COPYRIGHT NOTICE

Addressing the Complexities of Big Data Analytics in Healthcare: The Diabetes Screening Case

Daswin De Silva
La Trobe Business School
La Trobe University
D.DeSilva@latrobe.edu.au

Frada Burstein
Centre for Organisational and Social Informatics
Monash University
frada.burstein@monash.edu

Herbet Jelinek
School of Community Health & Centre for Research in Complex Systems
Charles Sturt University
HJelinek@csu.edu.au

Andrew Stranieri
Centre for Informatics and Applied Optimization
Federation University
a.stranieri@federation.edu.au

Abstract

The healthcare industry generates a high throughput of medical, clinical and omics data of varying complexity and features. Clinical decision-support is gaining widespread attention as medical institutions and governing bodies turn towards better management of this data for effective and efficient healthcare delivery and quality assured outcomes. Amass of data across all stages, from disease diagnosis to palliative care, is further indication of the opportunities and challenges to effective data management, analysis, prediction and optimization techniques as parts of knowledge management in clinical environments. Big Data analytics (BDA) presents the potential to advance this industry with reforms in clinical decision-support and translational research. However, adoption of big data analytics has been slow due to complexities posed by the nature of healthcare data. The success of these systems is hard to predict, so further research is needed to provide a robust framework to ensure investment in BDA is justified. In this paper we investigate these complexities from the perspective of updated Information Systems (IS) participation theory. We present a case study on a large diabetes screening project to integrate, converge and derive expedient insights from such an accumulation of data and make recommendations for a successful BDA implementation grounded in a participatory framework and the specificities of big data in healthcare context.

Keywords: big data analytics; health informatics; clinical decision support; translational research; business analytics; information fusion

1 Introduction

A new demand, increasingly driven by cost pressures and evidence-based medicine, is pushing the healthcare sector to acquire, manage and disseminate relevant data to every stakeholder, from medical practitioners to patients and carers (Groves et al. 2013). The type and complexity of medical condition(s) has a direct impact on the volume of data accumulated. Groves et al. (2013) clearly identify this as a Big Data scenario. Reduced expenditure and improved patient outcomes are means of value generation in healthcare. The five pathways to value generation; right living, right care, right provider, right value and right innovation are empowered by the use of Big Data analytics (Groves et al. 2013). These value pathways are well-positioned with recent advances in clinical decision-support (CDS). CDS has evolved from medical data processing tools to complex decision support frameworks and infrastructure for clinical

knowledge management. Two prominent taxonomies for CDS architectures are presented by (Sen et al. 2012) and (Wright and Sittig 2008). In the former, the authors explore CDS architectures within the context of underlying technologies of information management, data analytics and knowledge management. In the latter, the authors define architecture as the form of interaction between related systems, with four distinct CDS phases; standalone, integrated, standards-based and the service model.

Further impetus for Big Data analytics arises from the role of clinical scientists in research into disease management. The declining role of clinical scientists in medical research has been identified as a potential reason for the critical gap (termed the 'valley of death' crisis) that lies between bench research and bedside treatment (Butler 2008). (Roberts et al. 2012) emphasise the increasingly important role of data and clinical scientists in translational research; research that converts laboratory discoveries into clinical interventions. Ambitious projects such as (Burton et al. 2007) aim to predict the likelihood of medical conditions based on molecular biomarkers derived from datasets of disease-related mutations as well as the simpler history-based approach proposed by (Davis et al. 2008) are further indications of the potential of Big Data.

We reflect on CDS architectures from the perspective of the updated IS participation theory (Markus and Mao, 2004), which allows to not just look at the necessary components of the BDA system, but suggest the roles and the level of participation expected to lead to successful system implementation and use.

The paper is organised as follows. The following section presents a theoretical underpinning for articulating the role of BDA in advancing user participation in development and use of such systems from the updated IS participation theory perspective. An account of Big Data analytics (BDA) is reported next with examples from healthcare, relevant technology components, use and integration. We then review the context of analytics in healthcare, scrutinize the complexities of applying Big Data analytics to healthcare data and propose a solution to address these complexities.

2 Value of BDA Platform from the IS Participation Theory Perspective

(Markus and Mao 2004) revisited traditional IS participation theory and proposed the key elements of a new theoretical framework that clearly articulates the key parameters for designing the systems "for users with users". In their view, the updated theory of IS participation, requires specific re-definition of system success, differentiation among actors' roles in systems development and implementation, and refinement of the concept of participation between each actor accordingly (Figure 1).

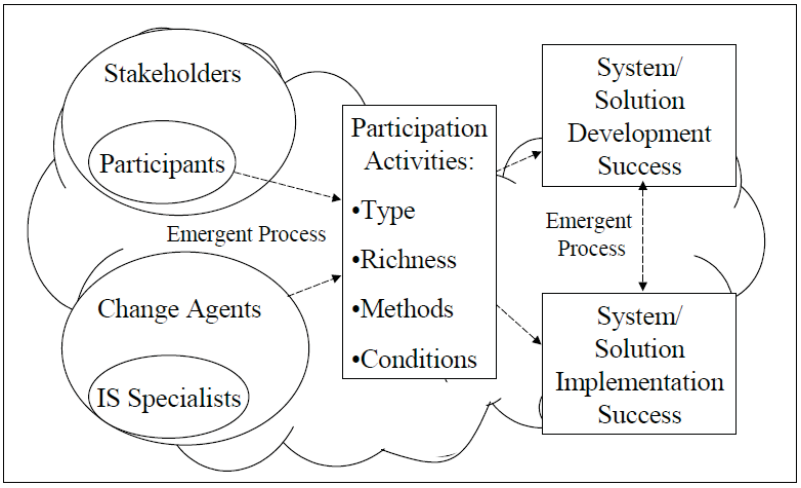

Figure 1: Updated Participation Theory (adapted from Markus and Mao, 2004)

The motivation of their work was to specify the effects of participation on various types of outcomes that are clustered together as system success. They identify three theories linking participation to system success (buy-in, system quality and emergent interactions) and determine conceptual gaps in these links that lead to the updated theory. Foundations of the updated theory lie in: 1) the distinction of system success into two concepts: system development success and system implementation success, with emergent reciprocal relations between them; 2) the description of groups of actors including stakeholders where participants are a subgroup, and change agents where IS specialists are a subgroup; 3) a reformulated behavioural concept of participation activities, characterized in terms of type and richness, methods and conditions; and 4) the hypothesis of emergent causal processes (Markus and Mao, 2004).

Following on from these foundations, (De Silva et al. 2013) adopted other dimensions of participatory design processes as suggested by (Bergvall-Kåreborn et al. 2010), e.g. "Designing for users" and "Designing with users". The latter provides an opportunity for closer engagement between the IS developers and users. Such participatory design process assumes the users' "voice" to be fully appreciated and better understood, together with the new opportunities that flow from full articulation of their needs in an act of active engagement from the planning phase to implementation, and commercialisation of the final product. In this sense new tools and techniques can be employed for "tracing user needs" by continuously monitoring their behaviour both implicitly and through shared data management, as well as by making users express their feedback as part of the systems deployment process.

