THE WINSTON CHURCHILL MEMORIAL TRUST OF AUSTRALIA

Report by  - LOUIS WANG  - 2015 Churchill Fellow

THE DR DOROTHEA SANDARS CHURCHILL FELLOWSHIP
to explore optimal systems and partnerships
to foster a local and national cardiac genomic network

Vanderbilt University Medical Center &
eMERGE (electronic medical records and genomics) Network
Nashville, TN & Cincinnati, OH, USA

I understand that the Churchill Trust may publish this Report, either in hard copy or on the internet or both, and consent to such publication.

I indemnify the Churchill Trust against any loss, costs or damages it may suffer arising out of any claim or proceedings made against the Trust in respect of or arising out of the publication of any Report submitted to the Trust and which the Trust places on a website for access over the internet.

I also warrant that my Final Report is original and does not infringe the copyright of any person, or contain anything which is, or the incorporation of which into the Final Report is, actionable for defamation, a breach of any privacy law or obligation, breach of confidence, contempt of court, passing-off or contravention of any other private right or of any law.

Signed    Louis Wang

Dated    2 October 2018
Key words

Genomic research
Cardiovascular
De-identification
Privacy
Electronic health record
Derived health record
Biobank
Network
Personalised medicine
Index

Indemnity clause .................................................................................................................. 1
Key words ............................................................................................................................. 2
Index ..................................................................................................................................... 3
Acknowledgements ............................................................................................................... 4
Executive Summary ............................................................................................................. 5
Programme ............................................................................................................................ 6
Key personnel met during the Fellowship .......................................................................... 6
The Vanderbilt Experience ................................................................................................. 7
The Vanderbilt BioVU .......................................................................................................... 7
The Synthetic Derivative at Vanderbilt University: a revolutionary and innovative example
of a derived and de-identified electronic health record ..................................................... 8
Lessons for an Australian environment ............................................................................. 9
Effective de-identification ................................................................................................. 9
Alteration of each “real” health record so that it becomes a “derived” patient record...... 9
Random exclusion of a percentage of records ................................................................ 9
Imposing a minimum dataset requirement for research. ................................................... 9
Advantages of a “derived” health record system ............................................................... 10
Disadvantages of a “derived” health record system .......................................................... 10
The rationale, logistics, and acceptability of an opt-out model ......................................... 10
The eMERGE (electronic medical records and genomics) Network ................................ 11
Organisation ....................................................................................................................... 11
Key observations: fostering a culture of collaboration ....................................................... 12
Ethical, legal and social issues in genomic research .......................................................... 12
Conclusions and Recommendations ............................................................................... 16
Local .................................................................................................................................... 16
National ............................................................................................................................... 16
Potential applicability to the My Health Record system .................................................. 17
References ........................................................................................................................... 19

Appendix: “The need for an Australian Derived Health Record: a practical solution to privacy
concerns surrounding secondary use of My Health Data.” Submission to the Senate
Community Affairs References Committee inquiry into the My Health Records
Amendment (Strengthening Privacy) Bill 2018 – 18/105 ..................................................... 20
Acknowledgements

I sincerely thank the Winston Churchill Memorial Trust and my sponsor, Dr Dorothea Sandars, for the Dr Dorothea Sandars Churchill Fellowship which enabled me to travel to Vanderbilt University Medical Center. I am grateful to Professor Diane Fatkin (Victor Chang Cardiac Research Institute) and Professor Anne Keogh (St Vincent’s Hospital) for encouraging me to apply for this program.

I especially thank Professor Dan Roden for his generous invitation to be based in his laboratory, and, in addition to meeting key personnel involved in the creation of the Vanderbilt BioVU and the eMERGE network, to also observe and participate in research activities resulting from the establishment of the Vanderbilt BioVU. This helped provide for me not only an appreciation of the workings of a cardiac genomic network (and a genomic network, in general) in the United States, but also offered insights into the potential downstream research and clinical applications that such a facility can provide. I also would like to thank Professor Ellen Clayton (Rosalind E. Franklin Professor of Genetics and Professor of Law), as well as members of the Dan Roden’s laboratory for their advice and hospitality. I also thank Dr John Harley and my hosts at the eMERGE for inviting me to be part of the eMERGE Summer Steering Committee meeting.

Where appropriate in this report, I have referenced published journal articles that are freely and publicly available via PubMed.gov (US National Library of Medicine, National Institutes of Health). These documents are available in the public domain for all parties interested in additional information regarding the design and implementation of the Vanderbilt Synthetic Derivative and BioVU, and how these programs addressed issues relating to oversight, community engagement and privacy.
Executive Summary

This Dr Dorothea Sandars Churchill Fellowship was undertaken at:

- Vanderbilt University Medical Center
  Nashville, Tennessee, USA
- The eMERGE (Electronic Medical Records and Genomics) Network
  Cincinnati, Ohio, USA

This fellowship, under the supervision of Professor Dan Roden at Vanderbilt University Medical Center, involved a short-term postdoctoral appointment and experience at a leading international genomic bank, exploring optimal infrastructure, information management systems, collection & storage practices, as well as infrastructure and systems aimed at promoting collaborative partnerships across other centres in Australia in the future. As the ethical, legal and societal issues (ELSI) associated with genomic research dictate the boundaries of what is permissible and achievable in this growing field, I explored in detail best practices in upholding highest standards of ethical genomic research with Professor Ellen Clayton and other key thought leaders in the United States. Collectively, this fellowship will lead to the dissemination of sustainable and ethical research practices, facilitate future compatibility of our information databases with major international sites, and promote collaboration between Australian and international researchers.

Arguably, however, the most useful and translatable aspect of information noted during this fellowship journey is the idea of developing a Derived Health Record, and its potential application to the Australian My Health Record system. This offers an innovative, potential solution to community concerns regarding researchers gaining access to data from the My Health Record system for approved secondary use.

As part of dissemination and implementation of these important concepts, and in order that this fellowship help act as a catalyst for the brave new world that is 21st century Australian big-data research, I have included this proposal in a submission to the Senate Community Affairs References Committee Inquiry into the My Health Records Amendment (Strengthening Privacy) Bill 2018 – 18/105. The submission is titled: "The need for an Australian derived Health Record: a practical solution to privacy concerns surrounding secondary use of My Health Data". The full submission has been included in the Appendix of this report.

The establishment of a nationwide, de-identified electronic health database that assures "privacy by design" is perhaps the single most important investment and safeguard for future Australian health research. Implementing a Derived Health Record that mirrors and coexists with the My Health Record has already been shown to be a practical and proven solution that leverages the maximum benefits that the systematic collection of electronic medical records can offer the community while protecting privacy concerns at an individual level. I look forward to discussing these ideas and working with all interested parties, organisations and government agencies in order to help bring this vision to life.

