Big Data in Medicine Is Driving Big Changes

F. Martin-Sanchez[1,2], K. Verspoor[2,1]

[1] Health and Biomedical Informatics Centre, The University of Melbourne, Parkville VIC 3010 Australia
[2] Department of Computing and Information Systems, The University of Melbourne, Parkville VIC 3010 Australia

Summary

Objectives: To summarise current research that takes advantage of "Big Data" in health and biomedical informatics applications.

Methods: Survey of trends in this work, and exploration of literature describing how large-scale structured and unstructured data sources are being used to support applications from clinical decision making and health policy, to drug design and pharmacovigilance, and further to systems biology and genetics.

Results: The survey highlights ongoing development of powerful new methods for turning that large-scale, and often complex, data into information that provides new insights into human health, in a range of different areas. Consideration of this body of work identifies several important paradigm shifts that are facilitated by Big Data resources and methods: in clinical and translational research, from hypothesis-driven research to data-driven research, and in medicine, from evidence-based practice to practice-based evidence.

Conclusions: The increasing scale and availability of large quantities of health data require strategies for data management, data linkage, and data integration beyond the limits of many existing information systems, and substantial effort is underway to meet those needs. As our ability to make sense of that data improves, the value of the data will continue to increase. Health systems, genetics and genomics, population and public health; all areas of biomedicine stand to benefit from Big Data and the associated technologies.

Keywords

Medical informatics, data mining, text mining, information systems, information storage and retrieval

Yearb Med Inform 2014:14-20
http://dx.doi.org/10.15265/IY-2014-0020
Published online August 15, 2014

1 Introduction

The Health Informatics Society of Australia (HISA), a member society of IMIA, organized a "Big Data" conference in Melbourne, Australia in April 2013 and 2014. The conference has addressed research, industry, government and clinical practice, introducing more than 200 clinicians, healthcare executives and managers, data and information professionals, health informaticians, health policy makers, and academics to the exploding world of Big Data in health and biomedicine.

During these two conferences, it was made very clear that Big Data is a hot topic in health care and biomedical research. The increased usage of the term "Big Data" in the biomedical literature is indicative of the emerging importance of large-scale data sets in health and biomedicine, and there is also an increasing awareness of the role that big data can play in scientific and clinical research. It was also observed that the term "Big Data" could mean different things to different groups of people. However, there was a common recognition that health care, biomedical research and population health are generating massive, complex, distributed, and often dynamic sets of data, and that the size and complexity of this data will pose both challenges and opportunities to health organizations.

The term Big Data is believed to have originated with Web search companies who had to query very large distributed aggregations of loosely structured data.[1]

This term has since been used to refer to the massive amounts of data collected over time, that are difficult to analyze and handle when using common database management tools.[2] While the term may seem to reference the volume of data, that isn't always the case. In a 2001 research report, Gartner (formerly META Group) analyst Doug Laney defined data growth challenges and opportunities as being three-dimensional, i.e. increasing volume (amount of data), velocity (speed of data in and out), and variety (range of data types and sources).[3] Much of the industry continues to use this "3 Vs" model for describing big data. In 2012, Gartner updated its definition as follows: «*Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization*." Additionally, a new V for "Veracity" has been added by some organizations to describe it. The term Big Data may also refer to the technology (such as storage facilities, tools and processes) that an organization requires to handle large amounts of data.[4] A more pragmatic definition defines Big Data in terms of a requirement for analytic applications to handle new types of data that (an organization) wasn't previously tracking.[5]

[1] http://www.webopedia.com/TERM/B/big_data.html

[2] PC Magazine Encyclopedia. http://www.pcmag.com/encyclopedia/term/62849/big-data

[3] Laney, Douglas. "3D Data Management: Controlling Data Volume, Velocity and Variety". Gartner.

[4] "What is Big Data?". Villanova University.

[5] David McJannet, from Hortonworks, in Information Week, 8/26/2013

In this article we survey the significant developments in advanced data management, analytics and visualization over the past few years, stimulated by the demands and opportunities of these growing data sources. Our aim is to provide readers with insights into critical advances that can contribute to improve the efficiency and performance of health organizations and biomedical researchers.

The article opens in Section 2 with an overview of the two primary types of sources of Big Data: structured data, e.g. databases, and unstructured data, e.g. natural language text. Section 3 describes published work making use of structured sources of Big Data, considering clinical data sets, research databases, population health registries, and Section 4 summarizes large-scale studies using unstructured sources of Big Data, focusing on text and considering the uses of the biomedical literature, the narrative sections from electronic health records, and social media.

2 Structured and Unstructured Big Data in Health and Biomedicine

Since the publication of the first Human Genome sequence in 2003, the field of genomics has represented a primary driving force behind the generation of Big Data in Biomedicine [1, 2]. Progress in laboratory analytical techniques (e.g., DNA sequence analysis) and mobile technologies (e.g., data from physical activity monitors and apps) are currently largely responsible for an ever-increasing real time production of data in high volumes.