Having a BDA platform increases stakeholder participation in healthcare. It is not only the domain experts and technology experts but also other stakeholders, the patients, carers, advocacy groups and regulatory bodies that can partake in healthcare delivery and management. To this end, the updated IS participation theory proposed by De Silva et al, (2013) and contextualised in a participatory information management framework to facilitate patient-centred care is even more relevant to the case of BDA as it deals with even greater complexity of data and has to address the dynamic and diverse needs of multiple stakeholders. Foundations of the updated participation theory lie in: 1) the distinction of system success into two concepts: system development success and system implementation success, with emergent reciprocal relations between them; 2) a broad engagement of the stakeholder groups where participants are a subgroup, and change agents where IS specialists are a subgroup; 3) a reformulated behavioural concept of participation activities, characterized in terms of type and richness, methods and conditions; and 4) the hypothesis of emergent causal processes (Markus and Mao, 2004). The key participation activities, type, richness, methods and conditions are enriched by the BDA platform, which leads overall to improved healthcare outcomes.

3 Big Data Analytics

This discipline originated as Business Intelligence (BI) in the 1990s but was later renamed to business analytics at the turn of the century to reflect the major contribution of analytics and a separation from information management. More recently, Big Data and BDA have been introduced to comprehend the changing nature of data (Davenport 2006). The formal definition by (Gartner) clearly notes these changes, Big Data is high-volume, -velocity and -variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making (Beyer and Laney 2012). Frequently known as the 3Vs, volume (the large quantities of data), velocity (speed of input/output, rate of change) and variety (different types of data from varied sources) are also used as quantifiers for applying analytics to Big Data. BDA aims to derive value from datasets quantified by these parameters. Several other parameters have been introduced to account for further intricacies of Big Data. Table 1 presents definitions of these concepts along with examples from healthcare.

Parameter	Description	Examples from healthcare
Volume	Size of data	Cohorts of patients, multiple conditions and treatment plans
Variety	Different formats and types (numbers, images, text)	Medical, clinical and omics data and images from patients with diverse conditions
Velocity	The rate at which data arrives and changes (streams, batches, infrequent intervals)	Wearable sensors and diagnostics transmitting patient behaviour
Veracity	Unpredictability of innately imprecise data types	Patient feedback and clinician notes on patient's state
Variability	Different interpretations of the same data	Clinical data on the same condition affecting a diverse group of patients
Value	Inherent value addition to the organisation against the costs to acquire/accumulate.	Extent of value addition to clinical decision-support and translational research
Sparseness	Low density of useful content (missing or null values)	Variability of patient feedback on symptoms and progress
Complexity	Hierarchies, linkages between entities and recurrent data structures	Multi-pharmacy and multi-morbidity

Table 1. Characteristics of Big Data

(Chen et al. 2012) conducted a comprehensive review of business intelligence and analytics (BIA) evolution from structured content (BIA 1.0), unstructured content (BIA 2.0) to mobile and sensor-based content (BIA 3.0). They signify smart health and wellbeing as a promising and high-impact BIA application, alongside ecommerce, e-government, science and technology and security and public safety. In healthcare, they identify patient data and genomics as the two main sources of Big Data. These two sources pair very well with the expectations of CDS and translational research. Privacy preservation and ethical research are highlighted as challenges to knowledge discovery from healthcare Big Data. Interestingly, the nature of data largely determines phases of evolution while the analytics techniques remain fairly consistent. It is this nature of data that gives rise to complexities in healthcare BDA. BDA also presents a paradigm shift in computing with the innovation of independent and self-managed components that can be infinitely scaled to suit large volume computations, without the complication of shared resources.

The computing platform most often used for BDA is Apache Hadoop (https://hadoop.apache.org/); it consists of a distributed file-system (Hadoop Distributed File System; HDFS) and a robust programming model (MapReduce) that work together to distribute the algorithm to the data through task assignment (instead of the traditional approach of data into the algorithm) and to create schemas/ data models at run-time (instead of a static schema). The general-purpose Hadoop stack is illustrated in Figure 2.

The initial layer is for data import, with two packages, Flume designed to extract data from file-based systems (free-form text, log files, etc) and Sqoop, designed to extract structured content from enterprise systems and relational databases. Next, data moves to the storage layer, HDFS, which introduces redundancy and fault-tolerance in preparation for scalable computations. MapReduce introduces the computation in the form of Map and Reduce functions to distribute the computation across many nodes and then merge the results into a single set of outputs. On the same layer, HBase provides a database-style interface to HDFS to deploy programs that can read or write to specific subsets of data.

