to 18,000 genomes per year.

## INTRODUCTION

Sequencing of patient genomes is anticipated to have a large impact upon healthcare and the delivery of personalised medicine in three key areas: stratifying patients for appropriate cancer treatment; diagnosing inherited genetic disease; and tailoring prescriptions by anticipating adverse drug reactions.

Recently, the Kinghorn Centre for Clinical Genomics (KCCG) purchased the Illumina HiSeq X™ Ten sequencing system, which has the capacity to sequence 18,000 whole human genomes at an average of 30x coverage, per year. This will generate 150 genomes every 3 days, or 1.4 PB per year.

Although we anticipate that the storage issue can be addressed via currently available computing architectures, the new challenge lies in the delivery of this architecture in a manner that is both sufficiently versatile to keep pace with the rapidly changing bioinformatics landscape and rigorous enough to fulfil the stringent regulatory requirements for clinical data. This presentation will focus on the implementation of modern software development processes and infrastructure adopted by the thought leaders in IT[5], to meet NATA quality standards and allow us the flexibility to continuously improve our processes and analytics.

## DESCRIPTION

Our bioinformatic workflow includes phenotype capture, read alignment, mutation calling, variant annotation and filtering by inheritance pattern, rarity, predicted functional impact and known disease association. Each stage utilises one or more software components, most of which are developed externally. These are supported by information systems that manage clinical data, laboratory processes and logistics. Every study traverses this "Sabretooth" pipeline, from accession to result.

Systems and modules in this pipeline undergo continuous, research-driven change, resulting in increased accuracy and diagnostic sensitivity. As the state of the art advances, obsoleted components must adapt or be replaced. This continuous change has a flow-on effect on subsequent components, and on the middleware interconnecting them. It poses four major challenges: managing software change; adapting and modularising workflows; generating auditable records; and allowing re-runs of legacy pipelines. The first two of these apply equally to clinical and research genomics, whereas the latter two are specific to a clinical context.

To manage software and requirements changes, the KCCG has put an agile software development process in place to continuously improve the modules, applications and information systems that make up our pipeline. By implementing daily stand-ups, feature backlogs, test driven development, automated testing suites, continuous integration and continuous deployment we gain confidence not only in the quality of the software we produce, but in our ability to manage the rapid release/deployment cycle of our systems, recover from hardware failures and roll out new features to the clinical and research arms of our group. Our implementation of the agile process strongly addresses the requirements and recommendations cited as critical to the development of high-quality bioinformatics software in the scientific literature [1,2,5,6].

For high-level management of repeatable, modular workflows, the KCCG have entered into collaboration with

**Dr Liviu Constantinescu**

Information Architect
Garvan Institute of Medical Research

l.constantinescu@garvan.org.au

Liviu Constantinescu completed his PhD in computer science at the University of Sydney as part of the Biomedical and Multimedia Information Technology Research Group, specialising in software development and multimedia technologies. His research focuses on improving the practice of healthcare through state-of-the-art networking and software development methods.

the SeqWare working group at the Ontario Institute for Cancer Research (OICR). We've developed an in-house adaptation of their SeqWare framework, a set of infrastructure tools designed to guarantee the correctness of sequence analysis pipelines and deploy new versions on-the-fly. This framework supports a full hierarchy of functional, scientific and regression tests; retains history and metrics for every run; and incorporates a powerful query engine for interrogating our growing corpus of genome datasets[8].

Finally, a suite of agile process management and documentation tools centred around Atlassian's JIRA[3] augments our pipeline via automatic collection of business intelligence data regarding every stage of the process, guaranteeing end-to-end auditability and allowing clinical, analytical and management teams to tap into continuously updated information that traditional paper-based reporting cannot capture[4]. This information integrates release management, continuous integration and issue tracking, so the scope of every software and analytics change can be constantly monitored in terms of its impact on business and clinical outcomes.

## CONCLUSION

KCCG is leading the charge toward the implementation of large-scale clinical genomics in Australia. We present Sabretooth as a case study in balancing the demands of clinical-grade informatics against the need to manage continuous change, so as to deliver the benefits of the most recent genomic research to all Australian patients in a cost-effective and reliable way.

## REFERENCES


1. K. Rother, et al., A toolbox for developing bioinformatics software. Briefings in Bioinformatics 2011, 13(2), 244–257.
2. K. Beck, Test Driven Development: By Example. Addison-Wesley Professional, Boston, 2002.
3. Jira: Bug tracking, issue tracking, and project management. Available: http://www.atlassian.com/software/jira. Accessed 15/1/2013.
4. D. Larson, Agile Methodologies for Business Intelligence. Business Intelligence and Agile Methodologies for Knowledge-Based Organizations: Cross-Disciplinary Applications. IGI Global, 2012. 101-119. Web. 15 Jan. 2014
5. S. Baxter, S. Day, et. al., Scientific Software Development Is Not an Oxymoron. PLOS Computational Biology 2006, 2(9), e87.
6. D. Kane, M. Hohman, et al., Agile methods in biomedical software development: a multi-site experience report, BMC Bioinformatics 2006, 7:273
7. T. Nyrönen, J. Laitinen, et al. (2012). Delivering ICT infrastructure for biomedical research. Presented in the WICSA/ECSA '12: Proceedings of the WICSA/ECSA 2012 Companion Volume, ACM.
8. B. O'Connor, B. Merriman, et al., SeqWare Query Engine: storing and searching sequence data in the cloud, BMC Bioinformatics 2010, 11 Suppl 12, S2.


## Dr Sarah Dods

Health Services Research Theme Leader
CSIRO

sarah.dods@csiro.au


Dr Sarah Dods leads CSIROs research in health services delivery within their Digital Productivity and Services Flagship. In this role, Sarah leads multidisciplinary research teams focused on supporting the future sustainability of Australia's health system through evidence based digital services innovation to improve healthcare productivity, quality of care, and access to services for all Australians. Sarah has over 20 years experience in multidisciplinary innovation, including mining R&D, high-tech startups, and academia, spanning research and business roles. Her experience includes 13 years researching into future optical broadband networks, which are now becoming an everyday reality.


## Dr Sankalp Khanna

Research Scientist
The Australian E-Health Research Centre, CSIRO

sankalp.khanna@csiro.au


Sankalp completed a PhD in 2010 looking at intelligent techniques to model and optimise the complex, dynamic and distributed processes of elective surgery scheduling. He is currently a Research Scientist at the CSIRO Australian e-Health Research Centre. His research interests include applied artificial intelligence, prediction and forecasting, planning and scheduling, multi agent systems, distributed constraint reasoning, and decision making and learning under uncertainty.

# Big insights into patient flow


Sarah Dods[a], Justin Boyle[b,a], Sankalp Khanna[b,a], John O'Dwyer[b,a], David Sier[a], Ross Sparks[a], Norm Good[b,a], Derek Ireland[b,a], Christine O'Keefe[a], David Hansen[b,a]

[a]CSIRO Computational Informatics, Australia
[b]The Australian E-Health Research Centre, CSIRO, Australia



## SUMMARY

Improving patient flow in hospitals is a significant challenge for any efforts across the globe aimed at addressing overcrowding, improving service delivery and preparing for the rising demand for healthcare. The complexity of the healthcare system however demands multiple improvements across the gamut of the health service to work together if sustained improvements in patient flow are to be delivered. Needing significant volumes of data from disparate sources ranging from hospital information systems to twitter feeds to be processed, often in real time, innovation in patient flow presents significant big data challenges but offers the opportunity to deliver significant benefits to the process. In this manuscript, we present our efforts to deliver improvements to patient flow across various areas of hospital service delivery and demonstrate how this distributed modular approach can help achieve organisational improvements to patient flow.


## INTRODUCTION

The most visible challenge facing our healthcare system is overcrowding in hospitals, which has been labelled an 'international crisis'[1]. Overcrowding and long hospital waiting periods have a significant impact on the quality of patient care and patient experience. Our health services research team is helping hospitals meet performance targets recently introduced by the national health reform, whilst solving the challenge of overcrowding and system bottlenecks. In this manuscript, we provide an overview of the patient flow modelling research currently being undertaken and how our analytics, optimisation and operational decision support tools are working on patient flow solutions across hospitals to deliver big insights into this particularly big problem.


Figure 1. Big data analytics: supporting organisational improvements to patient flow

## DESCRIPTION

In this section, we describe some of our key research efforts aimed at solving various patient flow problems across hospitals to show how the various solutions can fit together to provide big insights and deliver enterprise wide patient flow improvements.

Emergency departments (EDs) are critically overcrowded and struggle to respond to day-to-day arrivals. Contrary to conventional wisdom that emergency patient volumes are unpredictable, the number of admissions per day can be predicted with remarkable accuracy. We have developed the Patient Admission and Prediction Tool (PAPT)[2], that employs historical data to provide an accurate prediction of not only the expected patient load but their medical urgency and specialty, and how many will be admitted and discharged. Our PAPT platform allows hospital management to accurately forecast service demands for inpatient and ED beds, well in advance.

One topic of much recent controversy is "optimum occupancy," or how close to 100% occupancy a hospital can operate at before service efficiency decreases. To help improve understanding of the effects of high occupancy we investigated inpatient and ED patient flow across Queensland public hospitals[3] and identified three stages of system performance decline, or choke points, as hospital occupancy increased. These were found to be dependent on hospital size, and reflected a system change from 'business-as-usual' to 'crisis'. The results indicate that modern hospital systems can operate efficiently above the often-prescribed 85% occupancy level, with optimal levels dependent on the size of the hospital. With this information, hospitals can characterise their individual choke points and determine their optimal occupancy. They are then able to design strategies to better cope when the hospital reaches that occupancy.