Dr Louis W. Wang
2015 Australian Churchill Fellow, and
Postdoctoral Fellow at Vanderbilt University Medical Center (Roden Lab, 2018)
St Vincent’s Clinical School, UNSW Sydney
Email: louis.wang@unsw.edu.au
Phone: 02 8382 1111
Programme

The time-line of the activities undertaken during the Churchill Fellowship were as follows:

- June 13 to July 25: Visiting Postdoctoral Research Fellow at Vanderbilt University Medical Center (Roden Laboratory), Nashville, TN
- June 22 to 24: Official observer at the eMERGE (Electronic Medical Records and Genomics) Network Summer Steering meeting at Cincinnati, Ohio.

Key personnel met during the Fellowship

- **Dan Roden MD CM**: Professor Roden is the Senior Vice President for Personalized Medicine at Vanderbilt University. A clinical cardiologist who works on the Arrhythmia Service at Vanderbilt University, and with a large research program in cardiac pharmacology, he studies cardiac electrophysiology, drug interactions and has pioneered the idea of achieving personalised medicine by incorporating genomic information into clinical decision making tools (1). He is a founder of the BioVU biobank at Vanderbilt University.
- **Ellen Clayton MD(Harvard) JD(Yale)**: Professor Ellen Clayton is the Rosalind E. Franklin Professor of Genetics and Professor of Law at Vanderbilt University. She is an internationally respected leader in the fields of law and genetics. Her research interests include the ethical, legal, and social issues in genomic research. Professor Clayton is a key thought leader in this field, and was the author of a major review article in the *New England Journal of Medicine* titled: ‘Ethical, legal and societal implications of genomic medicine’ (2).
- **Lisa Bastarache MS**: Manager, Application Development and Lead Data Scientist, Vanderbilt University Medical Center. A key member of the Vanderbilt program and with extensive experience in mathematics, computational science and computation linguistics, Ms Bastarache was responsible for the development and implementation of new Phenome-wide Association Studies (PHEWAS) and phenotyping methods.
- **John Harley MD PhD**: Principal Investigator for Cincinnati Children’s Hospital, eMERGE Network. Professor Harley was the local host for the eMERGE Network meeting in Cincinnati.
- **Richard Gibbs, AO PhD**: An Australian geneticist now based in Baylor College of Medicine, Houston TX. Professor Gibbs is the Founder and director of the Human Genome Sequencing Center established at Baylor College of Medicine.
- **Josh Denny MD MS**: Professor of Biomedical Informatics, VUMC and Principal Investigator for Vanderbilt sites in the eMERGE Network, Pharmacogenomics Research Network, and the Implementing Genomics Into Practice (IGNITE) Network. Professor Denny is a key member of the PREDICT (Pharmacogenomic Resource for Enhanced Decisions in Care and Treatment) program at Vanderbilt, which prospectively genotypes patients to tailor drug response.
- **Rong Ling MD PhD MPH**: Epidemiologist and Program Director, National Human Genome Research Institute & eMERGE, Bethesda, Maryland.
- **Javid Moslehi MD**: Clinical cardiologist and basic/translational biologist whose key area of research is studying cardiovascular complications associated with novel molecular targeted cancer therapies. Associate Professor Moslehi’s group uses Vanderbilt BioVU to explore answers and questions arising from clinical medicine.
- **Joe-Elie Salem MD PhD**: A French cardiologist and postdoctoral research fellow who works closely with Javid Moslehi and Dan Roden, and who has been very successful in using data from BioVU, the Vanderbilt synthetic derivative and eMERGE in answering clinically relevant questions.
The Vanderbilt Experience

I was privileged to have the opportunity to spend the six weeks of my Churchill Fellowship at Vanderbilt University Medical Center in Nashville, Tennessee (TN), where I worked with Professor Dan Roden MD CM, the Senior Vice-President for Personalized Medicine and Founder of BioVU, the internationally renowned and pioneering biobank of Vanderbilt University Medical Center (VUMC). Professor Roden is also a Principal Investigator for the eMERGE (Electronic Medical Records and Genomics) Network in the United States.

Genetic variation plays a large role in human susceptibility to disease and variation in treatment response, not only in the field of cardiovascular medicine, but in many other health disciplines. With the advent of next generation sequencing and whole genome sequencing, genetic testing is becoming more reliable, affordable and comprehensive in its coverage. This, together with advances in information technology systems that enable the systematic collection and electronic storage of multi-dimensional data (e.g. diagnoses, clinical results, prognostic information, and genomic data), has enormous potential benefits for both clinicians and health researchers.

Bringing about a local cardiac genomic bank requires the confluence of multiple factors: (1) a quaternary level referral centre for cardiovascular care, which would therefore enable to steady recruitment of patients; (2) a closely affiliated research centre that is actively involved in clinical and basic science research; (3) secure funding and established infrastructure to ensure the viability of a long term project, and (4) a team of highly motivated individuals with long-term clinical and research appointments. In Sydney, St Vincent's Hospital fits the first criteria as it is the quaternary heart failure referral centre and is home to the cardiopulmonary transplant program in New South Wales. Also located in the St Vincent's Hospital precinct is the affiliated Victor Chang Cardiac Research Institute, an established research institute housing over 300 scientists that focusses heavily in basic cardiovascular research, and the Kinghorn Centre for Clinical Genomics, a collaborative research centre that aims to use genomics to improve precision healthcare.

The Vanderbilt BioVU

Vanderbilt University Medical Center is located in Nashville, Tennessee, USA. This health system services Middle Tennessee and is home to BioVU, a large-scale de-identified DNA biobank, which was created to facilitate big-data epidemiological and genomic research. This pioneering system has been in operation since 2007. The BioVU uses genomic information obtained from discarded blood specimens left over from clinical care and linked with the electronic medical record. The facility collects 500 samples per week, and currently holds over 250,000 DNA samples (3).

The underlying principles crucial to the success of genomic research that aims at identifying gene-phenotype associations are to have 1) a very large number of specimens obtained from as diverse a cohort as possible, and 2) have this genomic information linked to phenotypic database that is both accurate and rich with clinical information.

