Web Based Cohort Identification across Large Healthcare Data Sets – Opening the Treasure Chest

Christopher Bain, Christopher Mac Manus, and Jarrel Seah

***Abstract*—**We have previously constructed a large informatics platform at a major teaching hospital through the aggregation of massive and disparate data sets. This has been a strategic initiative in order to serve multiple needs across the organization - including across the domains of operations, quality improvement and research- all of which are critical to the functioning of a large academic health centre. We have subsequently been able to leverage off the platform by creating user tools, including an intelligent web-based self-service query tool. The tool does and will serve several functions, but is predominantly designed to support the identification of cohorts of patients for operational or research purposes. In this paper we report on the rationale for the tool, it's design, construction, validation, and projected development path.

***Keywords*—**Big data, cohort, hospital, REASON, search.

I. Introduction

ALTHOUGH working with, and making sense of, "Big Data" is not without its dangers [1], the benefits of its use are thought to be significant [2]. Some of the purported advantages of the "Big Data" paradigm are the ability to mine data sets for patterns, identify uncommon events and to unearth interesting or valuable insights.

As previously described in the literature [3], we have constructed a platform in this paradigm called The REASON Discovery Platform®. We have also previously described some of the actual and potential value from the platform [4-8].

The concept of, and need for, identifying cohorts of patients is a common one both in the healthcare literature and in the operational delivery of healthcare. Let us explain further.

One example is for medical research- often research studies, for example of a new drug, require the identification of patients with some specific criteria - eg - diabetic with anaemia, or a heart attack in the past and hypertension - to be considered as candidates for the new drug or other intervention.

Another example is audit- typically but not only clinical audit. In this scenario internal or external auditors may seek to examine the evidence around care for a particular condition or group of patients. Often there is not an existing pre-canned report to return details on the necessary group of patients, and a manual adhoc data gathering or extraction exercise is required.

In this paper we outliner the development and validation of a Cohort Discovery Tool (CDT) that leverages the vast data sets in the REASON platform to try and address these needs.

Christopher Bain is with Alfred Health and Monash University, Melbourne, Australia (+61 3 9076 3079; e-mail: christopher.bain-info@alfred.org.au).

Christopher Mac Manus is with Alfred Health, Melbourne, Australia (+61 3 9076 5433; email: C.Macmanus@alfred.org.au).

Jarrel Seah is a final year medical student studying at Monash University, Melbourne, Australia (e-mail: jarrelscy@gmail.com).

II. Context

Our health service (Alfred Health (AH)) has 3 main facilities, and several smaller satellite facilities, under its control, as well as many ambulatory services. It also provides statewide referral services in the areas of adult trauma, adult burns and organ transplantation. The original setting for this work is the creation of a Health Informatics (HI) department at the health service in early 2011. The department was charged with assuming responsibility for the technical development and management of the data and reporting infrastructure, whilst another key business unit was charged with the responsibility for delivering data and reporting off the infrastructure. In the mid part of 2011, HI along with Health Information Services (HIS), Information Technology Services (ITS) and the Australian Centre for Health Innovation was brought under a single new business division – the Information Development Division (IDD). It was the vision of the new divisional head to continue to develop this technical infrastructure as part of a broader plan. This work is now being continued in the recently created Information Services Department (ISD).

It can be difficult for readers to get a full appreciation of what is being attempted through the construction of the REASON platform. There are 2 United States (US) based examples outlined below, that allow the reader to get a sense of the amount of work done to date to establish this infrastructure, and the direction of travel of the platform.

One US initiative is "Informatics for Integrating Biology and the Bedside" (i2b2), although this operates under a different kind of governance model to our platform, and arguably has a broader reach [9]. One of the primary aims of the i2b2 Centre is in "developing a scalable computational framework to address the bottleneck limiting the translation of genomic findings and hypotheses in model systems relevant to human health". This initiative is well known

internationally and there are even competitions to analyze data provided from the platform.