Having the right mix of beds is critical for hospitals in maximising efficient service delivery and patient care. We have developed simulation models for patients admitted to inpatient beds from ED. These models can be used to assess how changing the numbers of beds in different specialties affects the waiting times for inpatient beds. The models have been used to determine the percentage of patients discharged within four hours from ED, for a fixed number of beds assigned to specialties in different combinations. The model can also automatically adjust allocation of beds between specialties to find the overall minimum number of beds needed in the hospital to achieve a specified performance.

There is much supposition and guesswork in understanding how hospital occupancy relates to patient safety and minimal hard evidence to date that higher inpatient occupancy equates to a higher likelihood of adverse events. We have explored this important issue through examining the relationship between daily hospital occupancy rates and the occurrence of reported adverse events[4]. The study confirmed that increased hospital occupancy does increase the reported rate of adverse events; in general, for a 10% increase in hospital occupancy, the percentage increase in the incident rate of all reported adverse events was around 20%. This is an important factor to consider in developing capacity management strategies.

A widely recommended strategy for improving patient flow in acute hospitals is to schedule patient discharges for earlier in the day. In the face of little evidence to support this suggestion, some clinicians have questioned the actual benefits of this strategy. We have investigated the effects of varying inpatient discharge timing on ED length of stay and hospital occupancy, to determine the 'whole of hospital' response to discharge timing. We also constructed simulations to model the impact on occupancy levels of shifting all discharges earlier or later[3], providing a tool for hospital staff to see the effect of early discharge.

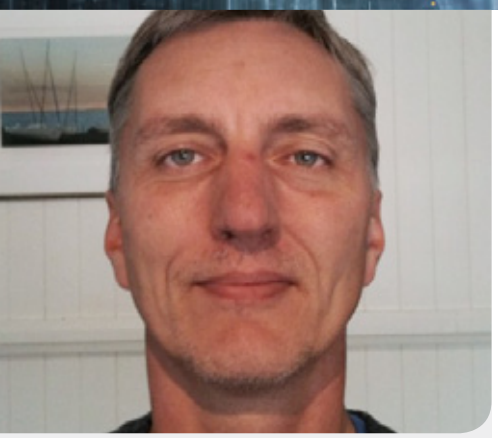
Our other research projects in that space are focused on linking health data from disparate sources in a privacy preserving way, modeling and visualising health data, improving real time disease surveillance using innovative sources such as social media, and employing predictive analytics to reduce unplanned hospital readmissions and support hospital decision making. Hospital administrators thus have a range of solutions available to choose from when addressing patient flow challenges at a service level. Out PAPT solution is currently available to public hospitals across Queensland and is being used to proactively manage bed demand. Our other research, including occupancy analysis, bed planning simulations and early discharge solutions have been used at several Australian hospitals to drive policy and process reform and deliver sustained improvements to patient flow.

## CONCLUSION

This abstract briefly touches on a number of the analyses that we have undertaken together with our hospital partners, and which are providing evidence based solutions to problems of overcrowding and bed capacity in hospitals. Bed demand may seem chaotic, but we have shown that hospital admissions are predictable when data techniques are applied properly to this complex system. Our models have provided information to hospitals to quantify the effect of early discharge on reducing peak occupancy, to show how having the right mix of specialty beds can reduce length of stay in emergency departments, and to show how understanding a hospital's "chokepoint" can inform hospitals as to when to trigger "high occupancy" strategies, to provide a better degree of control. The insights gained from each of these analyses have helped with understanding others too, and together, these collaborative analyses are helping deliver big insights into improving patient flow in hospitals.

## REFERENCES


1. Hoot NR, Aronsky D. Systematic review of emergency department crowding: causes, effects, and solutions. Ann Emerg Med. 2008 Aug;52(2):126–36.
2. Boyle J, Jessup M, Crilly J, Green D, Lind J, Wallis M, et al. Predicting emergency department admissions. Emerg Med J. 2012 May 1;29(5):358–65.
3. Khanna S, Boyle J, Good N, Lind J. Unravelling relationships: Hospital occupancy levels, discharge timing and emergency department access block. Emerg Med Australas. 2012;24(5):510–7.
4. Boyle J, Zeitz K, Hoffman R, Khanna S, Beltrame J. Probability of Severe Adverse Events as a Function of Hospital Occupancy. IEEE J Biomed Heal Informatics. 2014;18(1):15–20.



**Norm Good**

Senior Statistician
CSIRO

norm.good@csiro.au

Norm Good is a Statistician within the CSIRO's Computational Informatics Division. He has been based in the Australian E-Health Research Centre (Brisbane) for the past eight years conducting health related research. Mr Good has considerable expertise in developing and applying variable selection techniques and survival models to health data. Examples of this include novel approaches for estimating optimal colonoscopy screening intervals and developing a risk profiles for patients who are likely to be readmitted to hospital. He has also worked in the field of confidential health data, developing risk and utility measures for regression modelling and promoting the use of remote server statistical querying. Recently Mr Good has exploring the utility of visualisation methods for analysing high-dimensional geometry and multivariate data.

# Health informatics visualisation engine: HIVE

Norm Good[a], Chris Bain[b], David Hansen[a], Simon Gibson[a]

[a]The Australian e-Health Research Centre, CSIRO, Queensland
[b]Health Informatics, Alfred Health, The Alfred, Victoria

## SUMMARY

This project was designed to integrate, analyse, synthesise and present essential health and hospital information in a highly accessible, agile and visual form – because pictures are worth a thousand words.

We developed a prototype software tool that is;

- capable of drawing on standardised data files that replicate known industry standard, or are easily derivable from such standards
- provides the user (analyst, operational manager, financial manger, executive) with a customisable view of the relative outcomes of, and resources used in, care in a number of dimensions- clinical (LOS, number of adverse events, number of drug doses, attending doctor etc) and financial (surgical, pharmacy, nursing etc) - in one setting
- and identify outliers using advanced statistical modelling techniques.

This tool will generate immediate value for a hospitals' endeavour in continuous operational improvement and will be of particular interest to potential customers throughout Australia given the move to nationally provided Activity Based Funding for hospital services. The tool is a useful way to harness the power of "big data" through advanced analytics.

## INTRODUCTION

In Australia, there has been a paradigm shift from input to output (activity) based funding for hospital service delivery. A common output based payment system for managing healthcare provisions involves casemix. Casemix systems are information tools used to classify episodes of patient care. Diagnosis-Related Groups (DRGs) are the most popular casemix system. The rationale from going from an input to output based funding are: i) improving cost efficiency in providing healthcare services; ii) promoting equity for all people to access health services; and iii) increasing incentives to hospitals for providing efficient and quality services.

This shift in funding type has highlighted a number of issues. The one with a significant impact on the costs of care is focused on outliers[1,2]. An outlier in this context is an observation for a patient which is outside the "normal" range expected. Current practices of outlier identification involve a case-by-case analysis, which can be very time consuming. Another issue is to assess the relative clinical and financial value of components of care delivery. These components consist of equipment, drugs and specialist staff which are assessed against mortality and measures of morbidity. Once again, current practice is on a case-by-case basis. Significant cost savings and identification of efficient care models can be achieved if an appropriate software tool can be developed and utilised by front line care.

## DESCRIPTION

ANALYTICS ENGINE - Behind the software interface are two main analytics modules. In the first module we can develop models of predicted costs for treating individual patients. In addition to the standard cost attached to a specific DRG, extra costs based on clinical and demographic parameters can be estimated. The main purpose of this modeling was to develop a profile for the "average" patient. The second module focused on identifying outliers in a data set. Given the predicted cost of care calculated above, the actual cost of care is calculated. An outlier in this context is the difference between the actual and predicted cost for a patient which is outside the "normal" range expected. The current method for identifying outliers in Victoria is the L3H3 method[3]. It uses three times the average length of stay for a particular DRG as the high cut point for outliers. We will be using a more robust method based on statistical discordancy to identify outliers[4]. What is potentially more informative is the detection of "inliers". That is, people whose costs are much lower than predicted. Examination of these patients may reveal insights into optimal care models. Given the multivariate nature of the cost and clinical data, dimension reductions tools such as principal components analysis have been employed to project patients and their associated cost of care and clinical status onto a viewable two or three dimensional space.

The baseline data comes from clinical and financial databases from The Alfred Hospital, statistical analyses are undertaken with the R statistical package and the GUI is developed using web-based Java script.

## RESULTS

Figure 1 is a four dimensional representation of some of the hospital data. A simple point and click screen in the software produced this plot. The x-axis represents the costs associated with ICU, Allied and Pathology. These costs tend to be related to each other. However, the greatest cost is ICU. The y-axis represents costs associated with Nursing and Medical non-surgery. These axes are derived from a method known as "principal components". The size and colour of the bubbles are representative of total costs. The "Yes" and "No" overlaid onto the bubbles indicate whether the patient was alive at discharge. As you can see some of the biggest costs are associated with patients that died, "No". Figure 2 is a screenshot from the software tool showing a parallel coordinates plot. This plot easily shows outliers or inliers and associated clinical/financial measures. A "brush" can be applied to any vertical axis to select subsets of the data. This is only one example of showing the data. Another option is to undertake statistical discordancy analysis on a subset of disease related groups. This effectively "standardises" costs so that we can compare them in a relative way across patients.