The Vanderbilt BioVU helped achieve this by the following means. I have adapted the following points from a freely available publication which published this institution’s methodology (4):

1. **DNA specimens are obtained from discarded blood left over after routine clinical care**: The fact that genomic data could be also collected without requiring additional blood collection helped increase convenience and participation, while reducing waste and manpower costs associated with collection of additional specimens. It also allowed the program to be conducted at a large scale.
2. **Phenotypic data was obtained from the electronic health record**: Prognostic information is very useful in determining whether certain factors (e.g. genetic, lifestyle, biomarkers) are related to adverse or favourable outcome. Longitudinal patient information is usually obtained through cohort studies. These, however, are expensive and very labour intensive to maintain with regards to the personnel and research hours required to update clinical records. This also has additional risks in loss to follow up. Phenotypic data, a term that is used to describe the set of observable characteristics of a person resulting from the interaction of an individual’s genes and the environment, can be obtained from electronic health records. In other centres, however, the only electronic health data available may be hospital administrative data, which is defined as data derived from the operation of administrative systems. Health care administrative data is generated at every encounter with the health care system, ranging from a clinical consultation, diagnostic test or an admission to hospital. Although this data was initially designed to reimburse health care services and to evaluate health utilisation by various organisations, it is increasingly being used for epidemiological, effectiveness, and safety outcomes research. Although the accuracy of this data has been subject to debate, certain outcome measures (e.g. death, hospitalisation) are quite accurate.

3. **Each biospecimen is linked to a fully de-identified and altered electronic health record.** This was achieved through the establishment of the Synthetic Derivative, a de-identified mirror image of the Vanderbilt Medical Center Electronic Medical Record, which will be discussed in detail below.

**The Synthetic Derivative at Vanderbilt University: a revolutionary and innovative example of a derived and de-identified electronic health record**

A key component of the program is the creation of a “mirror image” of the electronic medical record (4). This database, known as the Synthetic Derivative, contains clinical information derived from the Vanderbilt Health electronic health record system. It contains patient records that have been altered with respect to dates and devoid of identifying details, such that the altered (“synthetic”) record only resembles the original record in terms of diagnoses, test results and prognostic information. The Synthetic Derivative continually accrues new clinical data as they occur over time. This system presents an innovative solution to de-identification and privacy concerns. As it contains synthetic patient records derived from actual patients, this database no longer contains real patient records, representing, in effect, a similar population in an alternate reality. At Vanderbilt University Medical Center, the Synthetic Derivative can be used by researchers as a stand-alone resource or to link clinical information with genomic data from the BioVU biobank. The Vanderbilt Synthetic Derivative contains data from over 2.2 million people. Its search interface allows users to input clinical and demographic information (e.g. diagnoses, procedure codes, medications, laboratory test values, age and gender) and returns de-identified data. The clinical information system that is used at Vanderbilt is a web-based with very strict security protocols. This system is fully compliant with the Security and Privacy Rules within the Health Insurance Portability and Accountability Act (HIPAA) 1996 (4), and operates under extensive institutional review board oversight. Since each record in this Derived Health Record is derived from real-life data (diagnoses, test results, procedure codes, outcome data) from actual people, inferences from analyses performed on the Synthetic Derivative are therefore applicable and generalisable to the Vanderbilt population. The Synthetic Derivative enables the study of disease-associations, patterns of disease incidence and prevalence, and provide insights into trends in healthcare access and resource uptake. The availability of this alternate database will reduce the need for people not involved in an individual’s health care to access that person’s health data, and provides an additional layer of protection to privacy.
Lessons for an Australian environment

Effective de-identification

The United States has a Safe Harbor standard under the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule which requires the exclusion or removal of 18 identifiers from a de-identified dataset (4, 5). At Vanderbilt University Medical Center, commercially available software (DE-ID, DE-ID Data Corp) was used to remove these personal identifiers from medical records.

Each medical record number in the true electronic health record is recoded by a secure hash algorithm (SHA-512, National Security Administration) to produce a Research Unique Identifier (RUI), a 128-character code that is unique to a particular medical record number. The nature of the secure hash algorithm is such that it is not possible to re-compute the original medical record number from the derived RUI.

The following identifying features of each health record should be permanently removed on entry into the de-identified database:
- Name(s) and aliases;
- Date of birth;
- Place of residence;
- Medical record number;
- Medicare and insurance card numbers;
- Contact numbers, email addresses and social media accounts;
- Vehicle identifiers and serial numbers, including license plate numbers;
- Biometric identifiers, including finger and voice prints, and
- Identifiable photographic or clinical images.

Alteration of each "real" health record so that it becomes a "derived" patient record

The exact dates of clinical episodes and other key personal events are altered. In the Vanderbilt Synthetic Derivative, all dates in the synthetic record are shifted 1 to 364 days into the past (4). The shift is different across different individual records. However, within each individual record in the Synthetic Derivative, the time-shift of events is consistent throughout the entire record (for all past and future events), thereby allowing accurate longitudinal follow-up.

Random exclusion of a percentage of records

This technique has been used in the Vanderbilt Synthetic Derivative (6). Random exclusion of a percentage of records from the Synthetic Derivative makes it impossible to know with certainty whether a particular individual is present in the database. This provides an additional layer of privacy by increasing the difficulty in being able to infer the presence of any particular individual in a dataset.

Imposing a minimum dataset requirement for research.

Inadvertent re-identification is easier when there are fewer numbers of subjects. To reduce the potential for re-identification of included subjects when the number of patients within a particular research project are small, the search program only returns a data set to an approved investigator if the number of matching records generated by a particular research question exceeds a specified minimum number (6).
Advantages of a “derived” health record system

One of the key benefits of using a permanently de-identified database is that patients cannot be re-identified. This protects privacy and confidentiality, especially in an era where multiple researchers from many different research teams may potentially gain access to data. The establishment of the Synthetic Derivative protects patient confidentiality as there is less need for researchers to view identified patient records and provides an alternative data resource for researchers when having identifiable data is not essential for the conduct of research (6). Studies that require time-dependent data (e.g., prognostic studies) will still be possible under the proposed system, as the random time-altering function will be uniformly applied to all events within an individual record, so that exact follow up time after a sentinel event or exposure can be obtained.

Disadvantages of a “derived” health record system

The time-shifting algorithm used in preserving anonymity in the Synthetic Derivative means that events tied to specific dates (e.g., studies of effects of natural disasters, policy changes, and other important events) cannot be evaluated with this database (4). These projects, instead, will require access to data from the true research database. Nevertheless, with the availability of the Synthetic Derivative, the vast majority of research projects can be completed without ever needing to access direct data from the Research Database.

For research using the Synthetic Derivative, contact with individual patients is not possible, and also likely to be prohibited. The design of the Synthetic Derivative means that results can only be inferred at a population level, and that there is no means of returning a significant result to an individual as the de-identification process is permanent. The lack of any identifiers means that it would also be impossible to contact any individual for further information, and only the information other already present within the derived health record will be available for the purposes of research. As an example, if specific biomarkers are found to be strongly associated with a disease in a permanently de-identified dataset obtained from the Synthetic Derivative, there will be no means of returning these results to the patients who originally contributed the data. Studies that aim to provide direct benefits for individual patients cannot be answered using the Synthetic Derivative, and should undergo the traditional pathway of formal, opt-in, informed consent. As per existing regulations, the informed consent process for these studies should explicitly ask the patients to nominate whether or not they wish to be contacted regarding any significant or incidental findings generated from their participation in the study.