The REASON platform however, is most analogous to STRIDE (**S**tanford **T**ranslational **R**esearch **I**ntegrated **D**atabase **E**nvironment), the Stanford based informatics platform. STRIDE "is a research and development project at Stanford University to create a standards-based informatics platform supporting clinical and translational research." [10]). There has been evidence published in the international literature pertaining to the benefits to health care processes, and patients, of such a platform. [11]

In order to set the scene a little further, let us describe in a broad sense some of the key data contained within the platform (see Table 1).

TABLE I
REASON- NUMBERS OF RECORDS BY TYPE

Record Type	Record Numbers
Number of Admissions	>900,000
Number of Emergency Encounters	>920,000
Number of Pathology Results- Atomic	>45,000,000
Number of Pathology Results- Textual	>700,000
Number of Patients	>1,900,000
Number of Radiology Reports	>770,000
Number of Radiology Test Orders	>800,000
Number of Surgeries Performed	>160,000

III. SYSTEM OVERVIEW

The AH REASON CDT enables the retrieval of complex cohorts from a Structured Query Language (SQL) database with minimal training. By creating a simple event based query language, cohorts of patients such as those with a diagnosis of X followed by a diagnosis of Y can be retrieved without the need for writing or even understanding SQL.

When opening the URL for the CDT, users are presented with a standard login screen prompting for a username and password. Only those users specifically added to an Active Directory (AD) group dedicated to the application can then access it. On the main search screen (see Figure 1) users are required to enter a maximum number of records to be returned by the search, as well as the search criteria which can be indicated to the system in 2 ways.

In relation to the maximum number of records to be returned, in most cases this can simply be set to an arbitrarily high number relative to the query. So for instance in the case of chordoma, this could be set at say 1000, since users would appreciate it is simply not possible that this many records could be returned. Users can however enter a number, in conjunction with a date range (in the main search text boxes) to act in concert in limiting the number of records returned.

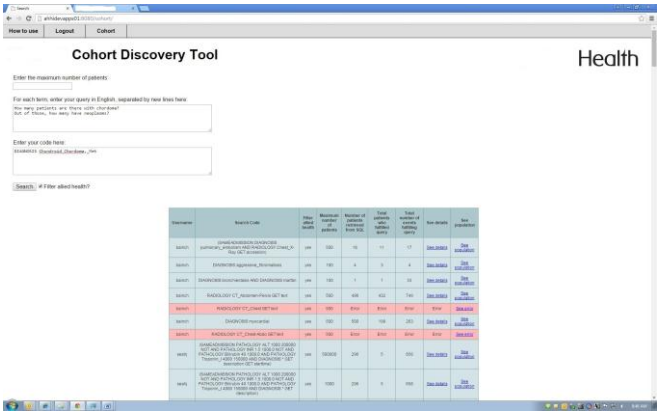

Fig. 1 CDT search entry screen

In relation to the main search text boxes (see Figure 1), users can enter plain English language statements - eg - "how many patients have chordoma ? " which the system will interpret and translate into machine understandable code. This machine understandable code is then displayed in the second text box and is fed into the application when the user hits the "Submit" button.

Alternatively, more informed users can enter the relevant machine understandable code eg – "DIAGNOSIS Chordoma" into the second text box and hit "Submit". Errors of syntax are handled by the application and are displayed in red in the "In progress" box (see Figure 1). Users can click on these to read what the nature if the error is, otherwise the desired query runs.

Fig. 2 CDT search result delivery screen

When the results of the search are returned, the user has 2 results screens they can open via hyperlinks against the row representing their specific search (see Figure 1). The first contains priority ranked lists of results that the user can view on the screen and scroll through (see Figure 2). They can also then export the results to Microsoft Excel. The second contains the details of the results in a similar scrollable format which can also be exported to Excel. So for example, the second screen will contain the list of relevant UR numbers (unique patient identifiers) and any attached details for each – eg - the relevant sodium level result, or the full text result of the relevant computerized tomography (CT) scan.