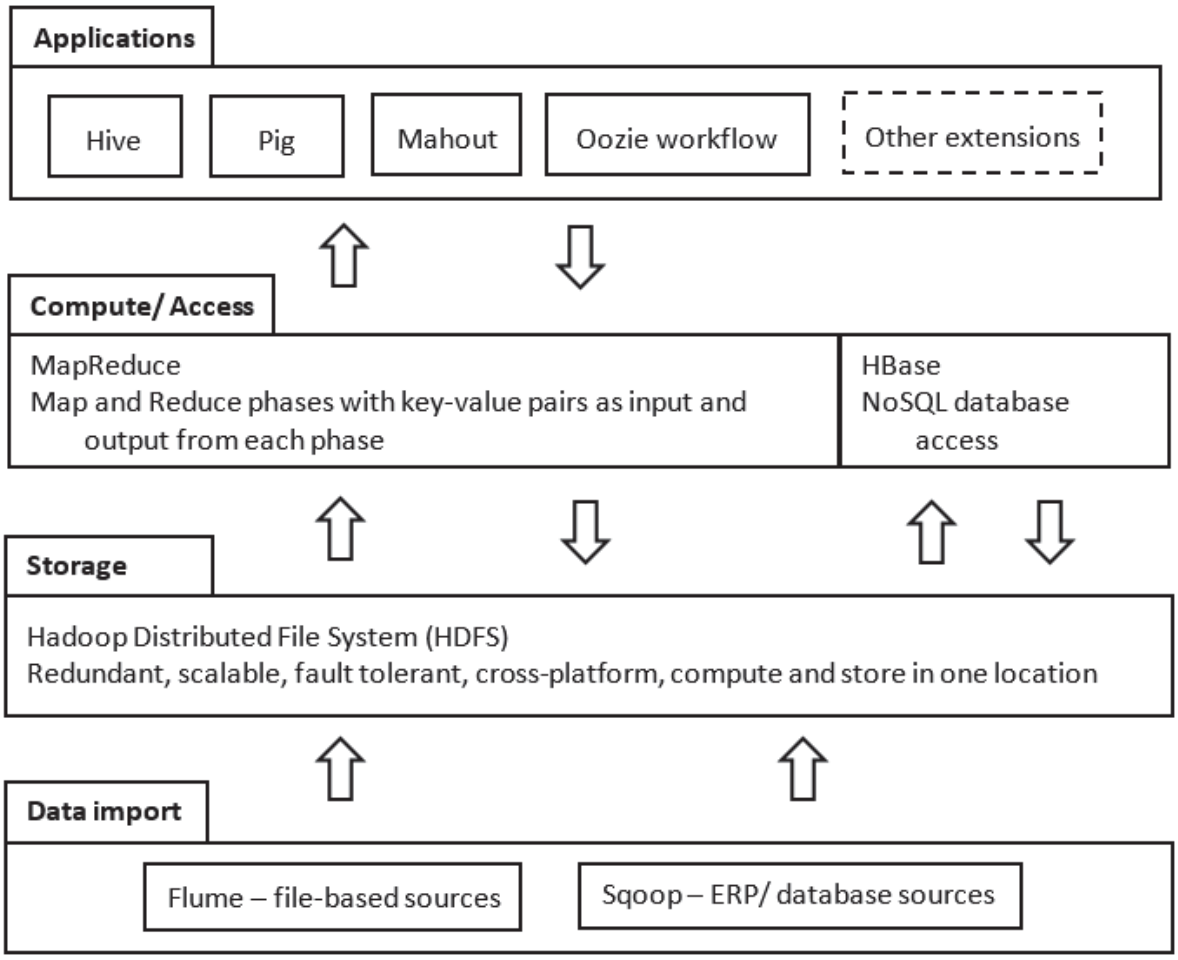

Figure 2. Apache Hadoop Ecosystem

The application layer presents end-user abstraction. Converting any computation to MapReduce format is complicated, the application layer abstracts this with a set of purpose-built packages; Hive supports declarative queries for warehouse-type batch operations, Pig provides a compiler for sequences of MapReduce sub-modules and Mahout a package of machine learning techniques designed to use Hadoop for scalable analysis of Big Data. Oozie is a workflow/coordination scheduler system, which can be programmed to execute several different jobs (Hive, Pig, Mahout) in sequence. Other extensions such as R connectors to Hadoop are available as early trials at the time of writing. Most major software vendors provide extended implementations of Hadoop with improvements for faster deployment and/or simplified management.

As a discipline related to healthcare, bio-informatics has been quick to utilise the strength of Hadoop with extended frameworks such as CloudBurst and Crossbow. These have been able to achieve record speeds for next generation sequencing via read mapping (Schatz 2009). However in contrast, Big Data management and analytics has had limited impact on healthcare. Adoption of electronic health records (EHR) in clinics and hospitals can be the much needed platform for BDA projects. (Jensen et al. 2012) present a complete review of the key issues in accumulating EHR data and integrating these with genetic data. (Hanauer et al. 2011) performed a simple symptom–disease–treatment association rule mining on a large collection of EHR through which they were able to identify clinically relevant and accurate associations for seven distinct diseases. Further impetus towards BDA is through large-scale online data collection projects driven by crowd intelligence, such as Daily Strength (http://www.dailystrength.org/) and PatientsLikeMe (http://www.patientslikeme.com/). BDA outcomes from the accumulated data can be a useful resource for patient empowerment

(Miller 2012). Other prominent examples do not link directly into BDA but through a support technology such as mobile phones or wearable sensors. Ginger.io is a BDA application (Ginger.io 2015) that utilises mobile sensors present in smartphones in order track, record and analyse physical movements, phone activity (such as calls, texts) in order to monitor behavioural health therapies for chronic illnesses. Another BDA application in healthcare, (PropellerHealth 2015) utilises a GPS-enabled tracker to monitor inhaler usage by asthmatics. The information collected is used to determine behaviours across populations and also integrates this information with known asthma catalysts to improve treatment and prevention of asthma. The lack of BDA having a direct impact on healthcare can be attributed to the complexities delineated in the following section.

4 Analytics in Healthcare

Analytics in healthcare is driven by the gradual shift from disease-centred to patient-centred care (PCC). PCC models transcend traditional boundaries that isolate patients from their clinical context. Patient-centred care was first featured in healthcare as one of the six aims for high-quality healthcare in a report 'Crossing the Quality Chasm' published by the USA Institute of Medicine (Bloom 2002). This report defines PCC as care that is "respectful of and responsive to individual patient preferences, needs, and values, and ensuring that patient values guide all clinical decisions". In the traditional approach, the clinician addresses the medical condition and thereby cures/improves the health of the patient whereas with PCC this becomes a shared responsibility between healthcare professionals, the patient and family members. Frequent communication and information sharing leads to a further accumulation of data besides the actual clinical data and can be used effectively to understand patients with similar circumstances.

Gartner's business analytics framework (Chandler et al. 2011) is a useful tool to identify the role of analytics in healthcare. The framework can be presented as a matrix to highlight its key elements (Table 2). The primary activities 'enable', 'produce' and 'consume' run across the main entities of 'people', 'processes' and 'platform'.

	People	**Processes**	**Platform**
Consume	Decision makers	Decision processes	Decision capabilities
Produce	Analysts/Data scientists	Analytic processes	Analytics capabilities
Enable	ICT Administrators	Information governance	Information capabilities

Table 2. Business analytics framework (Chandler et al. 2011)

BDA maintains its potential to fulfil the expectations of the 'platform' element. It can become the patient-centred healthcare platform for information management, analytics and decision-making capabilities. The following complexities need to be addressed for the eventuation of this platform. Based on our investigations, the key complexities impacting BDA in healthcare are granular data accumulation, temporal abstraction, multimodality, unstructured data, and integration of multi-source data. They are distributed across the 'enable, produce, consume' activities noted above. These will be explored bottom-up so that low-level explanations contribute to understanding at the high-level.

5 Granular data accumulation

From a general practitioner's desktop computer to cardiac monitors in an emergency room, a multitude of clinical information systems capture patient information. This information exists at different levels of granularity, in diverse formats and recorded at varying frequency. For instance, sensor readings from a cardiac monitor are well-defined in terms of grain, format

and frequency, however blood glucose measurements, although in the same format, can vary in terms of grain and frequency. A patient can record blood glucose levels at different times during the day when at home whereas a clinic may capture a single measurement but derive a different measure (glycated haemoglobin) to determine the three month average. This difference in granularity can be an extra dimension of information for BDA when paired with medication, demographic or behavioural information. Another example is recording medication information along with medical imaging. BDA can be used to identify changes in medical images and relate these to changes in medications or dosage. BDA is not limited by volume, velocity or variety so it is pertinent to capture and accumulate information at all levels of granularity.

Granular data accumulation also extends to capturing outcomes and feedback. Outcomes from a specific exercise routine, dietary modifications or change of medication need to be captured and recorded. Completeness of medical data from start to end of the patient lifecycle is crucial for successful BDA in translational research. It is equally important to capture patient feedback as this reflects their experience of the medical condition. BDA can be used to identify associations between outcomes, feedback, symptoms, medication and behavioural changes but the quality of the findings is heavily dependent on the completeness of the data accumulated.