However, the use of Big Data has now reached all areas of healthcare, biomedical research and population health. Health services researchers can combine administrative and clinical databases to develop predictive models to improve health policy [3]. The pharmaceutical industry manages huge repositories of clinical and molecular data for rational drug design and pharmacogenomics approaches [4]. In population health, disease registries and data from clinical records are being used for measuring the impact of health interventions [5]. Biomedical researchers have access to new sources of genomic data (e.g., microbiome, epigenomics) and can explore new hypotheses to understand the molecular causes of diseases [6]. Finally, environmental data are starting to be integrated with genetic and clinical data [7]. Such uses will be explored in more details in Section 3.

Since the early days of computing, the term 'data' has mainly been used to refer to structured data. However, in the last few years there has been a dramatic increase in the production of unstructured data, in a higher extend than the volume of available structured data. Experts from International Data Corporation have estimated that unstructured data accounts for more than 80 percent of currently available data. Section 4 explores the application of Big Data methods and tools to unstructured health and biomedical data.

Although there is no formal definition, generally the term 'structured data' refers to data with a defined schema or data model (i.e., explicit semantics). Data stored in a database is typically structured. Measurements and signals are examples of structured data. In contrast, unstructured data refers to data containing information that is not easily accessible to computational data management systems – the information it contains is not presented in a form with a clear data schema that enables direct computational interpretation and analysis. This type of data typically requires specialized analytical methods to extract the information it contains, and to transform it into computable form. Natural language text, images, and audio streams are examples of unstructured data.

In practice, structured databases may contain unstructured information (e.g., free text fields) and unstructured data may in fact have some structure (e.g., metadata associated with an image or markup embedded in a document). In addition, some data may not be clearly structured or unstructured. Nevertheless, the distinction reflects the differences in the immediate computability and interpretability of the data, and the distinction has implications for the methods that store and analyze the data.

There is a wealth of information relevant to understanding human health captured in unstructured resources. Texts, images, and audio and video streams all are commonly produced in the clinical context. These resources require the development of strategies for the extraction and the summarization of the information they contain, to impose structure and meaning by taking advantage of internal structure or patterns inherent to the source. Technological solutions to enable the automatic interpretation of such resources, in addition to assisting the people who are tasked with interpreting these data sources, allow to scale up the analysis and consider a much broader set of data – across a hospital, a population, or the scientific research community. This has resulted in some highly innovative research demonstrating the power of large-scale data analysis in medicine.

3 Structured Data Sources in Health and Biomedicine

There are three primary areas of work regarding the use of structured Big Data in biomedicine that have been explored in large-scale studies, and that we will review here: (1) molecular databases, bioinformatics, systems biology and personalized medicine, (2) clinical and healthcare applications, and (3) population and global health and policy applications.

Molecular Databases, Bioinformatics, Systems Biology and Personalized Medicine

Science and research are changing. The areas of bioinformatics, systems biology and personalized medicine are the areas in which we can perceive most clearly that a shift is underway from *hypothesis–driven* to *data-driven* research. Work in computational drug discovery and medicinal chemistry has clearly followed this trend [4]. The unprecedented generation of molecular data has also brought about new challenges in data visualization. Various software frameworks have been developed in order to facilitate data analysis and hasten time to discovery. This was the case for DIVE, a data visualization engine applied to the study of proteins [8]. Proteomics was also the focus of the Human

Proteome Organisation Proteomics Standards Initiative workshop, held in the UK in April 2013 [9]. Under the auspices of this organization, standards for data representation in proteomics were updated and refined to pave the way for Big Data approaches. The use of standards is also critical to Big Data sharing in genomics. The article by Tenenbaum et al [10] reinforces this idea and sustains that community-developed open data standards should be used. Sharing of genomic data among laboratories is pivotal to advances in the area of infectious disease surveillance. In [11] Iwasaki et al reported the application of novel bioinformatics strategies for the analysis of large influenza virus genome sequence files and the identification of potentially hazardous strains.

Advances in systems biology have also been facilitated by big data technologies. Understanding the etiology of mental diseases was the focus of the work described in [12], where a systems biology approach was taken for the study of psychiatric diseases. The development of complex models of networks of genes and proteins was reviewed by Fan and Liu [13] in the context of pharmacogenomics. Large datasets were analyzed with statistical methods to estimate complex correlations between biomarkers, genetic profiles and differential responses to drugs.

From a more technical perspective, Mohammed et al [14] reported the use of advanced computing paradigms (cloud, parallel processing) to improve the performance of molecular big data analysis processes. We know that traditional database technology is not able to process the large volumes of Big Data in an efficient way and emerging technologies such as Hadoop, MapReduce or advanced analytical and visualization tools are required.