IV. Methodology

A. System Architecture

The CDT uses SQL templates to generate a superset of patients that fulfil the criteria – e.g. to retrieve patients that have a diagnosis of X followed by Y, patients with both diagnoses are first retrieved from the SQL server without consideration of their temporal relationship. The CDT populates its own event-driven data model from the data retrieved from the SQL server, and further refines the search.
Technically, the tool is written in Python, supported by a number of open source libraries, and uses a Django framework to serve a HTTP front-end. Usability is improved upon by having server-side autocomplete functions that predict exact names of diagnoses and pathology test, obviating the need for users to look these up in a separate database.
The architecture of the CDT is displayed in the diagram below (Figure 3):

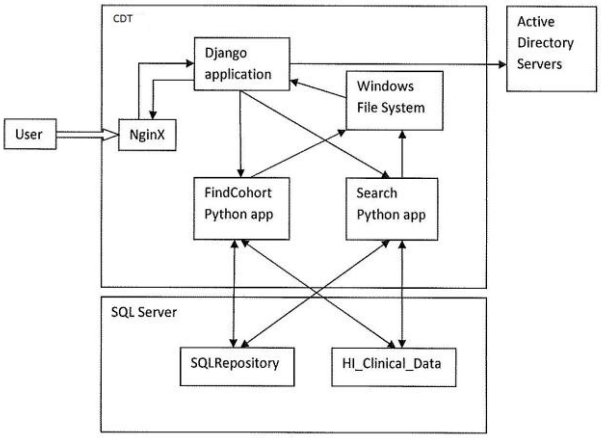

Fig. 3 CDT system architecture

The CDT extends upon this to provide an automatic odds-ratio calculation within cohorts to predict the likely diagnoses associated with said cohort, using commonly employed statistical tests such as Fisher's exact test to calculate significance values. These latter features of the tool have not been explored in this piece of research.

B. System Development Approach

The approach used in this work has been one of evolutionary prototyping underpinned by a heavily user-centred design philosophy. The department leading the work (ISD) is a heavy users and providers of data themselves, and so have been acting as test users of the evolving tool along with the developer, and whilst guiding next steps through each iteration of development.

C. System Validation

Chordoma was the initial "reference case" identified early in the development process. It is a rare tumour of the spinal axis that tends to occur at either end of the spine - the skull, or alternatively in the sacrum (base of the spine). It only occurs in about 1 person per million people per year. So in a city of 5 million people, 5 new cases per year would be typical. This condition was chosen in part on the following basis - if the tool could identify the likely handful of cases at our hospital over the last 5 years (start of 2010 to the time of writing), due to the rarity of the condition, then we could feel confident in its ability to search with high "sensitivity".

Coarctation of the aorta is a condition causing narrowing of the aorta (the massive artery leaving the heart and sending fresh blood around the body). It too is somewhat rare but is more common than chordoma. It was chosen for its rarity but also because of the likely availability of validation data sets from other hospital departments.

Finally, meningioma is a much more common condition. It is a tumour on the outside of the brain or spinal cord. Again it was in part chosen due to the likely availability of validation data sets from other hospital departments.

In each case additional data was sought via different departments to examine the number of cases of each condition over the 5 year time frame. In the case of chordoma, data was sought from the medical records department (HIS), as well as the hospital radiotherapy department and the state cancer registry. In the case of coarctation of the aorta HIS and the hospital cardiology department provided data. Finally, in the case of meningioma, HIS and the hospital neurosurgery department provided data.

V. Outcomes

It is inevitable that data queries performed by different individuals, and in some cases from different data sources, will not deliver identical results. What is important in this validation exercise however is that in each of the test cases there was a high degree of concordance amongst the different views provided from the different perspectives, of patient groups that had been treated at our health service over the last 5 years.