The rationale, logistics, and acceptability of an opt-out model

Two designs were initially considered, a conventional “opt-in” (consent) model, and an “opt-out” model. The opt-in approach has usually been used for prospective studies focusing on specific diseases or therapies. Recruitment is resource-intensive and slower than in opt-out methods. An opt-out approach has an advantage of being able to be applied at a large-scale. This can result in the recruitment of a large general population and the generation of very large datasets not limited by prospectively designed research questions. An opt-out consent model was originally adopted at Vanderbilt University Medical Center. For an opt-out system to be ethical, there needs to be proper amount of opportunity for potential participants to get information on the program, and plenty of opportunity to opt-out if they desired. Much of the original information was provided in the form of posters and pamphlets throughout the hospital, clinics and pathology collection areas (6). Surveys were conducted to ensure that the information reached the target audience. 90% of surveyed participants were comfortable with the idea of de-identified genetic data being used for research, while only a small number (~5%) dissented (7). The program observed an opt-out rate of ~2.5% of patients (4).
The eMERGE (electronic medical records and genomics) Network

As part of the experience of seeing how a multi-centre genomic network could function as well as its collaborative potential, I visited the eMERGE steering committee summer meeting. The eMERGE (electronic Medical Records and Genomic Research) network is a national genomic research network in the United States that is run and funded by the National Human Genome Research Institute (NHGRI). The network consists of several DNA biorepositories that combine genomic data with electronic health record systems (8). Unlike Vanderbilt BioVU, which was an opt-in system, patients enrolled in programs that are part of the eMERGE consortium are consented for broad research use, and strict regulations and certification is required for enhancing privacy of shared electronic health data.

One way this network functions is that if one group of researchers at one of the member sites finds that a particular genetic trait is associated with a particular observation, this can be re-tested (i.e. validated) in another member site (or multiple member sites) across the network. As validation of observations is important to further reinforce the validity as well as generalisability of any given observation, having the ability to test the hypothesis on multiple sites, or to increase the number of cases and controls through amalgamation or inclusion of other cohorts across the network is a powerful tool. The networks use similar software programs, to facilitate data exchange and testing.

Organisation

The eMERGE Network is comprised of the following individual sites, each of which having its own biobank which is linked to phenotypic data contained within an electronic health record:

- Children’s Hospital of Pennsylvania
- Cincinnati Children’s Medical Center
- Columbia University
- Kaiser Permanente Washington with University of Washington and the Fred Hutchison Cancer Research Center
- Harvard University
- Mayo Clinic
- Meharry Medical College
- Northwestern University
- Vanderbilt University (Coordinating Center)
- Baylor College of Medicine (central sequencing and genomic centre)
- Partners/Broad (central sequencing and genomic centre)

I met principal investigators and researchers from several facilities across the network (listed in Table 1 below). These differed in the types of patients that contributed samples, phenotypes studied as well as the biobank design:

Table 1: Several member sites within the eMERGE network (adapted from (9))

<table>
<thead>
<tr>
<th>Site</th>
<th>Biobank design</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marshfield</td>
<td>Population-based</td>
<td>20,000</td>
</tr>
<tr>
<td>Mayo Clinic</td>
<td>Study of peripheral artery disease. Cases identified from the Mayo non-invasive vascular laboratory database; control subjects identified from the Cardiovascular Health Clinic</td>
<td>2,000</td>
</tr>
<tr>
<td>Northwestern</td>
<td>Outpatient clinic and hospital-based population of patients with Type 2 Diabetes Mellitus</td>
<td>10,000</td>
</tr>
<tr>
<td>Group Health</td>
<td>Cohort study investigating Alzheimer’s Disease and dementia</td>
<td>4,000</td>
</tr>
<tr>
<td>Vanderbilt</td>
<td>Use of discarded blood/non-human subjects linked to electronic medical records</td>
<td>75,000</td>
</tr>
</tbody>
</table>
Key observations: fostering a culture of collaboration

To facilitate collaboration, external institutions may apply for affiliate membership to the eMERGE Network. Collaboration is encouraged by fostering collegiality and promoted by the following measures:

• Regular steering committee meetings
• Collaboration in the formulation, design, conduct of research
• Collaboration in troubleshooting problems related to bioinformatics, phenotyping, natural language processing, ethics, legal and social issues. Many of these involve committees consisting of members from various member sites.
• Sharing of new technologies that improve phenotyping or data interpretation.
• Generosity of authorship across institutions in downstream publications.
• Social events at the network steering committees. These allow members to meet each other outside of a formal business meeting, and provides an important setting for informal discussions relating to projects.
• Creation of membership: Information about affiliate membership such as benefits, criteria for participation, and application process can be found at the eMERGE webpage.

Ethical, legal and social issues in genomic research

During my fellowship, and through my discussions with key thought leaders both at Vanderbilt University Medical Center and at eMERGE, I gained further understanding of the deep ethical, legal and social issues that underpin the creation as well as ongoing success of these biobank repositories. Such is the important of the ethical, legal and social issues (termed ELSI) in genomic research that this network has a committee that focuses specifically on the oversight, troubleshooting, policy-making and performs research on these issues (9).