Clinical and Healthcare Applications

The application of Big Data methods and techniques is growing quickly in the domain of clinical medicine and healthcare administration. Electronic health record (EHR) data has been argued to be the "ultimate" repository for making discoveries through clinical data mining, supporting improvements in clinical practice and even increased the understanding of the genetic basis of diseases [15]. It is also arguably the "ultimate" clinical Big Data source. Every hospital admission, every prescribed drug, every symptom and diagnosis, every procedure, may eventually be catalogued in electronic form for every patient that visits a doctor. As a striking example of the potential of this resource, over 110 million EHRs across two continents have been analyzed for genetic disease research [16]. Patterns of deleterious genetic variants were determined exclusively from the statistical analysis of phenotype comorbidities in EHR data, without any additional genome sequencing of those patients.

The availability of electronic medical records and administrative datasets is enabling a wave of innovation with projects conducting health services research. In these studies, patients can be identified as being at increased risk for readmission [3], or their estimated length of stay at intensive care units can be modeled [17]. Many of these exercises of Big Data analytics require advanced computational frameworks for high data volume and intensive data processing. Dong *et al* [18] describe a Hadoop/MapReduce architecture for large scale clinical informatics applications highlighting its intrinsic advantages, including scalability, fault tolerance and high availability.

As far as clinical applications are concerned, Big Data methods, in conjunction with mobile technologies, can enable medical specialists to read patients' signals and images remotely, as well as store, deliver, retrieve and manage diverse medical files for teleconsultation and telediagnosis [19]. Cloud computing service architecture has been used to support analysis of Big Data in Epilepsy with good results [20].

Clinical and administrative data quality improvement represents a major concern in terms of the implementation of large health data repositories. To define more systematically electronic data quality, Dixon *et al* [21] propose a novel framework for data stewardship. This framework applies a systems approach to data quality with a particular emphasis on health outcomes.

The area of patient safety represents a great opportunity for Big Data, and sits at the intersection of clinical and administrative data. For instance, Chai *et al* [22] used statistical text classification to identify health information technology incidents within large databases.

Population and Global Health and Policy Applications

Occupational and environmental medicine has also been impacted by the developments in the Big Data space. The article by Sepulveda [23] reinforces the need to engage systems in communities for healthier workforces. Data linkage is a basic mechanism to integrate disparate data sources and enable biomedical research. The need to balance privacy protection concerns with the benefits derived from creating large data repositories was addressed in [24]. The authors report the development of a computerized third-party linkage platform for privacy-preserving and interactive record linkage.

Hay and colleagues [25] discuss the potential and challenges of producing continually updated infectious disease risk maps using diverse and large volume data sources such as social media. Lastly, in their 2012 article, Jalali *et al* present how leveraging cloud computing can contribute to address public health disparities. Through Big Data analytics, predictive modeling and cloud computing, they suggest an environment where emerging public health threats can be observed in real-time and reported to policy makers [5].

4 Unstructured Textual Big Data Sources in Health and Biomedicine

While images, and audio or video streams will play an important role in Big Data analytics for biomedicine, probably the most utilized large-scale source of unstructured information in the medical context to date is natural language text, i.e., documents that are written in human language (as opposed to computer language) and intended for communication of information among humans. There are a plethora of textual data sources in medicine, perhaps most obviously in EHRs

that contain substantial clinical narrative in free-text form, but also radiology and histopathology reports, nursing triage notes, clinical letters, discharge summaries, and so forth.

There are three primary sources of texts in biomedicine that have been explored in large-scale studies, and that we will review here: (1) the published biomedical literature (e.g., journal articles), (2) electronic health records, and (3) social media and other Web-based sources. Computers can access the information in these texts, e.g. through application of natural language processing techniques that make use of linguistic regularities in the text (for example, the domain-tailored clinical Text And Knowledge Extraction System, cTAKES [26]), or through strategies for the recognition of controlled vocabulary terms (e.g., the popular MetaMap program for the recognition of Unified Medical Language System terms [27]). An overview of methods and systems can be found in a recent book chapter [28]. Extraction of key concepts and relations from a text effectively provides a representation of the content of that text and facilitates searching, comparison, and aggregation of information from many texts. Methods to retrieve and classify texts from EHRs can be used to improve the identification of patients meeting criteria for a clinical trial [29], prioritization of the information relevant to a given topic in the literature [30, 31], and clinical question answering [32]. These complementary text-processing strategies can be combined in innovative ways to facilitate access to and analysis of text-based information.