6 Temporal abstraction

Time is the supporting dimension for granularity as patient information is collected over time. Temporal abstraction (TA) is defined as a process which takes in a set of time-stamped parameters, external events and abstraction goals to generate abstractions of the data to be used for interpretation of past/present data (Shahar 1994). The intention is to transform temporal data from a simple, numerical form to informative and qualitative descriptions, which can be understood by clinicians. (Stacey and McGregor 2007) present a comprehensive review of techniques used in clinical data analysis systems. They highlight several inadequacies, such as confinement to temporal trends and level shifts, limited dimensionality of abstraction output and lack of integration with other analysis outcomes. Complex TA is a further development to represent higher-level abstractions not directly from the data but from intermediate TA outcomes. (Keravnou 1997) describes complex TAs as compound time objects with a repeating element, a repeating pattern and progression pattern where the repeating element itself could be periodic. (Bellazzi et al. 2000) developed a complex TA in the domain of diabetes monitoring to detect two overlapping interval abstractions and also the successive increase and decrease of blood glucose levels. Another well-researched approach to complex TAs is to represent the changing nature of a timer series using characters and words as symbols. (Sharshar et al. 2005) separated a data stream into trend and stability for each data point and applied rules to convert the signal into transient, decrease, increase or constant categories. A further abstraction was applied when individual characters, defining the state of the signal, are merged to form words. These 'words' can be mined for clinically relevant associations and projections. TA and complex TA effectuate summarisation of highly granular data to abstract clinically relevant representations recorded over time for conventional data mining techniques. However with the advent of BDA the clinical knowledge embedded into TA of the same becomes equally useful to navigate the large space of granular data accumulation.

7 Multimodality

Mode is a resource for sense-making, it introduces context to an entity. Many contexts/modes can be found in a clinical environment to represent and also identify a patient. Modes such as demography, behaviour, symptoms, diagnosis, blood-gas measurements and medication are indicative of the patient's condition, disease trajectory and future well-being. Clinical decision-support greatly benefits from this multimodal representation of the patient's state. Capturing the multiple modes of an entire sample of patients is an equally rich resource for inference and prediction in translational research. Multimodality in business analytics is frequently addressed by data warehousing technologies. Despite its prevalence in many industries, its adoption by medical organizations has been limited. Early implementations of clinical data

warehousing (CDW) were aimed at solving specific clinical problems. There are still a few recent case studies that demonstrate the applicability of data warehouse concept in medical domain. For example, (Wisniewski et al. 2003) describe the use of a data warehouse for hospital infection control. It was populated with data from three hospitals and demonstrated to be useful for measurement of antimicrobial resistance, antimicrobial use, the cost of infections, and detection of antimicrobial prescribing errors. (Chute et al. 2010) present a review of the Enterprise Data Trust at the Mayo Clinic, which is a collection of all electronic data organized to support information management, analytics and high-level decision-making. In recent research endeavours (Hu et al. 2011; Lowe et al. 2009) have proposed and implemented data warehousing solutions to address the information needs of translational research, which can among other functions integrate pathology and molecular data with a clinical data model to support a breast cancer translational research program (Hu et al. 2011). STRIDE (Stanford Translational Research Integrated Database Environment) is an informatics platform for clinical and translational research. It consists of a data management system, a CDW and a development framework for new applications.

The complexity with healthcare data arises when designing a suitable dimensional model to encompass the variety of information (demographic to clinical) and type of information (structured and unstructured) accumulated.

8 Unstructured data

Healthcare data is inundated with unstructured content consisting mainly of textual records ranging from clinician comments to patient feedback. Textual records of this nature can exist at all levels of granularity noted earlier. Patient feedback can be collected before and after a surgery, during a period of new medication or a behavioural change. In addition to text, other unstructured formats include images, audio/video recordings and associative datasets. Given the inherent structure of the discipline, a majority of clinical text can be associated with well-formed ontologies. However, textual records received from patients need to be interpreted using a trusted knowledge base. The design and development of such user-warrant ontology becomes a complex task given the variety of terms that can be used to refer to medical conditions and symptoms. (Nguyen et al. 2014; Prakasa and De Silva 2013) have studied the extraction of an ontology from end-user feedback in healthcare. Many of the Big Data technologies have been developed to address the complexities of unstructured data. The central data models are key-value pair (KVP), column family stores, document databases and graph databases (Chen et al. 2012). The lack of a fixed schema makes these data models flexible and scalable, however they do not follow standard properties found in relational databases. Each type of unstructured data can be analysed independently for patterns of interest, however complexities arise when outcomes from such disparate sources needs to be integrated, in part or whole.

9 Information fusion

Information fusion is a widely researched field, which is an efficient method for automatically or semi-automatically transforming information from different sources and different points in time into a representation that provides effective support for human or automated decision making (Boström et al. 2007). Focus is largely on the transformation of information, which includes means for combining and aggregating to infer as well as reduce information. Much research has been conducted on multi-sensor fusion (Khaleghi et al. 2013), image fusion (Goshtasby and Nikolov 2007) and web information fusion (Yao et al. 2008). Independently, intelligent fusion techniques have been developed based on post-perceptual integration phenomena (Torra 2003) and cross-modal influence strategy (Coen 2005). In healthcare, levels of granularity and the temporal factor need to be well aligned with the purpose of information fusion. Structured and unstructured healthcare data accumulated at different levels of granularity will be processed by BDA within the mutually exclusive modals to generate analytics outcomes. Although these outcomes are beneficial on their own, they can be further fused to create a comprehensive account of a patient (for clinical decision support) or a medical

condition (for translational research). The computational requirements of such large-scale information fusion can be efficiently handled by Big Data technologies, however the complexities of the data model for effective information fusion is largely unaddressed.

10 The Analytics Team

It is pertinent to briefly discuss the analytics team most suited for BDA applications in healthcare. The team broadly consists of domain experts and analytics experts. A clinical scientist will be the primary contributor of domain expertise. Having a thorough knowledge of both the nature of medical data and expected outcomes, a clinical scientist can guide the knowledge discovery process. A physician is a preferred secondary role to further improve the domain expertise. The analytics experts will comprise a data scientist, a data analyst, a software developer and a project manager. Depending on the size of the organisation and the analytics effort, the team can expand or take multiple roles. The data analyst is generally in charge of data extraction and processing following which the data scientist and analyst will team together for the analytics phase. A software developer is required for new interface development or integration with existing systems. Another useful role for the analytics team is that fulfilled by a simplified version of the Business Intelligence Competency Centre (BICC) commonly found in large-scale analytics projects (Hostmann et al. 2006). The BICC is responsible for the strategic plan, prioritisation, data quality, governance and uptake of analytics outcomes in business activities. These responsibilities can be also assigned to the roles identified earlier. The next section presents a case study of BDA platform design and implementation as part of the Diabetes Screening Research Initiative (DiScRi) project with a review of outcomes and their benefits (Burstein et al. 2013).