It can be seen from the data in Table 2 that in the cases of both chordoma and coarctation of the aorta, this initial analysis revealed a high degree of correlation between the different data sources and extraction approaches, in terms of inpatient (patients who stayed in hospital) numbers with these conditions. This does not talk to the reason for the admission, just that at the time of admission they were noted to have had the relevant condition. It is somewhat surprising however, to see how few cases of coarctation were evident in the inpatient population, even in our adult facility, given that some sources put the incidence at 1 in 2000 live births [12] (versus the incidence of 1 case per million population per year for chordoma). A further discussion of reasons for this is beyond the scope of this paper, as it in in part is to do with the medical presentation and management of this condition.

In the case of meningioma, further work is going on to examine the number of unique patients identified from each source, but the initial results are again very encouraging despite first appearances. The variance of 49 records (380-331) is due to 2 measurable issues. One is a date cut off issue – so that the request for a given date range was interpreted by the HIS staff in a different way to how the CDT works. So

whilst the CDT looks at the discharge date of a patient admission to select records, the HIS staff used the admission date to select the relevant records. The more interesting kind of discrepancy, however, is the second one.

In the second case, the beauty of the CDT is that does not require users to know which International Classification of Disease (ICD-10) [13] codes, for example, to use to identify patients with certain conditions. So when the CDT was asked to find cases of meningioma - only that word, "meningioma", was used. As a result, the tool returned cases wherever that word was picked up in the appropriate fields. As a result cases of "meningioma not otherwise specified" (ICD 10 Code-PM9530/0) were returned as well as other variants of meningioma such as "atypical meningioma" and "meningioma- malignant". When the HIS staff went to address the query, they chose to use only the ICD 10 Code-PM9530/0 to identify cases and hence their result did not include a number of more rare variants of meningioma. These 2 issues combined accounted for the difference of 49 cases.

TABLE II
CDT RECORD VALIDATION BY TYPE

Condition (Inpatient records)	**Cohort Tool Numbers**	**Extra Source 1 (HIS)**	**Extra Source 2 (Clinical Dept)**	**Extra Source 3 (Cancer Registry)**
Chordoma (patients)	6	4	5	5 (notifications)
Coarctation of the aorta (patients)	7	6	6	Not applicable
Meningioma (Admissions)	380	331	Not yet available	Not applicable

This is by no means intended as a criticism of the HIS staff. This is the kind of thing that routinely occurs whenever a human performs a query such as this – they must, by definition, overlay their knowledge of the context onto the task, and this can result in variable results between individuals in trying to answer the same query.

This fact is interesting in itself and highlights one key aspect of the value of the tool. In many hospitals there are multiple source information systems, each with their own custodians or access points, that contain unique, or sometimes overlapping data. The implication of this is that if a customer has a given query, establishing the "absolute truth" pertaining to the data they need can be fraught. Humans will inevitably draw on their own knowledge and experience to answer such queries, as mentioned above, especially in the absence of strong corporate data governance - where all relevant definitions are agreed, known and documented- and even with the best of intentions. The net effect for the customer is that their single question or need may result in a range of answers or responses, depending on who answers their need, and how that person chooses to go about deriving a result. Tools such as this diminish this risk and in principle should increase the quality of provided data over time.

Such tools also reduce what we might call "the round-about effect" which we often see in response to seemingly simple queries from customers. What can happen in the case of ad-hoc queries, which are often defined by their nature as being new or rarely performed (as opposed to regular and common queries that are usually embedded in routine reports that are available to customers), is that there can be a disconnect between what a customer is trying to achieve, and what the analyst or other intermediary translates that into in terms of query code (eg - SQL), and output provided back to the customer.

As a result there may be 2, 3 or even more iterations of the process, if the need is important enough, before the customer finally gets data that addresses their need or helps them answer their underlying question. Clearly that is a suboptimal arrangement.

Tools such as ours are far from fool proof, but they can short cut this process by allowing a customer to explore the data, using language they can relate to, many times over if needed, in order to arrive at the desired output. This means in turn that any involvement from an analyst can be as a value-add by imparting their knowledge of the underlying data and its meaning, rather than wasting time on the mechanics of access and retrieval of data which has essentially been "outsourced" to the software.