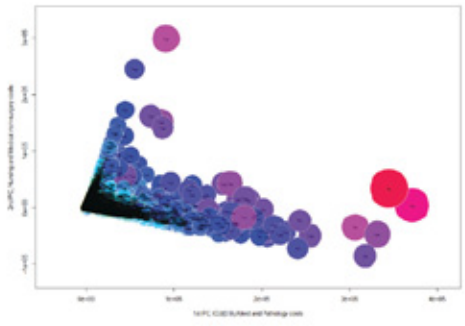


Figure 1. Principal component bubble plot

Figure 2. Parallel coordinates plot.
Different coloured lines represent DRG categories

## CONCLUSION

The development of advanced visual analytics capabilities especially those in the bioinformatics sphere[5] can give greater insight into an organisation data than standard reports alone such as those given by platforms such as Tableau ® and Qlikview®. Such capabilities could serve a very useful purpose when it comes to quickly gaining insights from "big data" data sets. Adding advanced multivariable reduction techniques such as principal components analysis and statistical discordancy can add additional insights into identifying outliers and inliers in hospital administrative and clinical data. It is hoped that this tool will aid in the timely identification of outliers and inliers and provide insight into reducing costs in the future.

## REFERENCES


1. M Harley, M.A. Mohammed, S Hussain, et al. Was Rodney Ledward a statistical outlier? Retrospective analysis using routine hospital data to identify gynaecologists' performance, BMJ (2005), 330:929.
2. E Polverejan, J.C. Gardiner, C.J. Bradely, et al. Estimating mean hospital cost as a function of length of stay and patient characteristics, Health Economics (2003), 935-42.
3. T.E. McGuire, J.A. Bender, C Maskel. Casemix episodic payment for private health insurance. Canberra: AGPS (1995).
4. A.H. Lee, J Xiao, S.R. Vemuri, Y Zhao. A discordancy test approach to identify outliers of length of hospital stay, Statistics in Medicine 17 (1998), 2199–206
5. Visualising Biological Data: VISBI, http://visbi.org [accessed 10/10/2014].

**Dr Adam Kowalczyk**

Principal Scientist
NICTA

adam.kowalczyk@nicta.com.au

Dr Adam Kowalczyk is currently a principal researcher in Victorian Research Laboratories of National ICT Australia (NICTA). He leads projects in molecular medicine and biology, leveraging many years of research and commercial experience in pure and applied mathematics, mathematical physics, artificial intelligence, telecommunications and, recently, bioinformatics.

# GPU enables search for 2-way and 3-way interactions in GWAS

Adam Kowalczyk[a,b], Qiao Wang[a,b], Fan Shi[a,b], Andrew Kowalczyk[a], David Rawlinson[a], Benjamin Goudey[a,b], Richard Campbell[a], Herman Ferra[a]

[a]NICTA, Victorian Research Laboratories, The University of Melbourne, Parkville, VIC 3010, Australia
[b]Computing and Information Systems, The University of Melbourne, Parkville, VIC 3010, Australia

## SUMMARY

Genome-wide association studies (GWAS) probe millions of DNA loci in an attempt to associate DNA mutations with a given disease. Complex aetiologies of many common diseases involve combinations of different genes which require individual evaluation of trillions (non-additive) combinations of loci for association in an average size study. We have developed solutions using a single GPU to evaluate association of each and every one bivariate feature within minutes (available via free webserver). Although an exhaustive tri-variate analysis requires currently a medium size GPU cluster, many focused tri-variate analysis tasks can be accomplished routinely on a single GPU within hours of computation.

## INTRODUCTION

In recent years, GWAS has been considered a fundamental instrument in unveiling the genetic aetiology of non-Mendelian complex diseases. More than 600 GWAS have been conducted for 150 different diseases and traits. Current genotyping technologies such as Affymetrix 6.0 allow GWAS assays of several million single-nucleotide polymorphisms (SNPs). In order to reveal the underlying biological mechanisms of disease, most analytical methods analyse each SNP individually for association with disease[1]. However, interactions between loci are believed to contribute to complex diseases with non-negligible joint effects[2], even while each SNP may show little effect independently[3]. As recently as 2010 it was considered technically impractical to exhaustively search for second-order (bivariate) interactions without access to state of the art computing facilities due to the large search space (of ~1011 for all pairs and ~1016 for all triples in typical GWAS with 500,000 SNPs). In response, researchers proposed several pre-filtering and stochastic partial search methods using univariate analysis for cutting the number of probes to be considered for multi-locus investigation. However, the worry remains that such methods may miss critical genuine interactions which may show only very weak marginal effects[3,1,5].

## DESCRIPTION

Spectacular progress in commodity computing technology in the last few years has led to development of a number of algorithms capable of exhaustive bivariate analysis not only on moderate computer clusters but also on standard desktop computers equipped with General Purpose Graphics Processing Units (GPUs). In order to fully exploit the potential of these devices for medical and biotechnological research there is a need for efficient software tools that reduce implementation and performance difficulties, so researchers can focus on comparison and evaluation of results rather than software tools development and low level algorithm tweaking; see a recent elaboration of this point in[6]. It has been demonstrated[7] that the number of novel putative epistatic loci can be detected using such techniques.

In this talk we present an efficient library of GPU kernels that can be used for fast implementation of any bivariate GWAS statistics that can be derived from contingency table counts. In order to illustrate the point we have implemented nine different algorithms from the literature. All algorithms can execute exhaustive search on typical case/control GWAS (500K SNPs, 5K samples) within 10 minutes. This performance makes comparative analysis of different statistical methods easy. These results are also significantly improved compared to original implementations of the nine algorithms considered. Speedup factors of over 300 are observed compared to some original GPU implementations in literature and even larger factors of over 10,000 are seen with respect to the CPU implementations, e.g. a popular Fast Epistasis algorithm in PLINK 1.07 software package.

Consider that timing scales quadratically with the density of genotyping markers used. Future high-density SNP arrays will include up to 5 million SNPs, and forthcoming GWAS based on NGS data will have even higher marker density and may include other technology such as methylation markers. Together, these expectations mean that the computational burden of exhaustive bivariate analysis will continue to be challenging: between one hundred and one thousand times more complex than existing GWAS. Thus our GPU approach, which

efficiently applies a battery of statistical tests to exhaustive search of all SNP pairs in minutes for current GWAS data, will still require hours or days for near-future data. This is achievable on a single GPU-equipped desktop or laptop computer, but time scales down accordingly if GPU clusters are used. Additionally, the users can define and add their own statistics to our platform and make use of our high performance library that generates contingency tables and ranks scores. To facilitate this we have insured compatibility with popular input data formats, a common interface for defining statistical tests. The runtime of ~10 minutes for a typical dataset and computer is fast enough that researchers can experiment with methods interactively, reviewing the effects of varied algorithms almost immediately.

The practical consequences of such an improvement in productivity are not a matter of degree. By enabling researchers to conduct GWAS experiments in less time than a coffee break it becomes possible to focus effort on statistical methods and results rather than avoiding performance bottlenecks. New ideas can be implemented and evaluated in very fast cycles, without the need to book time on shared high end computing resources.

Most recently, we have extended GWIS to exhaustive search for 3-way interactions, a previously impossible computational task. Using our methods, an exhaustive 3-way analysis of Celiac disease GWAS from UK containing ~310K SNPs and 2200 samples using a cluster of 200 GPUs requires 7 days of computing time. To our knowledge this is the first time such an analysis has been shown to be practical. The runtime reduces significantly for more targeted analysis, for example a specific DNA region or a preselected set of SNPs. Exhaustive filtering through all SNP-triplets in ~2500 SNPs, including the extended MHC region, requires <3 minutes on a standard PC with a single GTX470 NVIDIA GPU.

## CONCLUSION

In conclusion, analysis of two-way and three-way interactions in modern GWAS using multiple methods is practical today. Once such analyses are practiced in labs around the world faster progress in unveiling the genetic aetiology of complex diseases may result. To date, it has been difficult to compare methods across a range of datasets due to implementation difficulties and prohibitive runtime. Many existing benchmark studies were forced to use only small size, typically synthetic, datasets. What we claim is a paradigm shift, towards routine usage real life data for methods development, benchmarking and then “production” deployment of novel GWAS data analysis paradigms.

## REFERENCES


1. Cordell, 2009, Nat.Rev.Genet, 2009, 10(6), 392–404.
2. Marchini et al., 2005, Nat Genet, 37(4), 413–417.
3. Zhang and Liu, 2007, Nature Genetics, 39(9), 1167–1173.
4. Culverhouse et al., 2002, Am. J. Hum. Genet., 70, 461-471
5. Zhang et al., 2010, J. Comp. Biol. 17(3), 401–415.
6. Wilson et al. (2014, PLoS Biol, 12(1).
7. Godey et al. (2013), BMC Genomics 14.

**Dr Kiran Pohar Manhas**

Post Doctoral Fellow
University of Calgary (CANADA)

kiran.p.manhas@albertahealthservices.ca

Dr Manhas is Post-Doctoral Fellow at the University of Calgary, and funded by the Alberta Centre for Child, Family and Community Research. She has degrees in pharmacy, health research, law and bioethics. Her research interests are in paediatric bioethics, privacy, data sharing and stakeholder engagement.