Published Biomedical Literature

The published biomedical literature, as indexed in PubMed, can be considered to be the primary repository of biomedical knowledge. While there are increasing numbers of structured resources available in biomedicine, the source data for many databases and web portals is often the published literature. It typically requires manual effort to identify relevant literature, and extract and structure the information. Such manual effort is unsustainable given the exponential growth of the literature [33, 34]. To date, over 23 million articles have been indexed in PubMed, with nearly one million publications added in 2013 alone (924,687 as of December 24, 2013). This growth has driven substantial effort towards the development of tools for information retrieval and information extraction from the relevant literature[6]. Text mining applications that exploit this substantial data source have become an integral component of biomedical data analysis and contribute to knowledge discovery and hypothesis generation [35, 36].

One specific area of focus has been the use of text mining for the interpretation of molecular biology and genomic data, including genes, proteins, cells, tissues and whole organisms, with a range of applications to gene expression data [37], genome-wide association studies [38, 39], pharmacogenomics [40], interpretation of genetic variants such as singular nucleotide polymorphisms (SNPs) [41, 42], and systems biology network modeling [43]. These bioinformatics applications will have increasing relevance to the clinical context through personalized medicine, as the molecular-level understanding of diseases impacts diagnosis and drug treatment protocols [44].

Systematic reviews of the literature play a particular role in supporting evidence-based medicine [45], and text mining has been deployed in several ways to support systematic reviews, including retrieval of documents, identification of key sentences or phrases, and evidence synthesis [46, 47]. In a recent study, large-scale text mining was used for screening and selection of articles for inclusion in a systematic review, enabling a reduction of manual screening workload of up to a remarkable 90% [30]. While there remains work to be done to support sophisticated statistical meta-analyses across studies with text mining, the foundational technologies show substantial promise.

Clinical Narratives in Electronic Health Records

Increasing adoption of electronic health records is enabling a paradigm shift in the practice of evidence-based medicine, where the concept of *practice-based evidence* has emerged [15, 48]. In this paradigm, the record of practice is consulted as a source of evidence for monitoring clinical outcomes such as drug safety or the impact of specific treatments for particular types of patients, characterized by shared characteristics such as disease co-morbidities, age, or treatment history. Medicine is still practiced in an evidence-based manner, but the source of evidence can be broadened from the published experimental literature to the "natural experiment" (i.e., an empirical, typically observational, study in which experimental conditions are out of the control of researchers) of real patients in the broad population.

Text mining plays an important role in supporting this paradigm, as so much of the data available to provide evidence is available in the narrative of EHRs. Jensen et al provide a detailed review of both the great potential of the EHR for supporting clinical care, while highlighting some of the ethical, legal and technical challenges of working with EHR data [49]. Both small-scale and large-scale studies of EHRs have emerged that indicate the strong role that analysis of the content of EHRs can play in supporting extraction of practice-based evidence. Indeed, early work on breast cancer utilized manual analysis of EHRs (in the form of the Utah Population Database) to support genetic studies of the disease, leading to the discovery of the BRCA1 and BRCA2 genes [50]. Text search of the Stanford Translational Research Integrated Database Environment (STRIDE) [51] has been used to enable electronic chart review to identify the appropriate treatment course for a patient with a complicated presentation [52].

Textual parts of EHRs have been used for pharmacovigilance [53, 54], drug safety profiling [55], and recruitment and stratification of patients in clinical trials [56, 57]. The most recent studies [53, 55] have considered records from millions of patients, and converted the unstructured text content into structured records, utilizing simple vocabulary recognition strategies. Using data from real patients undergoing active treatment, coupled with text processing and sophisticated statistical analyses, the analysis of EHRs enable drug safety surveillance at an unprecedented scale.

[6] A list of biomedical literature search and retrieval tools is maintained at http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/search/

A recent special focus issue of the Journal of the American Medical Informatics Issue (JAMIA) was dedicated to EHR-driven phenotyping (JAMIA, December 2013, Volume 20, Issue e2). The editors highlight the recent transition from the use of the EHR for clinical decision support to the use of the EHR directly for research [58], and a number of the articles make use of natural language processing to process unstructured content in the EHR for phenotyping [59-63].

The use of EHR data for research and analysis purposes is growing at such a rate that we acknowledge we have had to leave many quality studies unmentioned. Despite these oversights, this work is relatively recent, and it is clear that we have only scratched the surface in identifying questions that might be asked of this data, and in developing strategies for taking advantage of the content of this rich clinical information resource. Furthermore, given that texts in the EHR are unlikely to be replaced with fully structured data [64], natural language processing and text processing will provide important enabling technologies for this growing body of work. As stated by Shapiro et al [65], "*Much of the meaning and inference that can be gleaned by the clinician through the use of narrative is lost when a rigidly structured template is used, and the ability to communicate complex ideas in an efficient and fluid manner diminishes*." By developing methods to extract and make use of this complex clinical narrative on a large scale, we will enable much more nuanced analysis of patient health through the EHR and ultimately form a more complete picture of complex sets of characteristics that impact on the diagnosis and treatment of diseases.