Standardized reporting and business intelligence systems - if properly constructed - can assist with this issue. Typically however, not all data that can be found across any given healthcare organization is architected within such systems. There is always a regular need for adhoc data queries either because of the nature of the data, or because of the nature (eg -complexity) of the query. This is especially true in the absence of a standardized clinical terminology (eg – Systematized Nomenclature of Human Medicine (SNOMED) [14]) in use across the organization.

VI. THE FUTURE

A. Further Testing

One of the very first next steps pertaining to the tool is to create further test cases and carry out further testing. Each of the existing and future test cases will allow regression testing as the system is modified and improved going forwards.

The next test cases that we will carry out are of 2 main kinds. The first is to look at the ability of the CDT to return more complex results against other verifiable outputs. So for instance the number, type and content of radiology and pathology results can also be obtained from the relevant departments, and these outputs can be compared against outputs from the tool.

The second is to examine the capabilities of the tool and its "proxy query language" to relate "events". This kind of testing will be more difficult. So for example, the ability of the tool to return all patients who had a chest x-ray (CXR) ***after*** they had previously had a diagnosis of a heart attack. In

principle it should be readily possible to obtain validation data sets from other areas of the business, or data sources, with which to compare. However, as soon as this extra level of complexity is introduced, the ability of others (humans) to perform such a query, and return an "absolute truth" or "gold standard", comes into question. Nonetheless we will attempt this.

B. Subsequent Use and Impacts

The potential users of the system are many and varied. In fact at the time of writing, having established a good level of robustness of the tool and it's functional capabilities, we are conducting a real-world pilot with a group of staff from across the health service. This group includes health information managers (the specialists in medical records), data analysts, researchers, clinical managers and doctors. These groups have been chosen as being representative of the greater pool of potential end users, and in turn this is based on 2 key criteria. The first is that they have a regular need to acquire data sets, often through ad hoc queries, for their own work purposes or to meet the needs of others. The second criteria is that each already has full access to the electronic medical record (EMR) system in use at the hospital, and to many other systems. What this means is that already have full access to highly sensitive information about patients in the course of their work- there is nothing new they can see through this tool in that sense, it's just that the tool answers questions and undertakes processing (eg - massive aggregation) that the human brain cannot.

We anticipate that as a result the need for human involvement in ad hoc data queries will diminish substantially, and that this in turn will generate a non-trivial labor saving. For example, it is known that the hospital reporting unit currently fields up to 100 new queries per month. Any given query may take many hours of time for an individual analyst to complete and check.

As seen already in the testing process to date, the tool is potentially capable of outperforming a human analyst, as it does not rely on human interpretation of code sets from the human language request. This feature does have potential drawbacks but it is still overall a very promising one.

C. Further Development

An initial next step for the tool is to improve the interface (eg- – with drop down selections for common searches), and embed it in a broader web-based portal which seeks to aggregate and centralize a number of web-based applications that feed into or off the platform.

Currently the tool is only fully functional when accessed via Google Chrome. Whilst this browser is available for use by hospital staff, it is not the primary browser in use at the hospital – that is Internet Explorer. Work will need to be done to remedy this situation.

The tool also needs to be readily accessible from mobile devices given the range of things it could be used for over time (eg - to potentially aid nurse decision making on the wards) – so this will be another key early aim of subsequent development efforts.

VII. Conclusion

In this paper we have described the design and development of a web based cohort identification tool. The tool has been shown to work well on a selected range of conditions from the rare to the moderately common. Whilst the tool needs further validation and operational testing, it shows enormous promise as a means to increase the accuracy of data extracts for staff, and the efficiency with which such extracts can be provided. Future development will also include interface improvements and the ability to access queries and results in a format more compatible with mobile devices.