For the successful formation of any repository as well as network, many issues need to be considered (10). Over the next three pages, I have tabulated the issues, the experiences at Vanderbilt and other eMERGE sites, as well as the implications for research performed in the Australian context (Table 2).
<table>
<thead>
<tr>
<th>Issue</th>
<th>Vanderbilt / eMERGE experience</th>
<th>Applicability to Australia</th>
</tr>
</thead>
</table>
| **Data Access**
Who is going to get access to the research data? | In the United States, the Institutional Review Boards (IRB) provide clear guidelines regarding who can access data, even when the data is fully de-identified. Only authorised researchers registered under an IRB-approved project can apply. Each user must undergo mandatory computer-based training and certification in data privacy and ethical humans research, and sign a data use agreement which outlines the scope of use and responsibilities of the user. The penalties for misuse of data are severe. Compliance is monitored through electronic search audits. | This is a very good system that maximises the protection of privacy and minimises the risk of unauthorised use or out-of-scope usage of health data. |
| **Returning results**
Can results be returned to patients? | The original Vanderbilt BioVU and Synthetic Derivative is a fully de-identified system and, as a result, there is no capability to return results. eMERGE and subsequent studies from Vanderbilt contributing to this network use a database that contains identifiable data. eMERGE has discussed the complex issues of returning clinically significant incidental results. There is a return of results oversight committee that reviews cases. Some sites in the eMERGE network have, at this stage, elected not to return findings. For reasons discussed above, the return of results was not possible at Vanderbilt (11). | This is a complex field, and it depends on whether the database contains identifiable or fully de-identified data. The best solution would be to ask participants during the consent process whether or not they wish to have results returned. An opt-out waiver of consent model is problematic, as the issues of returning results for which an individual did not originally consent are complex. Having a fully de-identified database that allows no opportunity for re-identification - and therefore preventing any re-contact with the original patient who contributed the data - actually avoids this problem, as was the case at Vanderbilt. |
| What results are you going to return to patients? | | |
| **Patients accessing their own genomic data**
Will patients be able to get access to their own genotype data? | This is not possible at Vanderbilt as genomic data is completely de-identified, but this may be possible in other centres. This has some issues, as this could result in individuals testing their own genetic data in published phenotypic catalogues. This can be problematic if patients do not receive appropriate genetic counselling or are provided with information in a format that would facilitate understanding. Patients, however, have a right to access to their own health data. | This is a complex problem. Requests for genomic data from individuals may need to be honoured, documented in a formal request, and information provided in a format that can be understood by the participant. Patients will need to be counselled regarding the implications of knowledge of their genetic information (an example is insurance implications) prior to the delivery of information. Genetic counselling will be a crucial. |
<table>
<thead>
<tr>
<th>Issue</th>
<th>Vanderbilt / eMERGE experience</th>
<th>Applicability to Australia</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Consent</strong></td>
<td>Vanderbilt has experience with both the opt-in and opt-out systems. Initially, there was an opt-out waiver of consent model for broad clinical use. This was then followed by an explicit opt-in consent model for broad clinical use. For projects that were initially narrow in their scope of research (i.e. patients not consented for broad research), patients are required to be re-consented if their data or samples will be used for subsequent, unrelated projects. As an example, investigators at Group Health and the University of Washington re-consented participants so that de-identified research data could be deposited in a centralised database, the database of Genotypes and Phenotypes (dbGaP) (12).</td>
<td>A limited consent model runs the risk of the need to re-apply for consent for all contactable study participants. This is very cumbersome and has the potential to hamstring future research projects due to delay. This situation is probably best avoided by having broad consent at the outset. An opt-out model is ideal, but is only possible if there was infrastructure that guarantees permanent de-identification. The Synthetic Derivative at Vanderbilt is an example, but this is very expensive to implement and maintain. Such a program would not be feasible for a single site in Australia given the limited funding available in the research space.</td>
</tr>
<tr>
<td><strong>What form of consent will you employ?</strong></td>
<td></td>
<td>Given the complexity of the ethical and legal issues, it would be worth having this in Australia. Genetic counselling can be used to help with the delivery of genetic test results for study participants. There is a Clinical Genomics Unit (CGU) at the St Vincent’s Campus that offers multidisciplinary clinics in cardiovascular genetics, neurogenomics, vascular genomics, endocrine genomics, immunogenomics and pharmacogenomics.</td>
</tr>
<tr>
<td><strong>When might it be necessary to obtain new consent for the use of previously collected samples?</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Formalised system of genetic counselling</strong></td>
<td>This would be helpful for interpreting test results, counselling patients. A formalised system may be helpful if study results are to be returned to participants. This formalised program does not exist in several centres.</td>
<td></td>
</tr>
<tr>
<td><strong>Is there a need for a centralised genetic counselling service attached to the genome research programs?</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Data sharing</strong></td>
<td>In the United States, the National Institutes of Health (NIH) strongly encourages genomic studies funded by the NIH to be placed in a central database called the database of Genotypes and Phenotypes (dbGaP) so that it can be shared across the broader research community. Best practices have been developed across the eMERGE network for sharing of genomic data and electronic health record-derived phenotypes while protecting the privacy of participants (8). Prior consent is required for sharing of this information.</td>
<td>Protection against hacking and unauthorised access by law enforcement and other federal regulatory bodies remains an important issue in Australia, as it is in the United States. This will likely require protection under legislation to limit the potential for unauthorised access.</td>
</tr>
<tr>
<td>Issue</td>
<td>Vanderbilt / eMERGE experience</td>
<td>Applicability to Australia</td>
</tr>
<tr>
<td>--------------------------------------------</td>
<td>-----------------------------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Community oversight</td>
<td>Extensive community input was sought for the Vanderbilt system, not just through community representation on governance committees, but also through surveys of study participants and the general public.</td>
<td>Community engagement and oversight is very important for the long-term viability of these programs. Australia should set up a similar system.</td>
</tr>
<tr>
<td>And what role should communities play in long-term oversight and governance of these projects?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Legal protection of the database</td>
<td>In U.S., the protection is mandated through law. People who breach confidentiality of the research databases are breaching privacy laws.</td>
<td>Similar protections are most likely to be required in Australia.</td>
</tr>
<tr>
<td>How will you protect the confidentiality of the research databases?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conclusions and Recommendations

• Cardiovascular disease is a leading cause of death in Australia and remains one of the largest burdens on the economy. Advances in the field of cardiac genomics will help explain why some patients respond better than others to treatment and different disease trajectories in heart disease.

• Local genomic studies should ideally be linked to an electronic health record, with appropriate digital security.

• In sites where an electronic medical record system is not present, retrospective linkage through linked data of administrative datasets might prove useful and a compromise solution. Due to issues relating to identification and individual privacy, however, patients will need to have explicit and prior consent for data-linkage.

• The eMERGE network in the United States offers a successful example of multiple local genomic research programs joining together in a collaborative network.

• Ethical, legal, and society issues are an important consideration during genomic research. Successful programs in the United States have invested considerable resources in ensuring that there is appropriate engagement, informed consent, oversight, legal protection and security applied during the set up and conduct of genomic research.

Local

The Vanderbilt system of having a synthetic derivative applied to a single Australian hospital site is too large and expensive to set up and maintain. However, many of the lessons learnt from their experience and those in eMERGE sites can be applied locally.

Example:

Patients with advanced heart failure are seen and treated at a large teaching hospital. Patients who are at advanced stages undergo detailed assessment which involves collection of blood (approximately 10-15 mL of blood). Clinical information is also collected at the time of assessment and entered into an electronic database. The following principles would apply:

• Opt-in consent for broad use, to be implemented during recruitment.
• Samples obtained from leftover blood collected for clinical use.
• Clinical information collected in the electronic database. Different databases have been studied during the fellowship.
• De-identification processes applied to the electronic database.
• If there is any chance that in the future there would be a desire to share data (even de-identified), appropriate cross-institutional data-use agreements will need to be formed, and patients should receive informed consent at the outset that data may be used by authorised research personnel at external sites.