Social Media and the Web

Social media sources, including on-line discussion forums, social networking sites and twitter feeds, as well as more general World Wide Web resources such as blogs and even search query logs are also valuable information resources for certain health applications.

Public health officials, for instance, can benefit from the monitoring of these on-line sources for evidence of disease outbreaks. An early publication in this area focused on tracking trends in Google query logs to detect influenza activity in specific geographic regions with large populations of web search users [66], and many extensions of this initial study have appeared since, as recently reviewed by Polgreen and Velasco [67]. According to that review, most authors recommend that social media-based surveillance should support rather than supplant other surveillance programs due to some noise in the methods, but the methods have the advantage of enabling rapid detection of disease. While there are earlier studies than the Google-based work [66], that study was able to exploit a substantially larger set of data – hundreds of billions of searches from 5 years of Google web search logs – resulting in more comprehensive models and finer-grained estimates of flu activity in different regions. The scale of the study was the largest of its kind to that point, and illustrates the "*unreasonable effectiveness of data*" [68]. More recently, on-line news sources [69] and Twitter have been explored as a data source for epidemic or syndromic surveillance [70, 71], with the potential of outperforming query-based search analyses.

Another area where social media has been explored for health is in the monitoring of adverse drug events. On-line medical discussion forums in particular have been a focus of several studies aiming to identify adverse drug effects from patient or caregiver reports of drug reactions in their on-line communications [72, 73]. As with disease surveillance, large-scale search query logs have also been shown to provide valuable information about adverse drug events, in particular reactions caused by specific drug combinations [74].

5 Conclusions

In this review, we have explored the growing use of Big Data in health and biomedicine. We described applications of Big Data divided into two groups, according to two distinct types of data, i.e., structured and unstructured data. However, we are well aware that in the future we will see more and more examples of applications that integrate both types of data. The effective integration of diverse data types poses a substantial challenge to current methods. This issue is expressed well by Mudunuri [75], who stated, "*In order to achieve useful results, researchers require methods that consolidate, store and query combinations of structured and unstructured data sets efficiently and effectively*". An early example of the application of Big Data principles to integrated structured and unstructured data can be found in [6], where tumor images and tumor mathematical parameters are jointly analyzed with a clinical prognostic purpose. Analytical methods that can simultaneously handle numerical, discrete, and categorical data, or integrate features from several modalities and data types – and can do so in the context of unprecedented data volumes – are still largely undeveloped.

There are still many unknowns around the practical use of Big Data in medicine. It has been said that our ability to produce data outpaces our ability to analyze it. This in turn raises issues related to the workforce and the need to train a new generation of data scientists with new skillsets including mathematics and statistics, IT, informatics and computer science, in combination with biology and medicine. Some people have begun talking about the value of Small Data (e.g., those generated by individuals through self-monitoring devices and apps). There also exist important challenges in the areas of privacy and security [76]. However, we cannot afford to have so many sources of data stored in different places, without the possibility of using them in research projects to improve our delivery of health services and our understanding of disease.

This article has highlighted the notable changes that are happening in clinical practice and biomedical research, enabled by both the increasing amounts of data that are being produced, and the development of innovative technical approaches to harnessing the information available in that data. The shifts to practice-based evidence for medicine, and data-driven rather than strictly hypothesis-driven biomedical research, represent the Big Changes driven by Big Data. It is an exciting time to be working at the interface between informatics and biomedicine.

Acknowledgements

FM is funded by the Institute for a Broadband-Enabled Society (IBES) of the University of Melbourne and the Australian NHMRC. He was the Chair of the Scientific Program Committee for the first HISA Conference on BIG DATA, held in Melbourne, Australia in April 2013.

KV was the Chair of the Scientific Program Committee for the second HISA conference on BIG DATA, held in Melbourne, Australia in April 2014. She began this work while at NICTA, which is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program.