11 Addressing the complexities: the case study

A prototypical trial was conducted on data accumulated by the Diabetes Screening Complications Research Initiative (DiScRi) run at a regional Australian university (Jelinek et al. 2006). It is a diabetes complications screening program in Australia where members of the general public participate in a comprehensive one-stop health review. The screening clinic has been collecting data for over ten years and includes close to a hundred features including demographics, socio-economic variables, education background, clinical variables such as blood pressure, body-mass-index (BMI), kidney function, sensori-motor function as well as blood glucose levels, cholesterol profile, inflammatory markers, oxidative stress markers and use of medication. The dataset is reflective of typical Big Data accumulation in a clinical environment to be used for both clinical decision support and translational research. Figure 3 illustrates the solution used to address the stated complexities to BDA.

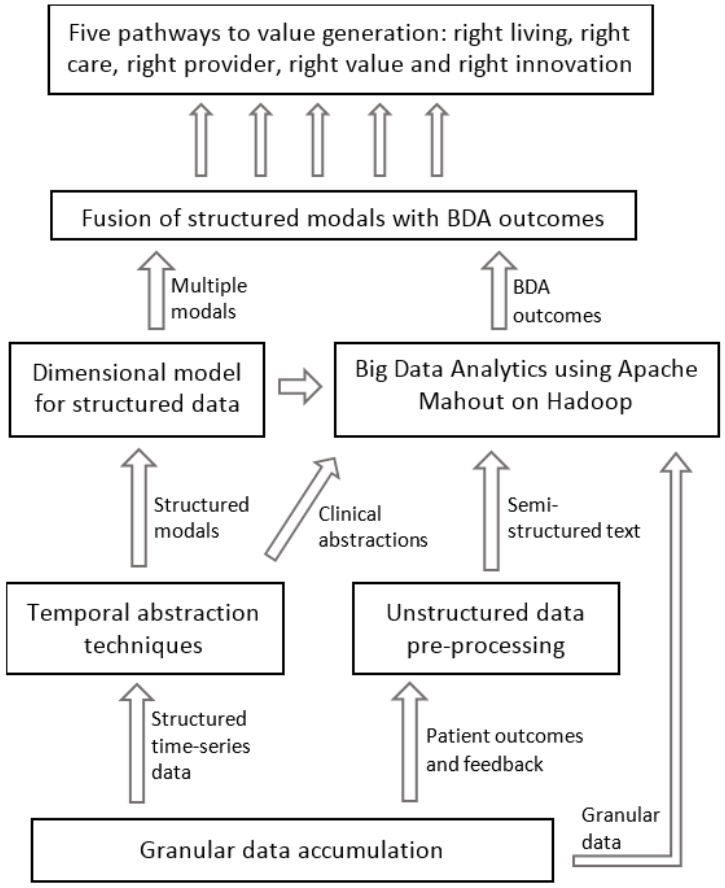

Figure 3. BDA solution - DiScRi experiment

At the lowest layer, data is accumulated at varied grains. High level data such as patient demographics and diagnosis are recorded by medical practitioners while low level blood glucose measurements, blood pressure, body mass index and others are recorded by lab technicians at the screening clinic and by patients and carers in their own homes. Structured data is maintained in a relational database while unstructured data, mainly free-flow text such as patient feedback and clinician notes, is saved in flat files. Clinical scientists derive standard statistics (patient numbers, average age, average clinician visits etc.) from the granular data for routine reporting. In the secondary layer, time-series data is fed into temporal abstraction techniques in order to embed clinical knowledge into the sequence of data points. Blood pressure recordings, blood glucose levels and sensori-motor function data are transformed in this manner. Free flow text is pre-processed to a format suitable for analysis; stop word removal, lemmatisation and named entity recognition are conducted in this phase. Optionally, a verified ontology as specified by (Nguyen et al. 2014; Prakasa and De Silva 2013) can be used to add context to patient comments and feedback. The third layer consists of two functions, dimensional modelling and the actual BDA. Temporal abstraction generates structured information with embedded clinical knowledge. For instance the number of visits per patient and the outcomes of each such visit are useful to determine the patient's current condition and disease trajectory. Temporal abstraction is necessary to extract this information that occurs over time at different stages of the medical condition. Integrating time-dependent information with static data (such as demographics) necessitates a multi-dimensional structure. This inherent structure can be captured in a dimensional model and implemented as a data warehouse.

A novelty in this dimensional model is the cardinality dimension. Individual patients attend the screening clinic multiple times, the cardinality dimension captures each visit. Except for Personal Information (which contained attributes such as gender, family history of medical

conditions), all other dimensions were composed of attributes recorded for each visit/test. While the fact table would distinguish between records, the cardinality dimension was necessary to distinguish between patients. Figure 4 presents the DiScRi dimensional model.

BDA is the second function. The primary input is semi-structured text data containing patient feedback and clinician notes. Clinical abstractions and granular data form the secondary input.