National

The Vanderbilt style of a synthetic derivative could better serve alongside an established national cohort study. Here, the participants would need to have been consented so that their data can be used for broad research use. This idea is more attractive as the infrastructure costs associated with a derived health record are substantial. For the same investment in infrastructure, there would be the maximum benefit – genomic data is not exclusive to cardiac diseases. The data obtained from whole exome chip or whole genome sequencing would be able to be used not just by cardiovascular researchers but researchers from all fields. Suitable storage sites in New South Wales include the NSW Biobank housed the Marie Bashir Building at Camperdown, and the Garvan Institute / Kinghorn Center for Clinical Genomics.
Potential applicability to the *My Health Record* system

The true value of the Vanderbilt style of genomic research, however, is in the potential application of the principles behind its Synthetic Derivative to the Australian *My Health Record* system so that a mirror, de-identified database can be created for the purposes of research and other secondary use.

Vanderbilt University Medical Center also uses a public database that allows patients and their treating practitioners to share medical details, named *My Health at Vanderbilt*. This electronic database of health information ensures that treating physicians have the correct information, and reduces the risk of missing important diagnoses, medication error and unnecessary duplication of tests.

The Australian *My Health Record* system aims to achieve a similar positive effect on healthcare delivery, and proposes to provide a secure, portable digital health record for each person in Australia. Systemic implementation of a digital health record will help safeguard and future-proof healthcare in Australia by allowing individuals and authorised health care providers obtain and input important information on a centralised electronic database.

Many of the concerns relating the *My Health Record*, however, relate to its potential use for purposes other than direct primary health care, collectively referred to as *secondary use of My Health Record data*. Many people have concerns about researchers gaining access to their health records, irrespective of whether or not these records are identifiable or non-identifiable. This issue has contributed to patients opting out of the *My Health Record*. Finding a solution that addresses these community concerns will ensure community satisfaction, ongoing engagement and the long-term viability of this program.

A derived and de-identified electronic health record, designed along similar principles to the Vanderbilt Synthetic Derivative, would provide insight into the health of all Australians while protecting the privacy of individual Australians.

Principles:

- Creation of a mirror database for the *My Health Record*.
- This database consists of altered patient records which serve as “avatars” of the true electronic health records of actual patients from which the specimens were originally obtained.
- Names and other identifiers will be removed.
- Birthdates and dates of events will also be altered by a fixed time-period.
- Patients will not be able to be identified in the database.
- All research and other secondary use items should be performed using this *Derived Health Record*.

Having a *Derived Health Record* will allow studies using large-scale, fully de-identified data to be performed, not only at a local level, but also population-studies at a national level. The presence of this digital infrastructure would also be crucial in advancing the original goal of fostering a genomic network, but in this case, it would benefit not only cardiovascular health but all health disciplines.

As per the Parliament of Australia website, ([https://www.aph.gov.au/Parliamentary_Business/Committees/Senate/Community_Affairs/MyHealthRecordsystem](https://www.aph.gov.au/Parliamentary_Business/Committees/Senate/Community_Affairs/MyHealthRecordsystem)), on 15 August 2018, the Senate referred the *My Health Record* system to the Senate Community Affairs References Committee for inquiry and report, with particular reference to the following items:
• the expected benefits of the *My Health Record* system;
• the decision to shift from opt-in to opt-out;
• privacy and security, including concerns regarding:
  • the vulnerability of the system to unauthorised access,
  • the arrangements for third party access by law enforcement, government agencies, researchers and commercial interests, and
  • arrangements to exclude third party access arrangements to include any other party, including health or life insurers;
• the Government’s administration of the *My Health Record* system roll-out, including:
  • the public information campaign, and
  • the prevalence of ‘informed consent’ amongst users;
• measures that are necessary to address community privacy concerns in the *My Health Record* system;
• how *My Health Record* compares to alternative systems of digitising health records internationally; and
• any other matters.

On this basis, and as part of dissemination and implementation of these important concepts, I made a submission to the Senate Community Affairs References Committee titled: “The need for an Australian Derived Health Record: a practical solution to privacy concerns surrounding secondary use of *My Health Data*”. In this submission, and using the Vanderbilt system as a model, I propose a potential solution to concerns about researchers gaining access to the *My Health Record* system.

As I have previously stated at the end of the Executive Summary, the establishment of a nationwide, de-identified electronic health database that assures “privacy by design” is perhaps the single most important investment and safeguard for future Australian health research. Implementing a Derived Health Record that mirrors and coexists with the *My Health Record* has already been shown to be a practical and proven solution that leverages the maximum benefits that the systematic collection of electronic medical records can offer the community while protecting privacy concerns at an individual level. I look forward to discussing these ideas and working in the future with all interested parties, organisations and government agencies in order to help bring this vision to life.
References

Appendix

Wang L. The need for an Australian Derived Health Record: a practical solution to privacy concerns surrounding secondary use of My Health Data. Submission to the Senate Community Affairs References Committee inquiry into the My Health Records Amendment (Strengthening Privacy) Bill 2018 – 18/105, 1 October 2018 (Full report included in the following pages).
The need for an Australian Derived Health Record:
a practical solution to privacy concerns surrounding
secondary use of My Health Record data

Dr Louis Wang MBBS(Hons) MMed(Hons) PhD FRACP FCSANZ FACC

Executive summary

There remain significant concerns regarding the impact of secondary use of My Health Record data on individual privacy. As an Australian Churchill Fellow, I visited Vanderbilt University Medical Center in Tennessee, USA. Health-related research involving data from the electronic health record at this internationally renowned centre for personalised medicine is predominantly conducted within a de-identified and altered “mirror image” of the true electronic health record. This innovative design offers a practical and proven solution to many of the community concerns regarding privacy and secondary use of health data.

This submission therefore proposes the implementation of a system of two parallel databases: (1) the existing My Health Record, which will be used for direct patient care, and (2) a de-identified and altered mirror database, which for the purposes of this submission will be called the ‘Derived Health Record’. The latter database, intended for approved projects that fall under the category of secondary use of My Health Record data, will help maximise protection of individual privacy, while at the same time providing useful data for medical researchers and health organisations.

Contents

1. Relevance of this submission ................................................................. 2
2. Background................................................................. 3
3. The Synthetic Derivative at Vanderbilt University: a revolutionary and innovative example of a derived and de-identified electronic health record .................................................. 3
4. A Derived Health Record would provide insight into the health of all Australians while protecting the privacy of individual Australians ........................................................................................................... 3
5. Key principles in the creation of a Derived Health Record .................................. 4
6. Advantages of a Derived Health Record .......................................................... 6
7. Disadvantages of a Derived Health Record .......................................................... 6
8. Conclusion .................................................................................................... 7
9. Acknowledgements ..................................................................................... 7
10. References.................................................................................................. 7

Corresponding author
Dr Louis W. Wang
2015 Australian Churchill Fellow, and
Postdoctoral Fellow to Professor Dan Roden,
Vanderbilt University Medical Center, 2018
St Vincent’s Clinical School, UNSW Sydney
Email: louis.wang@unsw.edu.au
1. Relevance of this submission

On 15 August 2018, the Senate referred the My Health Record (MHR) system to the Senate Community Affairs References Committee for inquiry and report. This submission has special relevance to the inquiry’s terms of reference (Table 1). It discusses the key principles used in an established, widely published and internationally renowned synthetic electronic health record system at Vanderbilt University in Tennessee, USA (1).