# Parent perspectives on the secondary use of birth cohort data

Kiran Pohar Manhas[a,b], Stacey Page[a], Nicole Letourneau[a], Suzanne Tough[a,b]

[a]University of Calgary, Alberta, Canada
[b]Alberta Centre for Child, Family & Community Research, Alberta, Canada

## SUMMARY

Parents who have given permission to enroll their children in longitudinal birth cohorts and research repositories have agreed to provide researchers with access to a vast amount of data on themselves and on their children's growth and development. These parents' perspectives on data sharing are critical: they are gatekeepers to this data and they are the guardians of child research participants. Their opinions and preferences have not been captured to this point in the literature. Consultation with research participant stakeholder groups is essential for establishing and maintaining respectful and mutually-beneficial research relationships.

## INTRODUCTION

"The value of data lies in their use"[1]. Worldwide, funders and custodians of public research encourage, if not mandate, the sharing of data[2-4]. The recognised benefits of secondary use of research data, typically held in Research Data Repositories (RDRs), include (a) increased diversity, novelty and complexity of research opportunities; (b) cost savings through economies of scale to benefit the public, funders, researchers and trainees; (c) lessened risk of "failure to discover"; (d) promotion of intra- and inter- disciplinary research; (e) maximisation of research participants' contributions; and (f) lessened future research and respondent burdens[2-5]. As RDRs are relatively new, processes that facilitate re-use while honouring participant preferences have not yet been fully developed. Issues that must be addressed include consideration of privacy, risk and consent over time, especially for child participants; access and ownership; and, regulatory governance[6-7].

RDRs can contain extensive information such as parental and child data, collected across the lifespan including biological, epidemiologic and longitudinal data on health, lifestyle, development and service utilisation. RDRs must address concerns relating to the operational and practice standards for maintaining and using these data. Some commentators believe the standards for biological vs. non-biological data to be quite divergent, given their manner of collection, potential for de-identification and storage requirements [8]. In applying standards of biobanks to non-biological research repositories, the latter may be subjected to overly restrictive or inappropriate requirements; conversely, important considerations may be overlooked. Those providing the data, the research participants, have a vested interest in determining standards of practice for research repositories including those relating to privacy, consent, access and communication. Current research on parent perspectives has focused primarily on biobanking; little has been undertaken in the area of non-biological RDRs. In this research, we have undertaken qualitative research to understand parent opinions about data repositories; these findings are critical to build and sustain trust in RDRs.

## DESCRIPTION

We used qualitative methods to describe the perspectives of parent who participate in a longitudinal pregnancy cohort on the secondary use of their, and their child's, non-biological data[9]. The study sample of parent participants was drawn from two Alberta pregnancy cohort research studies. Purposive sampling was used to identify parent participants who were fathers, maternal ages both older and younger than 30 years, visible minorities, and new immigrants. Thirty-seven people consented to take part in this study, 19 in individual interviews and 18 in focus groups (4 groups of 3-6 participants). Semi-structured interview guides were used to elicit parental perceptions regarding (1) the nature of research; (2) motivations to participate in research; (3) the benefits and risks to sharing research data and RDRs; (4) the strengths and weaknesses of RDR governance strategies; (5) 5 alternatives for RDR consent; and (6) the role, if any, of the burgeoning maturity of child research participants. Interviews and focus groups were audio-recorded and confidentially transcribed. A coding framework was developed using methods described by Patton and informed by focus group methodology[10]. Institutional ethics approval was obtained prior to the start of this study.

Both positive and negative opinions towards non-biological research data sharing and RDRs were revealed. There are several points of contention for parents, which will directly impact RDR implementation. Under positive perceptions, parents recognised (a) the overwhelming value and benefits to society, researchers,

participants and funders for data to be retained and shared, and (b) the rigour, trustworthiness and protectiveness of RDR governance strategies that control data access using applications, access criteria, committee oversight, and external regulation. Under negative perceptions, parents expressed concerns that (a) parent and child privacy is at risk and identity protection is a paramount concern; (b) 2 consent alternatives were inappropriate and inefficient: traditional, opt-in consent and opt-out consent; and, (c) governance strategies are weak around ensuring accountability of secondary researchers once access is granted. Under contentious issues, parents disagreed on (a) how child data varies from adult data; (b) how to recognise the burgeoning maturity of child research participants; (c) the most appropriate consent option (i.e. no clear preference between broad, broad-periodic, and tiered consent); and, (d) how to involve the public in RDR governance (see Table 1 for sample transcript quotes).

| THEME | SAMPLE QUOTE |
|---|---|
| Positive Perceptions | "I think in general if [data's] used properly and … for a noble cause then I [sic] totally agree with [data sharing], because we can save a lot of time and there is … a lot of things that are already collected to certain questions that you can reuse on [sic] certain parameters, that you can use for different research, so [sic] they can have a base and I'm sure every researcher is going to continue to dig further but why go back to square one when you have already some grass roots there." [mum, visible minority, new immigrant] |
| Negative Perceptions | "… if you [the RDR] don't know exactly who's using [the data in the future] then I wouldn't want certain things associated with my data, I know that makes the data maybe less useful but, yeah for me it would be more like name and, and things like that, age I see how that's beneficial." [mum, < 30 years] |
| Contentious Issues | "Well I think that whenever it's children [sic], you want sort of higher safeguards right and, and more checks and more security, particularly around the personal identification type of stuff because it's not [sic] them deciding, it's their guardians or whatever." [mum, ≥ 30 years]<br>"I just don't think that [sic] children's data versus adult's data should be something differently. I think they should all be. If someone's participating in the study, you're going to protect that information. It should be protected whether its child or an adult." [mum, ≥ 30 years]<br>"If you started doing that, if you contacted [children] when they became an adult or whatever that would be another logistical [issue]. I think as a parent you just trust that, you just trust the consent of the parent when the child was a child, from my opinion anyways." [mum, < 30 years]<br>"I don't really think [the child] should be involved in decision making, you know I think that she should have the option that [at] whatever 18 or whatever it is in the province to have her data opted out, or taken out if she's really that you know passionate about [it]." [mum, ≥ 30 years] |

Table 1. Sample participant quotes representing major themes recognised in transcripts

## CONCLUSION

Our findings suggest that research participants in a community-based, descriptive, non-intervention longitudinal cohort are supportive of non-biological data retention and sharing. Parent participants expressed trust in the original research team, which seems to extend to the research enterprise and RDRs when accompanied by altruistic motivations and detailed governance strategies. Parent participants dislike extreme consent options that are too active or too passive for parents. RDRs must prioritise protections of participant privacy and mechanisms to ensure secondary researcher accountability. Our findings suggest that parents are not universal on their preferred approach to handling the uniqueness of child data and the burgeoning autonomy of child research participants. We must understand the nuances of this divergence so that we can develop strategies to effectively and appropriate address it. Future research is necessary on child and adolescent perspectives on data sharing and consent to RDR participation, and on clarifying the preferred consent model for parent research participation. The implementation of leading edge data repository governance and data sharing strategies that protect privacy and address consent will enhance the development of new knowledge. This new information can be applied to policies and programs to improve outcomes for children and families.

## REFERENCES


1. Committee on Transborder Flow of Scientific Data, National Research Council. Bits of power: Issues in global access to scientific data, National Research Council, Washington, 1997.
2. Social Sciences and Humanities Research Council. Research data archiving policy. http://www.sshrc-crsh.gc.ca/about-au_sujet/policies-politiques/statements-enonces/edata-donnees_electroniques-eng.aspx, Updated 2012, Accessed February 28, 2013.
3. Medical Research Council. MRC policy and guidance on sharing of research data from population and patient studies, MRC, London, 2011.
4. National Institutes of Health. NIH data sharing policy and implementation guidance. http://grants.nih.gov/grants/policy/data_sharing/ data_sharing_guidance.htm, Updated 2003, Accessed February 28, 2013.
5. Office of the Information & Privacy Commissioner for BC, Report of the roundtable discussion on access to data for health research, Office of the Information & Privacy Commissioner for BC, Victoria, 2012.
6. J. Samuel, N.M. Ries, D. Malkin, B.M. Knoppers, Biobanks and longitudinal studies: Where are the children? GenEdit 6(3) (2008), 1-8.
7. A. Cambon-Thomsen, E. Rial-Sebbag, B.M. Knoppers, Trends in ethical and legal frameworks for the use of human biobanks, European Respiratory Journal, 30(2007), 373-382.
8. B. Brakewood, R.A. Poldrack, The ethics of secondary data analysis: Considering the application of Belmont principles to the sharing of neuroimaging data, NeuroImage, 82(2013), 671-676.
9. M. Sandelowski, What's in a name? Qualitative description revisited, Research in Nursing & Health, 33(2010), 77-84.
10. M. Patton, Qualitative research and evaluation methods, 3rd. ed., Sage Publications, Inc., Thousand Oaks, CA, 2002.

**Mahtab Mirmomeni**

Software Engineer
IBM Research Australia

mahtabm@au1.ibm.com

Mahtab Mirmomeni is a software engineer in IBM Research Australia. She was previously studying Master of Science (computer science) at the University of Melbourne.