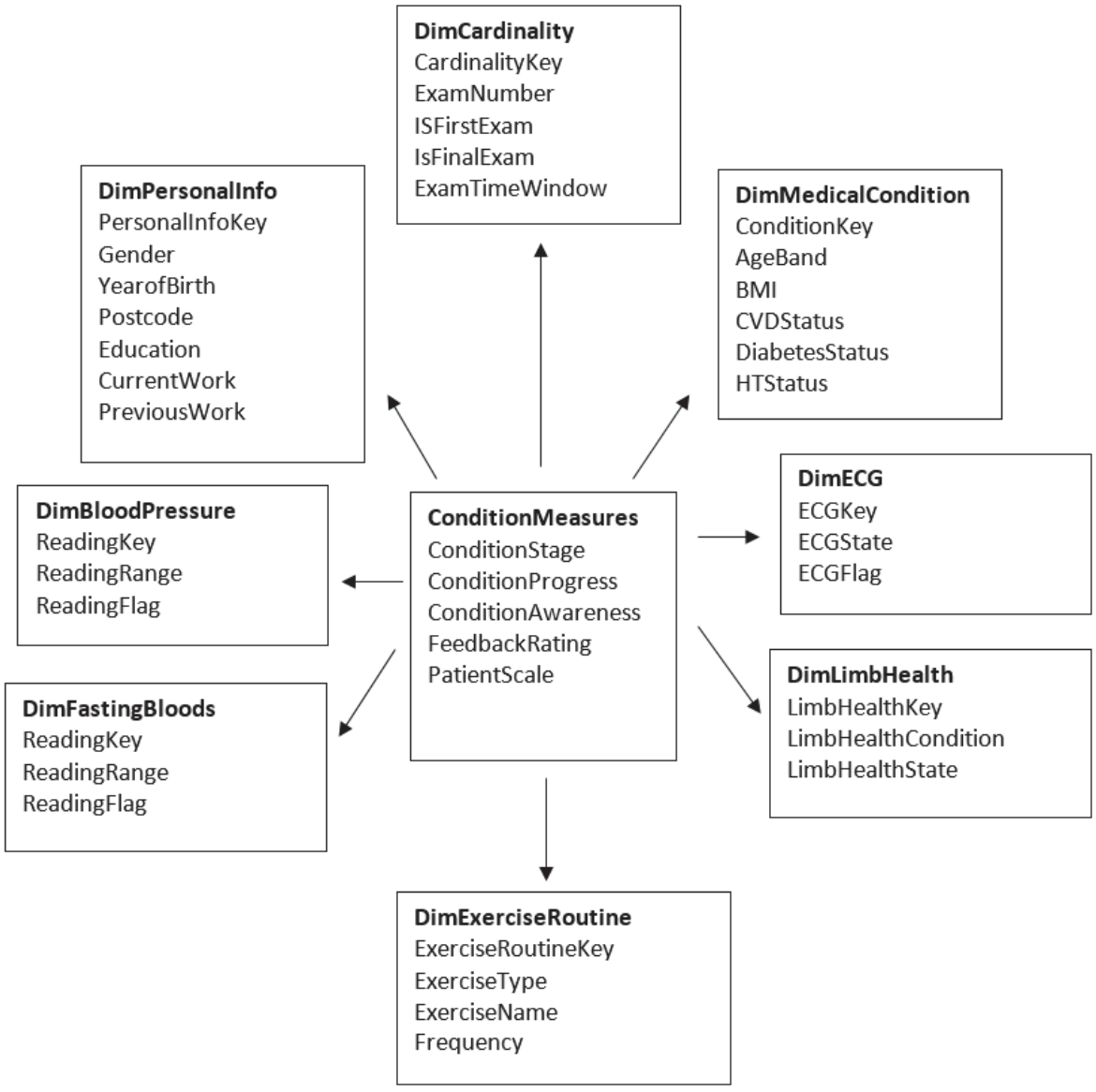

Figure 4. DiScRi dimensional model

The analytics outcomes from BDA range from clustering, classification, association rules, predictions, summarisation and visualisation. This primary output is fused with the outputs from clinical abstractions and models of structured information. For instance, the association of patient age with feedback provided can be understood by fusing the BDA outcome with the demographic dimension in the warehouse. Similarly, fusion of other structured with the unstructured information leads to expedient insights on patients and their conditions. These insights can lead to the five pathways of value generation in healthcare noted earlier, right living, right care, right provider, right value and right innovation.

The following figures present an interesting instance of fused outcomes. Figure 5 illustrates the use of a temporal abstraction scheme and drill-down queries on the data warehouse to extract

information on hypertension from granular data. Patients with hypertension were identified by their age groups and by the number of years since diagnosis of hypertension. The drill-down feature in age groups detects a significant drop in the number of 5-10 year hypertension cases in the age sub-groups of 70-75 and 75-80.

Separately, clinician notes were clustered using Apache Mahout running on a Hadoop instance. The 70-80 age group was found to be closely grouped and further intra-cluster exploration uncovered a distinction in multi-word term association. The same group with a drop in 5-10 year hypertension linked into multi-word term associations related to renal conditions. Fusion of information from the structured and unstructured analysis led to this outcome.

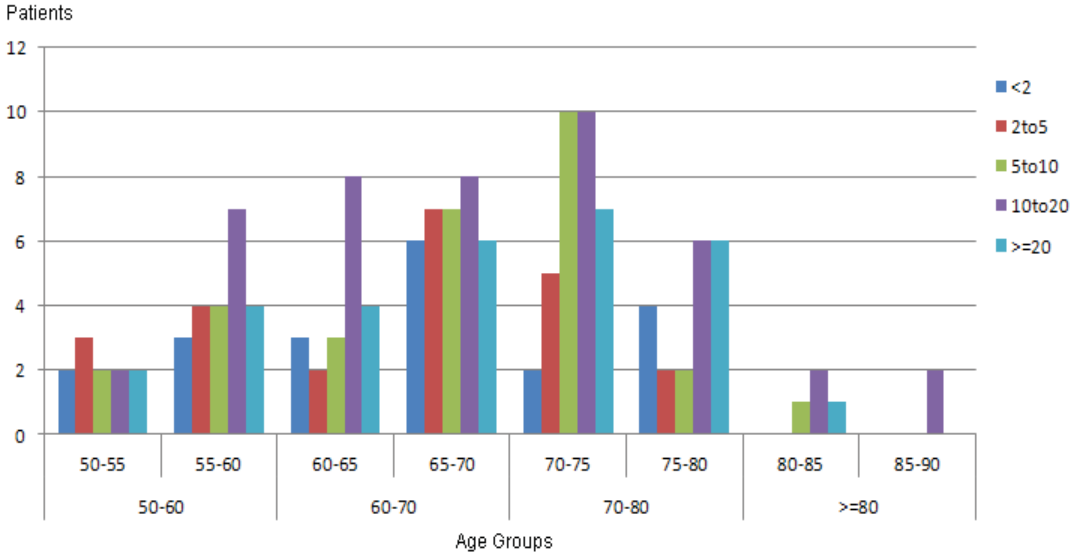

Figure 5. Distribution of number of years since diagnosis of hypertension by age groups.

Although one would normally suggest that the drop in the number of cases in the 80+ age group with greater than 20 years hypertension is largely due to mortality, the current BDA has found that this is possibly related to renal disease. Results from this large screening study provide further evidence for the role of BDA in translational research.

Figure 6 indicates that urinary and kidney disease play a role in the reduction of people over 80 years attending screening clinics due to medical care moving from observation as is the case in a screening clinic to intervention in primary health care.

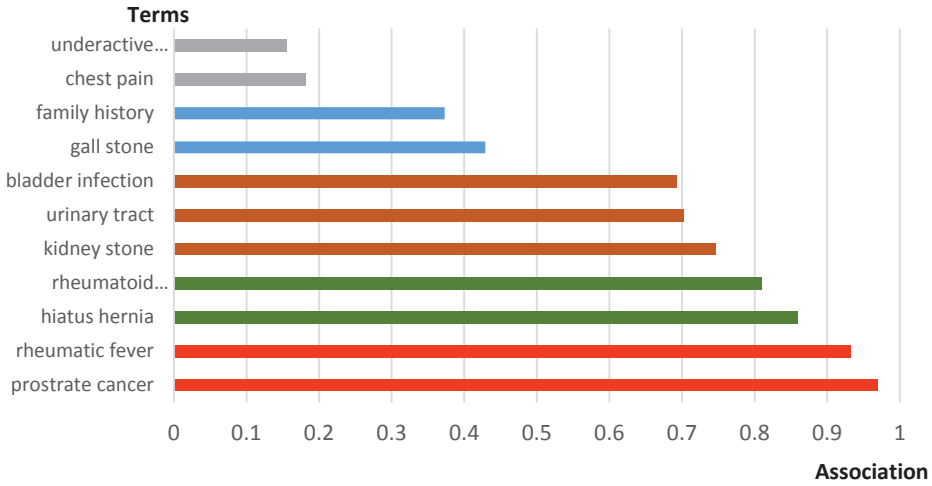

Figure 6. Multi-word term association for 70-80 age group cluster, with the sub-group affected by urinary tract infections coloured in brown.