This synthetic health record, effectively a mirror database derived from real-life health data, has been shown to be successful in addressing community privacy concerns regarding secondary use of real-life health data (2). Translation of some of the key principles from this program to the Australian My Health Record system has the potential to solve many of the existing issues raised during this inquiry.

Table 1: Relevance of this submission to the Inquiry’s terms of reference

<table>
<thead>
<tr>
<th>Terms of reference</th>
<th>Relevance of this submission</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. the expected benefits of the MHR;</td>
<td>• This submission offers an alternate method of facilitating secondary use of MHR data (and its expected benefits) while at the same time addressing community privacy concerns.</td>
</tr>
</tbody>
</table>
| B. the decision to shift from opt-in to opt-out; | • Satisfactorily addressing community privacy concerns may help reduce the numbers of Australians deciding to opt-out.  
• Incidentally, the Vanderbilt program, when conceived, was also an “opt-out” model. |
| C. privacy and security, including concerns regarding: | • The Vanderbilt program has been widely successful in addressing issues relating to the real and actual risks of vulnerability of the system to unauthorised access, and dealing with the legal and ethical issues relating to third party access by government agencies. |
| i. the vulnerability of the system to unauthorised access, |  |
| ii. the arrangements for third party access by law enforcement, government agencies, researchers and commercial interests, |  |
| iii. arrangements to exclude third party access arrangements to include any other party, including health or life insurers; |  |
| D. the Government’s administration of the MHR system roll-out, including: | • The Vanderbilt program studied the prevalence of ‘informed consent’ amongst users of health care during their roll-out program (3, 4). |
| i. the public information campaign, |  |
| ii. the prevalence of ‘informed consent’ amongst users; |  |
| E. measures that are necessary to address community privacy concerns in the MHR; | • The proposal addresses community privacy concerns by reducing the need for researchers to access real-life data. |
| F. how MHR compares to alternative systems of digitising health records internationally. | • This submission describes the systems in place at an internationally renowned centre for personalised medicine. Information relating to the Vanderbilt system is available in the published literature. Several key papers have been included in the reference list (1, 2). |
2. Background

The My Health Record system provides a secure, portable digital health record for each person living in Australia. Systemic implementation of a digital health record will help safeguard and future-proof healthcare in Australia by allowing individuals and authorised health care providers obtain and input important information on a centralised electronic database. This will help avoid medication error and unnecessary duplication of tests, and will have significant economic and direct patient benefits.

Many of the concerns relating the My Health Record, however, relate to its potential use for purposes other than direct primary health care, collectively referred to as secondary use of My Health Record data. Many people have concerns about researchers gaining access to their health records, irrespective of whether or not these records are identifiable or non-identifiable. This issue has contributed to patients opting out of the My Health Record. Finding a solution that addresses these community concerns will ensure community satisfaction, ongoing engagement and the long-term viability of this program.

3. The Synthetic Derivative at Vanderbilt University: a revolutionary and innovative example of a derived and de-identified electronic health record

Vanderbilt University Medical Center is located in Nashville, Tennessee, USA. This health system services Middle Tennessee and is home to BioVU, a large-scale de-identified DNA biobank, which was created to facilitate big-data epidemiological and genomic research. This pioneering system has been in operation since 2007. A key component of the program is the creation of a “mirror image” of the electronic medical record (1). This database, known as the Synthetic Derivative, contains clinical information derived from the Vanderbilt Health electronic health record system. It contains patient records that have been altered with respect to dates and devoid of identifying details, such that the altered (“synthetic”) health record only resembles the original health record in terms of diagnoses, test results and prognostic information. The Synthetic Derivative continually accrues new clinical data as they occur over time. This system presents an innovative solution to de-identification and privacy concerns. As it contains synthetic patient records derived from actual patients, this database no longer contains real patient records, representing, in effect, a similar population in an alternate reality. At Vanderbilt University Medical Center, the Synthetic Derivative can be used by researchers as a stand-alone resource or to link clinical information with genomic data from the BioVU biobank. The Vanderbilt Synthetic Derivative contains data from over 2.2 million people. Its search interface allows users to input clinical and demographic information (e.g. diagnoses, procedure codes, medications, laboratory test values, age and gender) and returns de-identified data. This system is fully compliant with the Security and Privacy Rules within the Health Insurance Portability and Accountability Act (HIPAA) 1996 (1), and operates under extensive institutional review board oversight.

4. A Derived Health Record would provide insight into the health of all Australians while protecting the privacy of individual Australians

The key features of this proposal, which guarantees privacy by design, are illustrated in the Figure on the next page. Essentially, it involves two databases, (1) the My Health Record, and (2) the Derived Health Record (Table 2). The My Health Record contains true records from real-life people, and this database, with rare exception, should be reserved for primary use only. Each My Health Record will be de-identified and altered to create a “derived” health record in a mirror database, termed the Derived Health Record. As these records are synthetic but derived from real people, the proposed Derived Health Record can be thought of as a database of electronic health records from Australia belonging to an “alternate reality”, populated by avatars of actual people living in real-life Australia who develop medical conditions, attend health care, have tests, receive treatment, and experience health events.
Since each record in this Derived Health Record is derived from real-life data (diagnoses, test results, procedure codes, outcome data) from actual people, inferences from analyses performed on the Derived Health Record are therefore applicable and generalisable to the Australian population. This will help leverage the maximum benefit that having a population-level electronic digital record can provide for epidemiological research. It will enable the study of disease-associations, patterns of disease incidence and prevalence, and provide insights into trends in healthcare access and resource uptake. The availability of this alternate database will reduce the need for people not involved in an individual’s health care to access that person’s health data and provides an additional layer of protection to privacy.

5. Key principles in the creation of a Derived Health Record

These principles from the Vanderbilt model can be adapted to the Australian system to minimise potential harm as well as addressing privacy concerns.

A. Effective de-identification

The United States has a Safe Harbor standard under the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule which requires the exclusion or removal of 18 identifiers from a de-identified dataset (1, 5). At Vanderbilt University Medical Center, commercially available software (DE-ID, DE-ID Data Corp) was used to remove these personal identifiers from medical records (1).
Each medical record number in the true electronic health record is recoded by a secure hash algorithm (SHA-512, National Security Administration) to produce a Research Unique Identifier (RUI), a 128-character code that is unique to a particular medical record number. The nature of the secure hash algorithm is such that it is not possible to re-compute the original medical record number from the derived RUI. A similar secure hash algorithm can be applied for an Australian Derived Health Record.