References

1. O'Driscoll A, Daugelaite J, Sleator RD. 'Big data', Hadoop and cloud computing in genomics. J Biomed Inform 2013;46(5):774-81.
2. Dai L, Gao X, Guo Y, Xiao J, Zhang Z. Bioinformatics clouds for big data manipulation. Biol Direct 2012;7:43; discussion 43.
3. de Lissovoy G. Big data meets the electronic medical record: a commentary on "identifying patients at increased risk for unplanned readmission". Med Care 2013;51(9):759-60.
4. Lusher SJ, McGuire R, van Schaik RC, Nicholson CD, de Vlieg J. Data-driven medicinal chemistry in the era of big data. Drug Discov Today 2014 Jul;19(7):859-68.
5. Jalali A, Olabode OA, Bell CM: Leveraging Cloud Computing to Address Public Health Disparities: An Analysis of the SPHPS. Online J Public Health Inform 2012;4(3).
6. Wang LW, Qu AP, Yuan JP, Chen C, Sun SR, Hu MB, et al. Computer-based image studies on tumor nests mathematical features of breast cancer and their clinical prognostic value. PLoS One 2013;8(12):e82314.
7. Duncan LE, Keller MC. A critical review of the first 10 years of candidate gene-by-environment interaction research in psychiatry. Am J Psychiatry 2011;168(10):1041-9.
8. Bromley D, Rysavy SJ, Su R, Toofanny RD, Schmidlin T, Daggett V. DIVE: A Data Intensive Visualization Engine. Bioinformatics 2014 Feb 15;30(4):593-5.
9. Orchard S, Binz PA, Jones AR, Vizcaino JA, Deutsch EW, Hermjakob H. Preparing to work with big data in proteomics - a report on the HUPO-PSI Spring Workshop: April 15-17, 2013, Liverpool, UK. Proteomics 2013;13(20):2931-7.
10. Tenenbaum JD, Sansone SA, Haendel M. A sea of standards for omics data: sink or swim? J Am Med Inform Assoc 2014 Mar-Apr;21(2):200-3
11. Iwasaki Y, Abe T, Wada Y, Wada K, Ikemura T. Novel bioinformatics strategies for prediction of directional sequence changes in influenza virus genomes and for surveillance of potentially hazardous strains. BMC Infect Dis 2013;13:386.
12. Mewes HW: Perspectives of a systems biology of the brain: the big data conundrum understanding psychiatric diseases. Pharmacopsychiatry 2013;46 Suppl 1:S2-9.
13. Fan J, Liu H. Statistical analysis of big data on pharmacogenomics. Adv Drug Deliv Rev 2013;65(7):987-1000.
14. Mohammed Y, Mostovenko E, Henneman AA, Marissen RJ, Deelder AM, Palmblad M. Cloud parallel processing of tandem mass spectrometry based proteomics data. J Proteome Res 2012;11(10):5101-8.
15. Shah NH: Mining the ultimate phenome repository. Nat Biotech 2013;31(12):1095-7.
16. Blair DR, Lyttle CS, Mortensen JM, Bearden CF, Jensen AB, Khiabanian H, et al: A Nondegenerate Code of Deleterious Variants in Mendelian Loci Contributes to Complex Disease Risk. Cell 2013;155(1):70-80.
17. Choi M, Lee J, Ahn MJ, Kim Y. Nursing critical patient severity classification system predicts outcomes in patients admitted to surgical intensive care units: use of data from clinical data repository. Stud Health Technol Inform 2013;192:1063.
18. Dong X, Bahroos N, Sadhu E, Jackson T, Chukhman M, Johnson R, et al. Leverage hadoop framework for large scale clinical informatics applications. AMIA Summits Transl Sci Proc 2013;2013:53.
19. Hsieh JC, Li AH, Yang CC. Mobile, cloud, and big data computing: contributions, challenges, and new directions in telecardiology. Int J Environ Res Public Health 2013;10(11):6131-53.
20. Shen CP, Zhou W, Lin FS, Sung HY, Lam YY, Chen W, et al. Epilepsy analytic system with cloud computing. Conf Proc IEEE Eng Med Biol Soc 2013:2013:1644-7.
21. Dixon BE, Rosenman M, Xia Y, Grannis SJ. A vision for the systematic monitoring and improvement of the quality of electronic health data. Stud Health Technol Inform 2013;192:884-8.