This highlights the necessity to include more urinary and renal testing in screening clinics to identify these conditions early in the older age group and provide timely personalised intervention.

The aforementioned outcomes from temporal abstraction, information fusion and text analytics can be further scrutinised using conventional data mining techniques such as classification, association and Bayesian techniques. An ideal scenario that can be realised is the capacity to predict the likelihood of clinical outcomes based on current knowledge of the patient and similar patients' information.

12 Discussion

Revisiting the three entities of the business analytics framework (Chandler et al. 2011), people, processes and platform, it is evident that BDA maintains capacity to fulfil the expectations of the platform entity across the three primary activities of consume, produce and enable. Big Data in healthcare demonstrates distinct characteristics in comparison to other domains. Each patient's healthcare record represents a temporal sequence of events, and each event maintains a number of attributes (of varying modalities) captured over the duration or a specific point in time. The complexities delineated in this paper arise due to this distinctive nature of healthcare Big Data. As reflected in the results reported from our experiments, expedient insights can only be derived in collaboration with the domain expert – a clinical scientist in our case study. Access to the adequately designed BDA platform addresses these complexities, and allows domain experts to interrogate the data and dynamically manipulate hypotheses formulation, thus fulfilling the requirements for decision support. .

The evolving role of CDS into clinical knowledge management is well positioned to leverage on the capabilities of such a BDA platform (Burstein et al, 2013). The fusion of highly structured domain knowledge with insights derived from unstructured patient data, for which no fixed schema can be developed in advanced, is a significant contribution towards updated clinical knowledge. The BDA platform addresses complexities arising due to the nature of Big Data described in Table 1. The parameters of volume, variety, veracity, sparseness and complexity are exemplified in the reported experiment. Resolving the Big Data complexities in this manner lays a robust foundation for a complete business analytics framework as shown in Table 2 with improved processes and empowered people.

From the IS Participation theory perspective, online accessibility to real-time patient information, analytics outcomes and updated clinical knowledge leads to increased stakeholder participation in healthcare delivery and management. Participation activities of type, richness, methods and conditions of all stakeholders from healthcare professionals, caregivers to patients themselves are empowered by access to relevant and reliable information. Therefore it is very convincing that the five pathways to value generation are sustained by the strong foundation of a BDA platform.

13 Conclusion

Significant clinical knowledge and a unique understanding of disease patterns can be acquired by utilising Big Data technologies for data management and analytics in the healthcare discipline (Chen et al, 2012). The nature of data and expectations of the healthcare professional has led to complexities when developing this inter-disciplinary focus area. The primary aim of this paper was to examine these issues in-detail from the perspectives of the participation theory (Markus and Mao, 2004) and illustrate how a potential solution was applied to the Diabetes screening case.

The paper presented the Big Data paradigm, supporting technologies and their relevance to BDA in healthcare. Thereafter, complexities to BDA in healthcare were discussed in detail followed by a proposed solution that addresses these complexities and unifies structured and unstructured data collected within a healthcare environment. The method and outcomes from a prototypical trial conducted on the DiScRi project were presented to implicate the

significance and contribution of the proposed solution. The inherent limitations of most data collections, such as missing data, null values, incorrect values and unmatched records were observed and accounted for in the BDA process. The fusion of structured and unstructured data is aptly demonstrated in the outcomes and is of significant value in a clinical context. To enhance translational research, data obtained from the annual screening clinic needs to be interpreted in terms of outcome measures following diverse treatment options present in the screening cohort. New associations between personal health status, intervention and individual outcome that also reflects a wider population use is an essential part of current healthcare research. Following the updated participatory theory led to engaging domain experts in the process of BDA platform design and implementation. We also confirmed that IS Specialists played a role of Change Agents in facilitating the adoption of BDA platform and leading to a better appreciation of the possibilities of data analytics for clinical decision support through this emergent process of engagement.

BDA architecture for healthcare applications should overcome the complexities of granular data accumulation, temporal abstraction, multimodality, unstructured data and integration of multi-source data in order to provide a robust platform for effective workflows and improved engagement. The stakeholders of the BDA solution have to be clearly identified and fully involved in the process of BDA platform design. Their role will be in defining the boundaries and expected outcomes of the platform, identifying the right data sources, transformation of unstructured data, integration with structure content and developing analytics pathways for CDS. In our case study their participation led to the emergent success of the BDA platform as it overcame the limitations to unified data exploration and led to analytics outcomes previously unknown to stakeholders. We also confirmed that strategic benefits of BDA are not visible upfront. Instead, an overall success of data analytics is generated by long-term use and appears of significant value in terms of knowledge management in a clinical setting. The authors are currently involved in expanding this work to include real-time sensor data into the BDA process to further empower clinical decision support. Future work can include expansion of this approach to other complex medical conditions which record a multitude of data points, such as cognitive decline (Alzheimer's disease) and psychosis (schizophrenia and depression).

References

Bellazzi, R., Larizza, C., Magni, P., Montani, S., and Stefanelli, M. 2000. "Intelligent Analysis of Clinical Time Series: An Application in the Diabetes Mellitus Domain," *Artificial intelligence in medicine* (20:1), pp. 37-57.

Bergvall-Kåreborn, B., Howcroft, D., Ståhlbröst, A., and Wikman, A. M. 2010. "Participation in Living Lab: Designing Systems with Users," in *Human Benefit through the Diffusion of Information Systems Design Science Research*. Springer, pp. 317-326.

Beyer, M. A., and Laney, D. 2012. *The Importance of 'Big Data': A Definition*, Stamford, CT: Gartner).

Bloom, B. S. 2002. "Crossing the Quality Chasm: A New Health System for the 21st Century," JA*MA: The Journal of the American Medical Association* (287:5), pp. 646-647.

Boström, H., Andler, S. F., Brohede, M., Johansson, R., Karlsson, A., Van Laere, J., Niklasson, L., Nilsson, M., Persson, A., and Ziemke, T. 2007. "On the Definition of Information Fusion as a Field of Research," *IKI Technical Reports*, HS- IKI -TR-07-006, pp.8-14.

Burstein, F., De Silva, D., Jelinek, H. F., and Stranieri, A. 2013. "Multivariate Data-Driven Decision Guidance for Clinical Scientists," *IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*, 2013: IEEE, pp. 193-199.

Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D. P., McCarthy, M. I., Ouwehand, W. H., and Samani, N. J. 2007. "Genome-Wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls," *Nature* (447:7145), pp. 661-678.

Butler, D. 2008. "Translational Research: Crossing the Valley of Death," *Nature News* (453:7197), pp. 840-842.

Chandler, N., Hostmann, B., Rayner, N., and Herschel, G. 2011. *Gartner's Business Analytics Framework*, Gartner Research:G00219420).