The following identifying features of each health record should be permanently removed on entry into the de-identified database:

- Name(s) and aliases;
- Date of birth;
- Place of residence;
- Medical record number;
- Medicare and insurance card numbers;
- Contact numbers, email addresses and social media accounts;
- Vehicle identifiers and serial numbers, including license plate numbers;
- Biometric identifiers
- Identifiable photographic or clinical images.

B. Alteration of each “real” health record so that it becomes a “derived” patient record

The exact dates of clinical episodes and other key personal events are altered. In the Vanderbilt Synthetic Derivative, all dates in the synthetic record are shifted 1 to 364 days into the past (f). The shift is different across different individual records. However, within each individual record in the Synthetic Derivative, the time-shift of events is consistent throughout the entire record (i.e. for all past and future events), thereby facilitating accurate analysis of longitudinal follow up data. A similar time-shift of up to several years can be performed for derived records in the proposed Australian Derived Health Record system.

Table 2: Key features of My Health Record and the proposed Derived Health Record

<table>
<thead>
<tr>
<th>Key features</th>
<th>My Health Record (MHR)</th>
<th>Derived Health Record (DHR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is in this database?</td>
<td>The MHR contains the records of real patients</td>
<td>This proposed database will contain de-identified and “altered” records. Each patient in the DHR will be based on a real patient in the MHR, but no key identifiers from the original patient will remain. Dates of key events will be time-shifted.</td>
</tr>
<tr>
<td>Intended purpose?</td>
<td>This database (MHR) will be used solely for primary use (i.e. direct health care).</td>
<td>This proposed database (DHR) will be available for approved secondary use</td>
</tr>
<tr>
<td>Is there capacity or availability for secondary use?</td>
<td>There will be occasions where research or analysis will require data from the MHR (e.g. investigating event-based outcomes, or outcomes following the roll-out of a particular program). When MHR data is required, application for access may be requested, and only granted if the project and research personnel satisfy strict guidelines and regulatory oversight.</td>
<td>It is envisaged that for most projects that are currently classified as secondary use and allowed under the current legislation, researchers will be using data derived from the proposed DHR. Applications for access to this data (which is permanently de-identified) will be made as per the existing legislation.</td>
</tr>
</tbody>
</table>
C. Random exclusion of a percentage of records

This technique has been used in the Vanderbilt Synthetic Derivative (2). Random exclusion of a percentage of records from the proposed Derived Health Record database makes it impossible to know with certainty whether a particular individual is present in the database. This provides an additional layer of privacy by increasing the difficulty in being able to infer the presence of any particular individual in a dataset.

D. Imposing a minimum dataset requirement for research.

Inadvertent re-identification is easier when there are fewer numbers of subjects. To reduce the potential for re-identification of included subjects when the number of patients within a particular research project are small, the search interface only returned a data set to an approved investigator if the numbers of matching records exceed a specified minimum number (2). A similar minimum dataset requirement would be useful within the Australian system. Australia’s geography offers additional challenges to the protection of privacy, especially in areas with sparse population, where it may be theoretically possible to infer the true identities of supposedly “de-identified data” when limited samples are linked with their geographic location (e.g. postcode). To improve anonymity, having a minimum dataset requirement for research requiring geographical information will help to reduce the risk of re-identification.

6. Advantages of a Derived Health Record

One of the key benefits of using a permanently de-identified database is that patients cannot be re-identified. This protects privacy and confidentiality, especially in an era where multiple researchers from many different research teams may potentially gain access to data. The establishment of a Derived Health Record protects patient confidentiality as there is less need for researchers to view identified patient records and provides an alternative data resource for researchers when having identifiable data is not essential for the conduct of research (2).

Studies that require time-dependent data (e.g. prognostic studies) will still be possible under the proposed system, as the random time-altering function will be uniformly applied to all events within an individual record, so that exact follow up time after a sentinel event or exposure can be obtained.

7. Disadvantages of a Derived Health Record

The time-shifting algorithm used in preserving anonymity in a Derived Health Record means that events tied to specific dates (e.g., studies of effects of natural disasters, policy changes, and other important events) cannot be evaluated with this database (1). These projects, instead, will require access to data from the My Health Record database. Nevertheless, with the availability of the Derived Health Record, it is anticipated that the vast majority of research projects can be completed without ever needing to access direct data from the My Health Record.

For research using the Derived Health Record, contact with individual patients is not possible, and also likely to be prohibited. The design of the Derived Health Record means that results can only be inferred at a population level, and that there is no means of returning a significant result to an individual as the de-identification process is permanent. The lack of any identifiers means that it would also be impossible to contact any individual for further information, and only the information other already present within the Derived Health Record will be available for the purposes of research. As an example, if specific biomarkers are found to be strongly associated with a disease in a permanently de-identified dataset obtained from the Derived Health Record, there will be no means of returning these results to the patients who originally contributed the data. This is in keeping with the existing principles guiding the approval of My Health Record data for secondary use. Future studies that
aim to provide direct benefits for individual patients cannot be answered using the proposed Derived Health Record, and should undergo the traditional pathway of formal, opt-in, informed consent. As per existing regulations, the informed consent process for these studies should explicitly ask the patients to nominate whether or not they wish to be contacted regarding any significant or incidental findings generated from their participation in the study.

There will be significant costs associated with the creation, curation, storage and maintenance of a second, mirror database. There will be at least twice the amount of database storage required, as well as ongoing costs related to digital infrastructure and security. This cost, however, should be offset by the enormous value and flow-on benefits that this digital infrastructure would provide for Australian big-data research.

8. Conclusion

The establishment of a nationwide de-identified database that assures privacy by design is perhaps the single most important investment and safeguard for future Australian health research. Implementing a Derived Health Record that mirrors and coexists with the My Health Record is a practical and proven solution which leverages the maximum benefits that the systematic collection of electronic medical records can offer the community while still protecting privacy concerns at an individual level.

9. Acknowledgements

I sincerely thank the Winston Churchill Memorial Trust and Dr Dorothea Sandars, Professor Dan Roden (Senior Vice-President for Personalised Medicine and Founder of BioVU) and Professor Ellen Clayton (Rosalind E. Franklin Professor of Genetics and Professor of Law) at Vanderbilt University for the Churchill Fellowship and my experiences at Vanderbilt University Medical Center.

10. References

These key articles are publicly available via PubMed.gov (US National Library of Medicine, National Institutes of Health). They provide additional information regarding the design and implementation of the Vanderbilt Synthetic Derivative, and how this program addressed issues relating to oversight, community engagement and privacy.