# Resolving ambiguity in genome assembly using high performance computing

Mahtab Mirmomeni[a,c], Thomas Conway[a], Matthias Reumann[b] , Justin Zobel[c]

[a]IBM Research Australia
[b]IBM Research Zurich
[c]Department of Computing and Information Systems, The University of Melbourne

## SUMMARY

DNA sequencing has revolutionised medicine and biology by providing insight into the nature of living organisms. High-throughput shotgun sequencing creates massive numbers of reads in a short period of time and de novo assembly attempts to reconstruct the original sequence, as closely as possible, using these reads. Longer pieces reconstructed by assemblies, shed more light on the underlying organism's biology. Repetitive sequences in the DNA, create ambiguities in the assembly which result in shorter fragments. In this project, we explore the search space of the assembly graph construction using the high performance computing capability of an IBM Blue Gene/Q and develop an algorithm that improves assembly quality through deeper search for valid longer sequences around repeat areas. Our results show that we can increase N50 of contigs by 4% and the number of contigs over 1000bp by up to 7%, however, this extension comes at the cost of using a great deal of computing power.

## INTRODUCTION

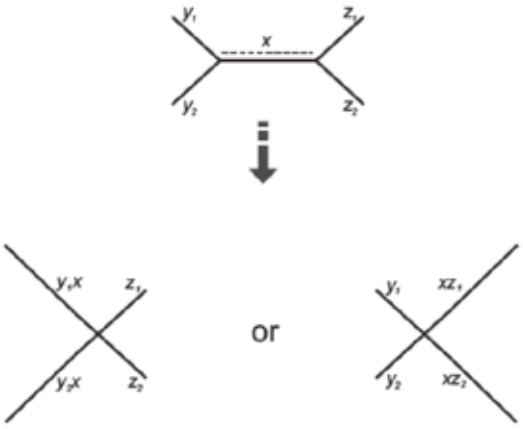
Genome sequencing has become an indispensable part of biomedical research. It enables researchers to understand the structure of genomes and the relationship between species. High throughput sequencing technologies produce large numbers of reads from the DNA, in a short period of time. These reads need to be merged to rebuild the original DNA sequence. This process is called assembly; de novo assembly is a type of assembly where the reads are the only information used in this reconstruction. The programs that perform assembly are called assemblers. Many assemblers create a graph (such as the de Bruijn graph) from the reads and traverse the graph to output unambiguous fragments called contigs. To construct a de Bruijn graph for assembly, reads are first decomposed into k-mers, which are strings of k nucleotides where k is a positive integer, which would become the nodes of the graph. Two nodes are connected together if the suffix of the source node shares an exact match of length k - 1 with the prefix of the destination node. Each read induces a path in the graph and the process of assembly will be mapped to finding an Eulerian path in the graph.

Assemblers have limitations when dealing with repetitive regions of DNA, since it is not be clear how many copies of a repetitive sequence exist and how the graph should be traversed. In ambiguous situations, the assemblers rely on heuristics or break the consigns at points of uncertainty. The most common heuristics used in common assembly programs exploit local graph topology in a greedy manner. In this project, we explore the opportunity of extending the length of contigs by further searching the assembly graph performing a semi-global optimisation to identify valid sequences to produce more meaningful assemblies.

## DESCRIPTION

In order to investigate the effect of an extensive search in the assembly graph around repetitive regions, on the length of the contigs reported by the assembler, we studied the assembly graph produced by Gossamer [1] using the human genome reads published by Gnerre et al. [2]. The assembly graph in Gossamer, called the supergraph, is the result of identifying the Eulerian paths in the de Bruijn graph. The edges of the supergraph are the contigs, which are the output of our assembly. Repetitive elements, create complexities: for example, a repeated sequence presented by contig x can be preceded by contigs y1 and y2 and followed by contigs z1 and z2. These situations are called tangles. When Gossamer encounters a tangle, it outputs x, y1, y2, z1, and z2 separately. Given that x can be followed with either z1 or z2, both x.z1 (x followed by z1) and x.z2 are correct sequences in the underlying genome.

Similarly y1.x and y2.x are both correct sequences. Reporting all four combinations can result in over-estimating the number of instances of x in the genome. We thus 'expand' x, either from left and report y1.x, y2.x, z1, and z2 as contigs or from right and report y1, y2, x.z1 and x.z2.




Given that the assembly supergraph in our human Gnerre dataset contains over 86 million contigs, we estimate that the amount of memory required for our Gnerre dataset is over 87GB. In addition, over 13 million tangles have to be expanded. To tackle this problem, we have divided the supergraph into smaller partitions and used high performance computing (HPC) to process each partition in parallel, in a reasonable amount of time. The cost of an exhaustive search in the supergraph to expand all tangles is exponential, and therefore requires an infeasible amount of computing power. Thus, instead of an exhaustive search to find the best set of tangle expansions in a partition of the supergraph, we have implemented a heuristic search, randomly expanding the tangles in that partition a number of rounds and recording the lengths of the produced contigs. Our algorithm ran on 512 CPUs for 50 hours. Our results show that it is possible to create longer contigs, however, we used around 8 times additional computing power to the assembly algorithm, to gain this improvement.

## CONCLUSION

In this project, we explored the possibility of producing longer, more meaningful contigs by extending contigs around repeat regions instead of breaking them into separate contigs. The repeat regions create complex structures in our assembly supergraph called tangles. We used the structure of the graph and searched more deeply in the assembly supergraph produced by Gossamer[1] to find the best set of expansions for the tangles. Because of the size of the Gnerre dataset, we had to partition it's supergraph and use high performance computing to process different parts of the graph concurrently.

## REFERENCES

1. Conway, T., Wazny, J., Bromage, A., Zobel, J. and Beresford-Smith, B. Gossamer -- a resource-efficient de novo assembler. Bioinformatics (Oxford, England), 28, 14 2012), 1937-1938.
2. Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., Sharpe, T., Hall, G., Shea, T. P., Sykes, S., Berlin, A. M., Aird, D., Costello, M., Daza, R., Williams, L., Nicol, R., Gnirke, A., Nusbaum, C., Lander, E. S. and Jaffe, D. B. High-quality draft assemblies of mammalian genomes from massively parallel sequence data.

**Dr Priscilla Rogers**

Manager
IBM Research Australia

prrogers@au1.ibm.com

Priscilla Rogers is a Research Staff Member at IBM Research - Australia, and the manager of the Laboratory's Healthcare Research team. In this role, she is responsible for driving the research agenda in healthcare, and is passionate about data-driven healthcare and pioneering technologies for transformative plays in the sector. Priscilla studied Mechanical Engineering at Monash University, and holds a PhD in acoustic microfluidics for lab-on-a-chip device applications, also from Monash.

# Spatio-temporal visualisation of disease incidence and respective intervention strategies

Stefan Von Cavallar[a], Matthew Davis[a], Kelly L. Wyres[a], Matthias Reumann[b], Martin J. Sepulveda[c], Priscilla Rogers[a]

[a]IBM Research – Australia, 204 Lygon Street, Carlton, VIC 3053, Australia
[b]IBM Research – Zurich, Saeumerstrasse 4, 8803 Rueschlikon, Switzerland
[c]IBM Research – Watson, 1101 Kitchawan Road, Yorktown Heights, NY, 10598 USA

## SUMMARY

The ability to effectively collect and leverage information offers rich new insights, which can greatly inform decision making within healthcare practice and health systems. In this study, we show how information can be collected, analysed and presented in new ways to inform key decision makers in understanding the prevalence of disease and the response to interventions. We demonstrate this for malaria, using data sources from Kenya, where malaria is one of the three biggest causes of mortality for children under the age of 5 in the sub-Saharan continent of Africa[1].

## INTRODUCTION

Health research and "big data" have the potential to provide powerful new insights. Frameworks, which have the ability to collect and store data in a secure manner, organise and prepare the data, conduct advanced analytics, and display the data in a visually compelling way offer many possibilities to improve healthcare planning and delivery. A pressing problem in developing nations is childhood mortality. Despite the availability of effective intervention strategies, pneumonia, diarrhoeal disease and malaria remain the top three causes of mortality for children <5 years in Africa[1]. In this region there remains a lack of accurate disease incidence and healthcare data with a spatiotemporal resolution appropriate for effective intervention planning and implementation. In this study, we developed a web-based visualisation portal using select and simulated data sources that would allow relevant bodies, such as public health officials, to visualise the disease burden of a particular disease and the potential impact of available and relevant interventions.

## DESCRIPTION

The Cognitive Healthcare and Health Systems Hub is a framework that provisions the ability to collect, store and analyse data from many data providers in a secure manner. The visualisation portal allows relevant bodies to obtain and visualise tailored insights into data within the Hub. This includes a portal to visualise the disease burden of a particular disease and the potential impact of available and relevant interventions. The portal employs a geospatial 3D environment in which to visualise data, enabling the user to view information from a sub-county/town/house level through to a country wide overview. By further augmenting the disease model data with information representing other entities, such as points of interest (i.e. healthcare clinics, pharmacies etc) we can obtain a deeper insight into the disease and appropriate resources available to the different regions.

At its most basic, the disease model data is visualised by utilising a combination of D3[2] and Cesium[3]. The D3 framework is used to process the visualisation data and transform it into a form ready for presentation. The Cesium framework is used to present the visualisation data within a geospatial 3D environment, for example displaying sub-county boundary lines, sub-county names, points of interest etc., right through to rendering the level of red shading within each sub-county region to represent the estimated incidence of disease. Enhanced user functionality is provided by the framework, allowing the user to both navigate the 3D environment and clearly see how the disease burden changes with space and time.

The Spatiotemporal Epidemiological Modeler (STEM), an open source tool to simulate models of disease, was used to generate data for an epidemic malaria outbreak and associated public health interventions in Kenya[4]. STEM contains a malaria disease model to simulate disease transmission by region using environmental and earth science data sources, including elevation, temperature, precipitation, and vegetation coverage[5,6].