Chen, H., Chiang, R. H., and Storey, V. C. 2012. "Business Intelligence and Analytics: From Big Data to Big Impact," *MIS Quarterly* (36:4), pp. 1165-1188.

Chute, C. G., Beck, S. A., Fisk, T. B., and Mohr, D. N. 2010. "The Enterprise Data Trust at Mayo Clinic: A Semantically Integrated Warehouse of Biomedical Data," *Journal of the American Medical Informatics Association* (17:2), pp. 131-135.

Coen, M. H. 2005. "Cross-Modal Clustering," *Proceedings Of The National Conference On Artificial Intelligence*: Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, p. 932.

Davenport, T. H. 2006. "Competing on Analytics," *Harvard Business Review*: 84, pp. 98-107, 134.

Davis, D. A., Chawla, N. V., Blumm, N., Christakis, N., and Barabási, A.-L. 2008. "Predicting Individual Disease Risk Based on Medical History," *Proceedings of the 17th ACM conference on Information and knowledge management*: ACM, pp. 769-778.

De Silva, D., Burstein, F., Stranieri, A., Williams, K., and Rinehart, N. 2013. "A Participatory Information Management Framework for Patient Centred Care of Autism Spectrum Disorder," *24th Australasian Conference on Information Systems (ACIS)*: RMIT University, pp. 1-11.

Gartner. *Gartner It Glossary - Big Data*. Retrieved 1 Mar, 2015, from http://www.gartner.com/it-glossary/big-data

Ginger.io. 2015. *Ginger.Io for Individuals*. Retrieved 1 Mar, 2015, from https://ginger.io/

Goshtasby, A. A., and Nikolov, S. 2007. "Image Fusion: Advances in the State of the Art," *Information Fusion* (8:2), pp. 114-118.

Groves, P., Kayyali, B., Knott, D., and Van Kuiken, S. 2013. "*The 'Big Data' revolution in Healthcare*," New York (NY): McKinsey Global Institute.

Hanauer, D., Zheng, K., Ramakrishnan, N., and Keller, B. 2011. "Opportunities and Challenges in Association and Episode Discovery from Electronic Health Records," *IEEE Intelligent Systems* (26:5), pp. 83-87.

Hostmann, B., Rayner, N., and Friedman, T. 2006. *Gartner's Business Intelligence and Performance Management Framework*, Gartner Inc.

Hu, H., Correll, M., Kvecher, L., Osmond, M., Clark, J., Bekhash, A., Schwab, G., Gao, D., Gao, J., and Kubatin, V. 2011. "Dw4tr: A Data Warehouse for Translational Research," *Journal of Biomedical Informatics* (44:6), pp. 1004-1019.

Jelinek, H. F., Wilding, C., and Tinely, P. 2006. "An Innovative Multi-Disciplinary Diabetes Complications Screening Program in a Rural Community: A Description and Preliminary Results of the Screening," *Australian Journal of Primary Health* (12:1), pp. 14-20.

Jensen, P. B., Jensen, L. J., and Brunak, S. 2012. "Mining Electronic Health Records: Towards Better Research Applications and Clinical Care," *Nature Reviews Genetics* (13:6), pp. 395-405.

Keravnou, E. T. 1997. "Temporal Abstraction of Medical Data: Deriving Periodicity," in *Intelligent Data Analysis in Medicine and Pharmacology*. Springer, pp. 61-79.

Khaleghi, B., Khamis, A., Karray, F. O., and Razavi, S. N. 2013. "Multisensor Data Fusion: A Review of the State-of-the-Art," *Information Fusion* (14:1), pp. 28-44.

Lowe, H. J., Ferris, T. A., Hernandez, P. M., and Weber, S. C. 2009. "Stride–an Integrated Standards-Based Translational Research Informatics Platform," *AMIA Annual Symposium Proceedings*: American Medical Informatics Association, p. 391.

Markus, M. L., and Mao, J.-Y. 2004. "Participation in Development and Implementation-Updating an Old, Tired Concept for Today's Is Contexts," Journal of the Association for Information Systems (5:11), p. 14.

Miller, K. 2012. "Big Data Analytics in Biomedical Research," *Biomedical Computation Review*, pp. 14-21.

Nguyen, B. V., Burstein, F., and Fisher, J. 2014. "Improving Service of Online Health Information Provision: A Case of Usage-Driven Design for Health Information Portals," *Information Systems Frontiers*, (17:3), pp. 493-511.

Prakasa, A., and De Silva, D. 2013. "Development of User Warrant Ontology for Improving Online Health Information Provision," *24th Australasian Conference on Information Systems (ACIS)*: RMIT University, pp. 1-12.

PropellerHealth. 2015. *Propeller Health - the Leading Mobile Platform for Respiratory Health Management*. Retrieved 1 Mar, 2015, from http://propellerhealth.com/

Roberts, S. F., Fischhoff, M. A., Sakowski, S. A., and Feldman, E. L. 2012. "Perspective: Transforming Science into Medicine: How Clinician–Scientists Can Build Bridges across Research's "Valley of Death"," *Academic Medicine* (87:3), pp. 266-270.

Schatz, M. C. 2009. "Cloudburst: Highly Sensitive Read Mapping with Mapreduce," *Bioinformatics* (25:11), pp. 1363-1369.

Sen, A., Banerjee, A., Sinha, A. P., and Bansal, M. 2012. "Clinical Decision Support: Converging toward an Integrated Architecture," *Journal of Biomedical Informatics* (45:5), pp. 1009-1017.

Shahar, Y. 1994. "A Knowledge-Based Method for Temporal Abstraction of Clinical Data," *Technical Report* Stanford University.

Sharshar, S., Allart, L., and Chambrin, M.-C. 2005. "A New Approach to the Abstraction of Monitoring Data in Intensive Care," in *Artificial Intelligence in Medicine*. Springer, pp. 13-22.

Stacey, M., and McGregor, C. 2007. "Temporal Abstraction in Intelligent Clinical Data Analysis: A Survey," *Artificial Intelligence in Medicine* (39:1), pp. 1-24.

Torra, V. 2003. "On Some Aggregation Operators for Numerical Information," in *Information Fusion in Data Mining*. Springer, pp. 9-26.

Wisniewski, M. F., Kieszkowski, P., Zagorski, B. M., Trick, W. E., Sommers, M., Weinstein, R. A., and Project, C. A. R. 2003. "Development of a Clinical Data Warehouse for Hospital Infection Control," *Journal of the American Medical Informatics Association* (10:5), pp. 454-462.

Wright, A., and Sittig, D. F. 2008. "A Four-Phase Model of the Evolution of Clinical Decision Support Architectures," *International Journal of Medical Informatics* (77:10), pp. 641-649.

Yao, J., Raghavan, V. V., and Wu, Z. 2008. "Web Information Fusion: A Review of the State of the Art," *Information Fusion* (9:4), pp. 446-449.