22. Chai KE, Anthony S, Coiera E, Magrabi F. Using statistical text classification to identify health information technology incidents. J Am Med Inform Assoc 2013;20(5):980-5.
23. Sepulveda MJ. From worker health to citizen health: moving upstream. J Occup Environ Med 2013;55(12 Suppl):S52-57.
24. Kum HC, Krishnamurthy A, Machanavajjhala A, Reiter MK, Ahalt S. Privacy preserving interactive record linkage (PPIRL). J Am Med Inform Assoc 2014 Mar-Apr;21(2):212-20.
25. Hay SI, George DB, Moyes CL, Brownstein JS. Big data opportunities for global infectious disease surveillance. PLoS Med 2013;10(4):e1001413.
26. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507-13.
27. Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 2010;17(3):229-236.
28. Doan S, Conway M, Phuong TM, Ohno-Machado L. Natural Language Processing in Biomedicine: A Unified System Architecture Overview. In: Trent RJA, editor. Clinical Bioinformatics. Springer; 2014.
29. Voorhees EM, Hersh W. Overview of the TREC 2012 Medical Records Track. In: The 21st Text REtrieval Conference (TREC 2012); 2012.
30. Shemilt I, Simon A, Hollands GJ, Marteau TM, Ogilvie D, O'Mara-Eves A, et al. Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. Research Synthesis Methods 2013:online preprint.
31. Arighi CN, Carterette B, Cohen KB, Krallinger M, Wilbur WJ, Fey P, et al. An overview of the BioCreative 2012 Workshop Track III: interactive text mining task. Database 2013;2013.
32. Cairns BL, Nielsen RD, Masanz JJ, Martin JH, Palmer MS, Ward WH, et al. The MiPACQ Clinical Question Answering System. In: AMIA Annu Symp Proceedings; 2011.
33. Lu Z. PubMed and beyond: a survey of web tools for searching biomedical literature. Database 2011(baq036).
34. Baumgartner Jr. WA, Cohen KB, Fox L, Acquaah-Mensah GK, Hunter L. Manual curation is not sufficient for annotation of genomic databases. Bioinformatics 2007;23:i41-i48.
35. Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. Nat Rev Genet 2006;7(2):119-29.
36. Rebholz-Schuhmann D, Oellrich A, Hoehndorf R. Text-mining solutions for biomedical research: enabling integrative biology. Nat Rev Genet 2012;13(12):829-39.
37. Tanabe L, Scherf U, Smith LH, Lee JK, Hunter L, Weinstein JN. MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. Biotechniques 1999;27(6):1216-7.
38. Johansson M, Roberts A, Chen D, Li Y, Delahaye-Sourdeix M, Aswani N, et al. Using Prior Information from the Medical Literature in GWAS of Oral Cancer Identifies Novel Susceptibility Variant on Chromosome 4 - the AdAPT Method. PLoS One 2012;7(5):e36888.
39. Raychaudhuri S, Plenge RM, Rossin EJ, Ng ACY, Purcell SM, Sklar P, et al. Identifying Relationships among Genomic Disease Regions: Predicting Genes at Pathogenic SNP Associations and Rare Deletions. PLoS Genet 2009;5(6):e1000534.
40. Hahn U, Cohen KB, Garten Y, Shah NH. Mining the pharmacogenomics literature: A survey of the state of the art. Briefings in Bioinformatics 2012;13(4):460-94.
41. Hakenberg J, Voronov D, Nguyen VH, Liang S, Anwar S, Lumpkin B, et al. A SNPshot of PubMed to associate genetic variants with drugs, diseases, and adverse reactions. J Biomed Inform 2012; 45(5):842-50.
42. Verspoor K, Jimeno Yepes A, Cavedon L, McIntosh T, Herten-Crabb A, Thomas Z, et al. Annotating the biomedical literature for the human variome. Database 2013;2013.
43. Li C, Liakata M, Rebholz-Schuhmann D. Biological network extraction from scientific literature: state of the art and challenges. Brief Bioinform 2013.
44. Roden DM, Tyndale RF. Genomic Medicine, Pre-