To model malaria in Kenya and integrate the intervention scenarios, we made several changes to the included models and data in STEM. First, we replaced the default Kenya data with higher resolution political and earth science data sources, enabling simulations at sub-county resolution. Next we modified the program to modulate disease model parameters by region. This allows us to target interventions for specific regions. For the “distribute bed nets” intervention, the malaria model’s biting rate was modified; for “spray insecticide”, we constrain the mosquito population. To feed simulation data directly to the visualisation portal, we also implemented a JSON data logger in STEM. Model parameters for the baseline and intervention scenarios were derived from literature.

The portal enables the user to select which scenario output to visualise. Such visualisation allows intuitive access to information essential for those interested in understanding transmission dynamics or planning disease control strategies. Future versions could also link directly to the simulation engine, allowing scientists to modify rate parameters real time to the simulation and actively test alternate hypotheses. Such comparisons will greatly inform public health policy-makers.

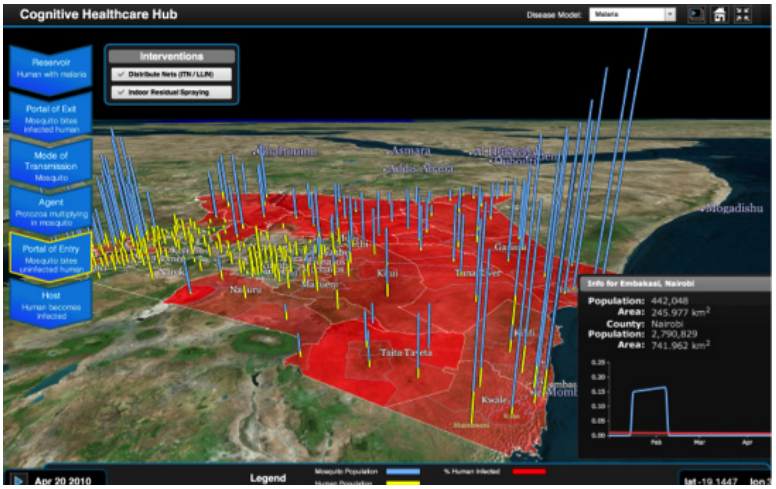



Figure 1. Visualisation portal displays simulated disease incidence in Kenya (no interventions are applied).

## CONCLUSION

In this study, we simulate disease incidence data for various scenarios using a mathematical model for analysing and visualising the infectious disease spread in human populations over time and over geographies. We have shown how leveraging information can be used to inform users on the prevalence of disease and the ways in which the diseases are best treated and prevented.

## REFERENCES


1. World Health Organization, Heslth Statistics and Informatics Department. Summary: Deaths by cause, in WHO Regions, estimates for 2008. 2011.
2. Cesium, Analytical Graphics Inc. http://cesiumjs.org/
3. D3, Data Driven Documents, Mike Bostock, http://d3js.org/
4. STEM, http://www.eclipse.org/stem
5. S. Edlund, M. Davis, J. Pieper, A. Kershenbaum, N. Waraporn, J. Kaufman, A global study of malaria climate susceptibility. Epidemics 3 (Boston, 2011).
6. G. Macdonald, The Epidemiology and Control of Malaria, London, Oxford University Press, 1957.


# Enhancing diagnostics for invasive Aspergillosis using machine learning

Simone Romano[a], James Bailey[a], Lawrence Cavedon[a,b], Orla Morrissey[c,d], Monica Slavin[e,f], Karin Verspoor[a,b]

[a]The University of Melbourne, Dept. of CIS
[b]NICTA (National ICT Aust.) VRL
[c]Alfred Health
[d]Monash University
[e]Peter MacCallum Cancer Centre
[f]Melbourne Health

## SUMMARY

Invasive Aspergillosis (IA) is a serious fungal infection and a major cause of mortality in patients undergoing allogeneic stem cell transplantation or chemotherapy for acute leukaemia[1]. The major contributing factor to the high mortality rates is that culture methods have limited sensitivity, only detecting 40-50% of IA cases. The currently accepted criteria used for diagnosing IA are CT scan findings, microbiology and risk factors. However, because of ease of use and improved sensitivity, biomarkers such as Aspergillus PCR and Galactomannan (GM) assays are used both increasingly and alternatively. These tests are performed at least twice weekly. Two consecutively positive PCR or GM results or greater than two intermittently positive PCR or GM results within a two-week timeframe is taken as an indicator of Probable IA irrespective of the results of other tests[4]. The frequent testing coupled with moderate specificity of these biomarkers can result in a number of false positive results. It can be difficult to ascertain whether an individual result is a true or false positive. Large amounts of data are collected during the treatment of high-risk haematology patients and we propose leveraging such data to produce more accurate predictions of IA diagnosis. We describe here the application of machine learning techniques to predict probability of IA, which can be used to enhance the interpretation of biomarker results. When the estimated probability of infection is low, we identify 26.5% PCR/GM positive tests in our data that did not lead to an IA diagnosis within a week (TNR = 28.9%, NPV=100%). For such cases, antifungal treatment may be safely avoided, minimising over-treatment and drug toxicity, and reducing associated costs.

## INTRODUCTION

Invasive Aspergillus (IA) has been associated with a 34-43% mortality rate[2] and a patient with IA incurs an added 7 days of hospital stay and extra $AU30,957 in hospital costs[3]. In a randomised controlled trial comparing different strategies for diagnosing IA[4], large amounts of data were collected from 240 patients undergoing allogeneic stem cell transplantation or chemotherapy for acute leukaemia between September 2005 and November 2009 at six Australian centres. All patients were tracked for 26 weeks from the beginning of their treatment, providing rich longitudinal data on daily and weekly tests for each patient. In total, the data consists of $240 \times 26 \times 7 = 43{,}680$ records, a large number that makes bed-side, interpretation a challenging task. In one strategy arm of the study twice weekly Aspergillus PCR and GM results were used to diagnose IA according to pre-defined criteria[4]. However, in some cases these biomarkers produced single positive results. Knowing that in these patients IA is associated with high mortality rates, a single positive result did trigger treatment with antifungal drugs which may have been unnecessary in some cases and may have resulted in avoidable toxicity and expense. We aim to aid the interpretation of a single positive result by providing information on the likelihood that it is a false positive by using machine learning techniques over the collected data to compute a value of infection probability when a single positive PCR/GM test is recorded.

## DESCRIPTION

We focus on discriminating between the positive PCR/GM tests that are associated with an IA diagnosis within a week and those that are not associated with an immediate IA diagnosis. When a single positive PCR/GM is detected, two types of data could be used to improve the reliability of prediction: data related to the subject characteristics (constant along the treatment) and data related to events occurring in the recent past. The latter includes results of daily tests (e.g. blood tests) as well as individualised treatment determined by the clinicians during the days preceding the single positive PCR/GM result. Due to the temporal aspect of the data, it is not possible to employ a simple statistical prediction model such as logistic regression. Rather, it is necessary to tailor a machine learning model that exploits the trajectories of values over time.

**Simone Romano**

PhD Student
University of Melbourne

sromano@student.unimelb.edu.au

Simone Romano is a PhD student at the Computing and Information System department at the University of Melbourne. His main interests are in data mining and machine learning applied to biomedical problems. He is currently working on prediction of invasive aspergillosis using classification techniques.

## METHODS

We used the Random Forests[5] machine learning algorithm for prediction due to its ability to cope with a large number of heterogeneous features. Our approach used both features that did not vary during the treatment, (e.g. gender, age, BMI, and smoking status) as well as features that varied over time, including neutrophil count, body temperature, corticosteroids doses, haemoglobin and platelet count. We generated the following features reflecting clinical intuitions about strategies for capturing the variation of data over time:

- Duration features: We counted the number of days the value each parameter lay within specified ranges. We limited our analysis to 30 days prior to a single positive PCR/GM reading, reflecting upper bound of the incubation time for IA[6]. We partitioned values for each parameter into percentiles or, for cases where we had specific knowledge about data type, we used pre-specified thresholds. For example, we divided temperature in 1 Celsius degree intervals, e.g. [36,37], (37,38] etc., and we counted number of days temperature occurred for a patient in each interval;
- Trajectories: In order to capture changes in the sampled value for a given parameter, we used the methodology proposed in[7]. We selected two days in the 3 week window preceding a single positive PCR/GM test and computed the mean value, the standard deviation, and the relative difference between those values. We performed this operation for each possible pair of days in the 3 week window. For example, we computed the mean value of temperature readings for the week prior to the single positive PCR/GM result, the standard deviation and the relative difference between the temperature measured one week prior to the positive test result and the day of the positive test result. We repeated this process for all possible intervals in the 3-week window. This technique has the capacity to capture subtle changes in the trajectories of values.

## RESULTS

Our training set was a collection of 358 single positive PCR/GM tests that precede the earliest label of IA infection according to either diagnostic strategy (standard culture-based or biomarker strategy). Respectively, 284 cases were related to positive PCR and 83 cases to positive GM tests. Just 29 of the positive PCR/GM were associated with a Proven IA or Probable IA label within a week according to either diagnostic test. Thus 329 results might be considered as false positives. We built a predictive model that exploits the data available at the occurrence of a positive PCR/GM to output a probability of infection within a week value. We validated this model by a patient-level cross-validation framework, where we took care not to use data relative to one patient to both model training and validation. Setting a low threshold on the model output probability to achieve high NPV (100%) we were able to identify 95 such tests that do not lead to an IA infection (TNR = 28.9%) within a week.