Acknowledgment

We would like to acknowledge the contributions of the following ISD staff members to this work: Bismi Jomon, Annie Gilbert, Hien Le, David Kelly and Robin Thompson. In addition we would like to acknowledge Ms Lise Hales from HIS, Dr Peter Bergin from Cardiology, Mr Phil Lewis from Neurosurgery, and the staff of the state cancer registry, for assistance with the validation exercise. Finally we would like to thank all of our colleagues in the IDD who have assisted with this work over time, in particular the EMR and Infrastructure teams.

References

[1] D.Lazer, R. Kennedy, G. King, and A. Vespignani. 2014. The Parable of Google Flu: Traps in Big Data Analysis. Science 343 (6176) (March 14): 1203–1205.

[2] T.Murdoch and A.Detsky. The Inevitable Application of Big Data to Healthcare. JAMA. 2013;309(13):1351-1352.
http://dx.doi.org/10.1001/jama.2013.393

[3] C.Bain and C.Mac Manus. Advancing data management and usage in a major Australian health service: The REASON Discovery Platform ™. Proceedings of the International Conference on Data Science and Engineering 2014.

[4] N.Good, C.Bain, D.Hansen and S.Gibson. Health informatics visualisation engine – HIVE. Big Data Conference, Abstract Book. pp30-31. Big Data Conference, Melbourne April 2014.

[5] D.Martinez, L.Cavedon, Z.Alam, C.Bain and K. Verspoor. Text mining for lung cancer cases over large patient admission data. Big Data Conference, Abstract Book. pp24-25. Big Data Conference, Melbourne April 2014.

[6] C.Bain, T.Bucknall and J. Weir-Phyland. A clinical quality feedback loop supported by mobile point-of-care (POC) data collection. IMMoA Workshop 2013. Trento, Italy August 2013.CEUR Workshop Proceedings Vol 1075.pp44-51.

[7] C.Bain, T.Bucknall, J. Weir-Phyland, S.Metcalf. P.Ingram and L.Nie. Meeting National Safety and Quality Health Service Standards – The Role of the Point-of-Care Audit (POC) Application. IJEEEE 2013 Vol.3(6): pp 507-512.

[8] C. Bain, J. Weir-Phyland, S. Metcalf, P.Ingram and T.Bucknall. Driving clinical safety initiatives through innovative technological feedback systems in an Australian academic health service. – Oral Presentation. IARMM 2nd World Congress on Clinical Safety. Sep 12-12 2013. Heidelberg, Germany.

[9] I2B2 - https://www.i2b2.org/about/index.html. Accessed 24/12/2013

[10] STRIDE -https://clinicalinformatics.stanford.edu/research/stride.html Accessed 24/12/2013

[11] J.Frankovich, C.Longhurst and S.Sutherland. Evidence-Based Medicine in the EMR Era. N Engl J Med 2011; 365:1758-1759
http://dx.doi.org/10.1056/NEJMp1108726

[12] Pediatric Heart Specialists Web site. http://pediatricheart specialists.com /articles/detail/coarctation_of_the_aorta

[13] ICD- WHO Website. http://www.who.int/classifications/icd/en/ Accessed 9/2/2015

[14] SNOMED- NEHTA Website. http://www.nehta.gov.au/our-work/clinical-terminology/snomed-clinical-terms Accessed 9/2/2015

Chris B. is the Director of the Information Services Department at Alfred Health in Melbourne, Australia and was the inaugural Director of Health Informatics there. In addition he is an Adjunct Associate Professor in the Faculty of IT at Monash University also in Melbourne. Chris has professional IT accreditation with the ACM in the US and the ACS in Australia. He is also an Associate Fellow of the Australasian College of Health Service Management and a Fellow of the Australasian College of Health Informatics.

Chris M. is the Manager of Data Services in the Information Services Department at Alfred Health, and is currently completing a Master of Public Health. Chris is a Computer Scientist and Software Engineer by training and has been working in related roles for 12 years.

Jarrel S. is a final year medical student at Monash University in Melbourne. In addition he has completed a Bachelor of Medical Science on computational diagnostic support of cervical spine injuries, and is an accomplished software developer, having developed the internationally known and award winning Eyenaemia app. (www.eyenaemia.com)