cision Medicine, Personalized Medicine: What's in a Name? Clin Pharmacol Ther 2013;94(2):169-72.
45. Sackett DL, Rosenberg WMC, Muir Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ 1996;312(7023):71-2.
46. Tsafnat G, Dunn A, Glasziou P, Coiera E The automation of systematic reviews. BMJ 2013;346:f139.
47. Cohen AM, Adams CE, Davis JM, Yu C, Yu PS, Meng W, et al. Evidence-based medicine, the essential role of systematic reviews, and the need for automated text mining tools. In: Proceedings of the 1st ACM International Health Informatics Symposium; Arlington, Virginia, USA. 1883046: ACM 2010. p. 376-80.
48. Pincus T, Sokka T. Evidence-based practice and practice-based evidence. Nat Clin Pract Rheum 2006;2(3):114-5.
49. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012;13(6):395-405.
50. Hurdle JF, Smith KR, Mineau GP. Mining electronic health records: an additional perspective. Nat Rev Genet 2013;14(1):75.
51. Lowe HJ, Ferris TA, Hernandez PM, Weber SC. STRIDE – An Integrated Standards-Based Translational Research Informatics Platform. In: AMIA Annual Symposium Proceedings 2009. p. 391–5.
52. Frankovich J, Longhurst CA, Sutherland SM. Evidence-Based Medicine in the EMR Era. N Engl J Med 2011, 365(19):1758-9.
53. LePendu P, Iyer SV, Bauer-Mehren A, Harpaz R, Mortensen JM, Podchiyska T, et al. Pharmacovigilance Using Clinical Notes. Clin Pharmacol Ther 201; 93(6):547-55.
54. Bates DW, Evans RS, Murff H, Stetson PD, Pizziferri L, Hripcsak G. Detecting Adverse Events Using Information Technology. J Am Med Inform Assoc 2003;10(2):115-28.
55. Leeper NJ, Bauer-Mehren A, Iyer SV, LePendu P, Olson C, Shah NH. Practice-Based Evidence: Profiling the Safety of Cilostazol by Text-Mining of Clinical Notes. PLoS One 2013;8(5):e63499.
56. Roque FS, Jensen PB, Schmock H, Dalgaard M, Andreatta M, Hansen T, et al. Using Electronic Patient Records to Discover Disease Correlations and Stratify Patient Cohorts. PLoS Comput Biol 2011;7(8):e1002141.
57. Korkontzelos I, Mu T, Restificar A, Ananiadou S. Text mining for efficient search and assisted creation of clinical trials. In: Proceedings of the ACM fifth international workshop on Data and text mining in biomedical informatics; Glasgow, Scotland, UK. 2064706: ACM 2011. p. 43-50.
58. Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. J Am Med Inform Assoc 2013;20(e2):e206-e211.
59. Lyalina S, Percha B, LePendu P, Iyer SV, Altman RB, Shah NH. Identifying phenotypic signatures of neuropsychiatric disorders from electronic medical records. J Am Med Inform Assoc 2013;20(e2):e297-e305.
60. Gundlapalli AV, Redd A, Carter M, Divita G, Shen S, Palmer M, et al. Validating a strategy for psychosocial phenotyping using a large corpus of clinical text. J Am Med Inform Assoc 2013;20(e2):e355-e364.
61. Davis MF, Sriram S, Bush WS, Denny JC, Haines JL. Automated extraction of clinical traits of multiple sclerosis in electronic medical records. J Am Med Inform Assoc 2013;20(e2):e334-e340.
62. Pathak J, Bailey KR, Beebe CE, Bethard S, Carrell DS, Chen PJ, et al. Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. J Am Med Inform Assoc 2013;20(e2):e341-e348.
63. Deleger L, Brodzinski H, Zhai H, Li Q, Lingren T, Kirkendall ES, et al. Developing and evaluating an automated appendicitis risk stratification algorithm for pediatric patients in the emergency department. J Am Med Inform Assoc 2013;20(e2):e212-e220.
64. Resnik P, Niv M, Nossal M, Kapit A, Toren R. Communication of Clinically Relevant Information in Electronic Health Records: A Comparison between Structured Data and Unrestricted Physician Language. In: Perspectives in Health Information Management. CAC Proceedings 2008.
65. Shapiro J, Bakken S, Hyun S, Melton G, Schlegel C, SB J: Document ontology: supporting narrative documents in electronic health records. In: AMIA Annual Symposium Proceedings: 2005. p. 684-6.
66. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2009;457(7232):1012-4.
67. Bernardo TM, Rajic A, Young I, Robiadek K, Pham MT, Funk JA. Scoping Review on Search Queries and Social Media for Disease Surveillance: A Chronology of Innovation. J Med Internet Res 2013;15(7):e147.
68. Halevy A, Norvig P, Pereira F. The Unreasonable Effectiveness of Data. IEEE Intelligent Systems 2009;24(2):8-12.
69. Collier N. Uncovering text mining: a survey of current work on web-based epidemic intelligence. Glob Public Health 2012;7(7):731-49.
70. Paul MJ, Dredze M. You are what you Tweet: Analyzing Twitter for Public Health. In: Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM 2011); Barcelona, Spain; 2011.
71. Gesualdo F, Stilo G, Agricola E, Gonfiantini MV, Pandolfi E, Velardi P, et al. Influenza-Like Illness Surveillance on Twitter through Automated Learning of Naïve Language. PLoS One 2013;8(12):e82489.
72. Wu H, Fang H, Stanhope S. Exploiting online discussions to discover unrecognized drug side effects. Methods Inf Med 2013;52(2):152-9.
73. Leaman R, Wojtulewicz L, Sullivan R, Skariah A, Yang J, Gonzalez G. Towards Internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. In: Proceedings of the Workshop on Biomedical Natural Language Processing; Uppsala, Sweden. Association for Computational Linguistics; 2010. p. 117-25.
74. White RW, Tatonetti NP, Shah NH, Altman RB, Horvitz E. Web-scale pharmacovigilance: listening to signals from the crowd. J Am Med Inform Assoc 2013;20(3):404-408.
75. Mudunuri US, Khouja M, Repetski S, Venkataraman G, Che A, Luke BT, et al. Knowledge and Theme Discovery across Very Large Biological Data Sets Using Distributed Queries: A Prototype Combining Unstructured and Structured Data. PLoS One 2013;8(12):e80503.
76. Kum HC, Ahalt S. Privacy-by-Design: Understanding Data Access Models for Secondary Data. AMIA Summits Transl Sci Proc 2013;2013:126-30.

Correspondence to:
Fernando Martin-Sanchez
Health and Biomedical Informatics Centre
The University of Melbourne
Parkville VIC 3010
Australia
E-mail: fjms@unimelb.edu.au