## CONCLUSIONS AND FUTURE WORK

The implemented predictive model seems to enable safe avoidance of antifungal therapy for a significant number of cases. It accurately discriminated between true and false single positive PCR/GM results. This tool can be an aid to further avoid over-treatment, reduce drug-toxicity, and reduce antifungal drug costs. Future work will aim to make the model more accurate in predicting when a positive PCR/GM is associated with an immediate infection to trigger the antifungal treatment earlier in time; search for alternative diagnosis when the outcomes are equally probable according to the model; and make the model output more interpretable to clinical practitioners, e.g. by identifying the trajectories in the data which generate a low or high probability of IA.

[†]NICTA is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program.

## REFERENCES

1. D. Neofytos, D. Horn, et al., Epidemiology and outcome of invasive fungal infection in adult hematopoietic stem cell transplant recipients: analysis of Multicenter Prospective Antifungal Therapy (PATH) Alliance registry, Clinical Infectious Diseases 48(3): 265-273, 2009.
2. WJ. Steinbach, K. A. Marr, et al., Clinical epidemiology of 960 patients with invasive aspergillosis from the PATH Alliance registry. J. Infection 65(5): 453—456, 2012.
3. MR. Ananda-Rajah, A. Cheng, et al., Attributable hospital cost and antifungal treatment of invasive fungal diseases in high-risk haematology patients: an economic modeling approach. Antimicrob Agents and Chemotherapy, 2011: 55(5): 1953-60.
4. C. O. Morrissey, SC. Chen, et al.,Galactomannan and PCR versus culture and histology for directing use of antifungal treatment for invasive aspergillosis in high-risk haematology patients: a randomised controlled trial, Lancet Infectious Diseases 13(6), 519-528, 2013.
5. L. Breiman, Random forests, Machine Learning 45(1), 5-32, 2001.
6. H. Deng, G. C. Runger, E. Tuv, M. Vladimir, A time series forest for classification and feature extraction, Information Sciences 239, pp. 142-153, 2013.
7. T. Benet, N. Voirin, et al., Estimation of the incubation period of invasive aspergillosis by survival models in acute myeloid leukaemia patients, Medical Mycology 51(2), 214-218, 2013.

**Prof Svetha Venkatesh**

Professor
Deakin University

svetha.venkatesh@deakin.edu.au

Prof Venkatesh is the chair in Complex Pattern Analysis at Deakin University and heads the strategic research centre for pattern recognition and data analytics. Venkatesh and her team have tackled problems in autism, security and aged care. The outcomes include three award winning start-up companies: Virtual Observer, iCetana and Tobyplaypad.

# HealthMap: A visual platform for patient suicide risk review

Santu Rana[a], Wei Luo[a], Truyen Tran[a], Dinh Phung[a], Svetha Venkatesh[a], Richard Harvey[b]

[a]Centre for Pattern Recognition and Data Analytics, Deakin University, Geelong, Australia
[b]Barwon Health, Geelong, Australia

## SUMMARY

Misjudging suicide risk can be fatal. Risk assessment is complicated by multiplicity of risk factors, none of which individually can reliably predict risk. This paper addresses the need for better clinical support, visualising risk factors scattered in raw electronic medical records. HealthMap is a visual tool that helps clinicians effectively examine patient histories during a suicide risk assessment. We characterise the information visualisation problems accompanying suicide risk assessments. A design driven by visualisation principles was implemented. The prototype was evaluated by clinicians and accepted into daily clinical work-flow.

## INTRODUCTION

"More British soldiers commit suicide than die in battle." "Suicide: Leading cause of Death among middle age Americans surpass car accidents, says CDC report". These are some recent headlines showing the severity of suicide in our modern society. Although most people have sought clinical help before committing suicide[1], the help they received is often insufficient. Intensive intervention, no matter how costly, is needed for the right people at the right time. To identify patients at risk, mental health practitioners rely on assessment organised through a list of questions covering major risk factors such as suicide attempts, suicide ideation, family history and sense of hopelessness[2]. Answers to the questions are based on interviews with the patient or the family members in combination with patient history. But the task of collecting relevant patient information is so complex that it is a "source of apprehension for clinicians"[3]. This complexity reflects the Variety dimension of big data. The bottle neck of retrieving patient information is, to a large degree, due to badly designed user interface. We need a computer tool for clinicians to quickly understand the psychosocial context and life experience of each patient[4]. In this talk, we present HealthMap, a visual tool supporting suicide risk assessment. We chart our journey from concept to acceptance of HealthMap in a large Australian hospital.

## METHODS

HealthMap (see Figure 1) exploits the routine information from Electronic Medical Records (EMR) to reconstruct a comprehensive picture of patients' mental history. By organising information around the patient, we facilitate optimal information retrieval. Coupled with machine learning predictions, the system presents data that address-"Is this patient is at risk?" and "Why?"

*Development and implementation*

HealthMap aims to present two types of information: the raw patient records and the summarised patient suicide risk based on machine prediction, which is generated through a data mining system previously validated and published in[5]. The design follows Sedlmair et al's 9-stage framework[6]. Tasks involved identification and subsequent filtering of Electronic Medical Records and other specialised databases to extract factors critical to suicide risk, prioritising to visually discriminate moderate and high risk elements, and finally presentation of that information during a risk assessment. We describe the interaction, feedback and evolution of the visualisation, conducted in 3 phases, with final adoption into clinical practice. Feedback collection is through presentation of the prototype at clinician meetings and following questionnaires.

In the final design, we organise raw patient data by their relevance to suicide risk. We borrow several ideas from[5]. They considered a wide range of variables that can be extracted from electronic medical records and found the most predictive variables are recent ED visits, recent admissions, and certain demographic information. We classify each patient event into three risk levels, in part through the ICD-10 codes generated from each event.

To achieve click-free navigation, we classified patient information into two tiers: event dates and risk categories are on the top tier; detailed diagnoses and clinical notes are on the bottom tier. Two tiers were mapped to two views: the left view shows temporal overview and enables event selection; the right view displays contextual information for the selected event. The layout/navigation design and visual encoding in HealthMap center around a two-tier information hierarchy that align with the mantra of "overview first; details on demand".

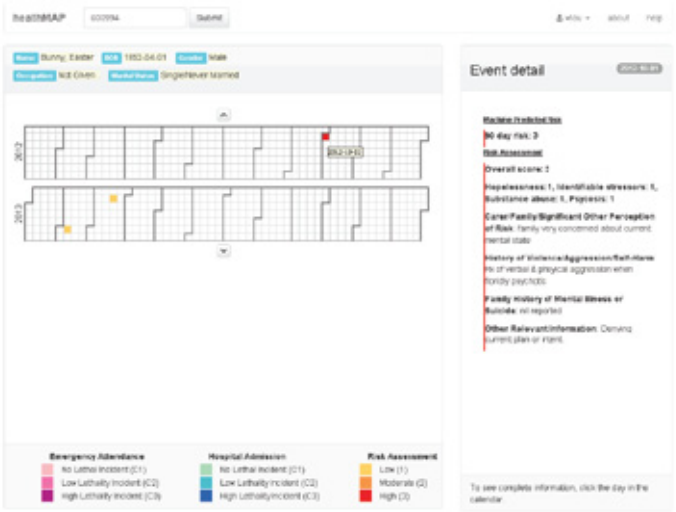



Figure 1. HealthMap, a visual tool for interactive overview of patient mental health history. Easter Bunny is a hypothetical patient. Machine prediction uses the same colour scheme as (clinician) risk assessment.

## RESULTS

HealthMap has been deployed in an Australian tertiary hospital after initial positive response from mental health clinicians. The main scientific contributions of this work include:

1. Characterisation of the information visualisation needs for suicide risk assessment, a daily task in mental healthcare. The social significance of improved suicide risk assessment is immense, as errors can be fatal.
2. Implementation of a visual tool to assist suicide risk assessment, iteratively refining the visualisation to fit the needs of clinicians.
3. Validation of the visual tool through clinical interactions and acceptance into clinical practice at a large Australian hospital. These factors are used as the basis to "gather a comprehensive picture of each individual patient"[7].

## CONCLUSIONS

We describe the visual design challenges in producing HealthMap, a supporting tool for suicide risk assessment. We present the journey from problem identification to specification and visual design, and finally, iterative refinement of prototypes. Our final prototype was accepted for adoption in clinical practice. Our experience yields interesting lessons for utilising EMR more effectively, and for introducing machine predictions into clinical practice.

[1]The Telegraph, 14/07/2013.

[2]International Business Times, 06/05/2013.

## REFERENCES


1. Luoma JB, Martin CE, Pearson JL. Contact with mental health and primary care providers before suicide: a review of the evidence. Am J Psychiatry 2002;159(6):909-16
2. Hawton K. Assessment of suicide risk. The British journal of psychiatry : the journal of mental science 1987;150:145-53
3. Hughes CW. Objective assessment of suicide risk: significant improvements in assessment, classification, and prediction. Am J Psychiatry 2011;168(12):1233-4 doi: 10.1176/appi.ajp.2011.11091362[published Online First: Epub Date]|.
4. Chehil S, Kutcher SP. Suicide risk management: a manual for health professionals: Wiley. com, 2012.
5. Tran T, Phung D, Luo W, et al. An integrated framework for suicide risk prediction. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. Chicago, Illinois, USA: ACM, 2013:1410-18.
6. Sedlmair M, Meyer M, Munzner T. Design study methodology: reflections from the trenches and the stacks. Visualization and Computer Graphics, IEEE Transactions on 2012;18(12):2431-40
7. Ryan CJ, Large MM. Suicide risk assessment: where are we now? The Medical journal of Australia 2013;198(9):462-3


## Dr Chen Wang

Senior Research Scientist
CSIRO Computational Informatics

chen.wang@csiro.au


Dr Chen Wang is a Senior Research Scientist at CSIRO Computational Informatics. He received his PhD from Nanjing University. His research interests are primarily in distributed, parallel and trustworthy systems. His current work focus on data analytics systems for drug adverse reaction discovery. His recent work include accountable distributed systems and cloud computing. He is also an Honorary Associate of the School of Information Technologies at the University of Sydney. Dr Chen Wang has industrial experience. He developed a high-throughput event delivery system and a medical image archive system, which are used by many hospitals and medical centres in USA.

# Causality driven data integration for adverse drug reaction discovery


Chen Wang, Sarvnaz Karmi

CSIRO Computational Informatics



### SUMMARY

We describe an ongoing effort in CSIRO for partially automating causality discovery in the Adverse Drug Reaction (ADR) detection process. The proposed method integrates data from multiple sources based on rules that indicate causality.


### INTRODUCTION

Drug adverse reactions are a major threat to public health and impose huge costs to healthcare systems. Postmarketing surveillance aims to reduce the effects of adverse drug events. There are two types of ADR discovery systems operating around the globe: passive discovery and active discovery. They differ in the data used for detecting unexpected harm caused by the normal use of a drug at the normal dosage as per label or prescription. Passive ADR discovery, which has been established for decades, uses reports that are voluntarily submitted by pharmaceutical companies, healthcare professionals and consumers. Regulatory bodies, such as TGA (Therapeutic Goods Administration) and FDA (Food and Drug Administration), maintain very large databases of such reports which they use to mine the potential safety signals. More recently, active discovery has been introduced. Active discovery monitors healthcare data from a variety of sources such as electronic health records, health insurance claims, medical literature, or even recently medical forums to identify potential signals automatically using text and data mining techniques. Active discovery intends to discover unexpected adverse events as early as possible and is therefore also called near real-time drug safety surveillance. An example of such system is recently proposed in FDA Sentinel initiative which relies on sharing deidentified patient data among a number of organisations.

Both types of ADR discoveries ultimately lead to establishing the causal relationship between a drug and unexpected adverse reactions. Often ADR discovery starts with data-mining techniques for disproportionality detection of the reports about a drug and an adverse reaction in comparison to other pairs of drugs and adverse reactions. These potential ADRs are then examined in medication safety review and assessment meetings. The main task of these meetings is to establish the causality between a drug and its adverse reactions. This is largely a manual process in the current practice and often generates wide variability in assessment[1,2,3]. Even though the shortcomings of the current process were recognised in 70s[1], there is not much improvement in the practice of establishing causality between a drug and an ADR so far. As ADR related data become increasingly accessible in electronic format and with the increase in processing power and techniques of dealing with big data, it is now possible to introduce carefully designed algorithms to assist the causality reasoning process and therefore automate some of the manual steps in this assessment to reduce variability. We note that the current process, endorsed by WHO, is still largely based on Naranjo's questionnaire[1] designed in 1981. To achieve this, there are two major requirements: first, it is essential to understand and capture the reasoning process in the existing practice. A good reasoning process tends to minimise the variability and inconsistency in assessment as shown in[1]. Second, integrating data from various sources is essential for reaching correct conclusions in the reasoning process, e.g. additional data about background of the patients in ADR reports may help to identify causes of an ADR. This is of course only possible with collaboration of multiple health agencies to make such data accessible. Below, we propose a causality detection method to address these requirements.

### DESCRIPTION

Previous work trying to establish causality between a drug and its unexpected adverse reactions used a well designed questionnaire to guide the assessment process[1]. The answers of these questions were assigned different scores and the total score of each rater determines the certainty of the rater on whether a drug D causes a reaction R. A consensus among raters served as an indicator of the causality of D and R.

Our proposed method contains two steps: (1) Design rules to capture the causality reasoning process using domain-expertise and the current known knowledge of ADRs per each drug or active ingredient; and (2) Process different data sources based on these rules to establish if a given drug D causes a specified adverse reaction R.

A starting point for rule identification is using the existing questionnaires, and also formalising the reasoning process within the review and assessment teams inside the regulatory. For instance, consider a specific drug D and its possible adverse reaction R and a given dataset S (e.g, electronic health records and clinical notes). The following rules could be considered for causality discovery:

1. Discontinued D, R still existed;
2. Discontinued D, R improved;
3. Readministered D, R reappeared;
4. Increased the dose of D, R became more severe;
5. Decreased the dose of D, R became less severe;
6. Factors F1, F2 and F3 cause R.

These rules capture common reasoning used in identifying whether D causes R. The set of rules are extensible. With these rules defined, the next step is to process data based on these rules to discover causality between a drug and a given adverse reaction. In order to achieve this, we first build a data model that contains necessary data fields required by these rules. For example, to support the rules above, we need information about actions taken on a drug by a consumer, or instructions of a medical professional to the patient, such as the discontinuing its use, changing its dose etc. as well as additional information about other factors that may cause the adverse reaction. After the rule list is completed, a table is constructed to capture the data model. See Table 1 for an example. The headers of Column 2 to Column 6 show a sample data model. Afterwards, we process each data source using information extraction techniques and assisted by medical ontologies and drug knowledge repositories to populate the table. The last column "D causes R" in Table 1 represents the decisions and is partially populated via existing knowledge.

Table 1. Summary of reports regarding drug D and suspected adverse reaction R

| REPORT | D DISCONTINUED | D READMINISTRATERED | DOSE CHANGE | OTHER FACTORS | R CHANGE | D CAUSES R |
|---|---|---|---|---|---|---|
| 1 | Yes | N/A | No | None | Improved | Yes |
| 2 | Yes | N/A | No | Unknown | No improvement | Unknown |
| 3 | Yes | N/A | No | F2 | No improvement | No |
| 4 | No | No | Decreased | F3 | Improved | Yes |
| 5 | No | No | Decreased | Unknown | No improvement | Unknown |

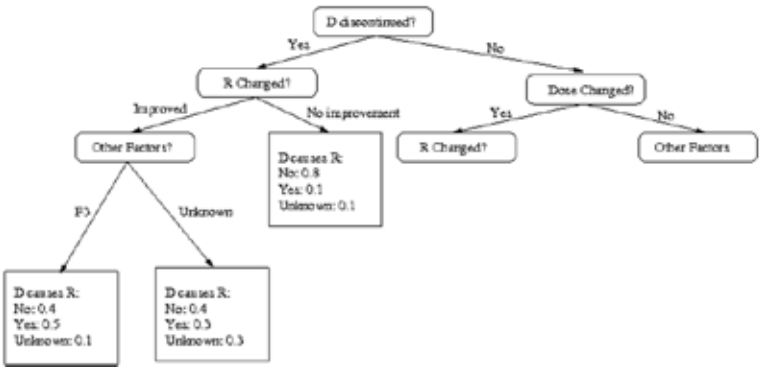



Figure 1. A sample decision tree for ADR causality discovery

Based on the table, we use a decision tree to classify the data. Human annotated data are used as the training set. A sample decision tree classifier is shown in Figure 1. Note that this tree only partially covers Table 1. The final decisions on causality (Yes, No, or Unknown) will be based on a threshold on the probabilities generated by decision. The decision tree evolves as the number of confirmed causality pairs increases. As the data model is independent of underlying data sources, our method is capable of dealing with multiple data sources as long as they contain at least some of information needed by the data model.

## CONCLUSION

Causality discovery is essential to detect potential adverse drug reactions. However, the implementation challenges are extracting high quality causality information from a variety of data and dealing with different level of credibility of information from different data sources.

## REFERENCES


1. C. Naranjo, U. Busto, E. Sellers, P. Sandor, I. Ruiz, E. Roberts, E. Janecek, C. Domecq, and D. Greenblatt. A method for estimating the probability of adverse drug reactions. Clinical Pharmacology and Therapeutics, 30:239–245, 1981.
2. R. P. Naidu. Causality assessment: A brief insight into practices in pharmaceutical industry. Perspect Clin Res 2013;4:233-6.
3. N. Anderson, J. Borlak. Correlation versus causation? Pharmacovigilance of the analgesic flupirtine exemplifies the need for refined spontaneous ADR reporting. PLoS One. 2011;6(10):e25221

australia's premier e-health conference

# melbourne

11 - 14 august

# hic 2014

## Investing in e-health:
## People, knowledge and technology for a healthy future

REGINA HOLLIDAY
Patient Artist Advocate, The Walking Gallery

DAVE DE BRONKART
@ePatientDave

DR DANNY SANDS
Co-Founder and Past President,
Society for Participatory Medicine

"Let patients help!
We're the most underused
resource in all health and care."

*Dave de Bronkart*

hisa

"Patient and provider
engagement and collaboration
are critical to increasing the
effectiveness of healthcare."

*Dr Danny Sands*

hisa.org.au/hic2014

Health Informatics Society of Australia (HISA)
1A / 21 Vale Street
North Melbourne VIC 3051
t: 03 9326 3311
e: bigdata@hisa.org.au
w: hisa.org.au

ISBN
978-0-9751013-2-2

hisa.org.au/bigdata2014