Technical Challenges for Big Data in Biomedicine and Health: Data Sources, Infrastructure, and Analytics

N. Peek[1 2], J. H. Holmes[3], J. Sun[4]

[1] Dept. of Medical Informatics, Academic Medical Center, University of Amsterdam, The Netherlands
[2] Centre for Health Informatics, Institute of Population Health , University of Manchester, Manchester, UK
[3] Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA
[4] College of Computing, Georgia Institute of Technology, Atlanta, GA, USA

Summary

Objectives: To review technical and methodological challenges for big data research in biomedicine and health.

Methods: We discuss sources of big datasets, survey infrastructures for big data storage and big data processing, and describe the main challenges that arise when analyzing big data.

Results: The life and biomedical sciences are massively contributing to the big data revolution through secondary use of data that were collected during routine care and through new data sources such as social media. Efficient processing of big datasets is typically achieved by distributing computation over a cluster of computers. Data analysts should be aware of pitfalls related to big data such as bias in routine care data and the risk of false-positive findings in high-dimensional datasets.

Conclusions: The major challenge for the near future is to transform analytical methods that are used in the biomedical and health domain, to fit the distributed storage and processing model that is required to handle big data, while ensuring confidentiality of the data being analyzed.

Keywords

Big Data, electronic health records, distributed computing, statistical analysis

Yearb Med Inform 2014:42-7
http://dx.doi.org/10.15265/IY-2014-0018
Published online August 15, 2014

1 Introduction

Big Data are driving a revolution in information and communication technology. Technological advances in nano-electronics, interconnectivity, and information-sensing devices are generating unprecedented amounts of data. These data collections become so large and complex that they can no longer be processed using traditional data management and analysis tools.

Big Data are encountered in diverse areas such as meteorology, finance, experimental physics, telecommunication, military surveillance, business informatics, environmental research, and social media. Also the life and biomedical sciences are not shielded from – and in fact, already massively contributing to – the big data revolution, due to advances in genome sequencing technology and digital imaging, growth of clinical data warehouses, increased role of the patient in managing his own health information and rapid accumulation of biomedical knowledge.

In this paper we review the main technical and methodological challenges for big data in biomedicine and health. We first focus on the various sources of big datasets, which range from traditional data collections, such as medical records and administrative data, to novel sources of information such as web search logs and social media. Subsequently, we review infrastructures for data storage and data processing, and discuss how these can be utilized to build analytic pipelines. Finally, we review analytical challenges associated with big data in biomedicine and health, such as selection of cases and controls, bias and confounding in observational data, and techniques for mining high-dimensional data. We conclude with recommendations for future research. Two major subfields of big data in biomedicine, natural language processing (NLP) and biomedical image processing, are not covered in this review. We refer the reader to the excellent overviews by Meystre et al. [1], Friedman and Elhadad [2], Deserno [3], and Rubin et al. [4] for an in-depth discussion of these subfields.

2 Sources of Data

There are a large number of sources of big data, and these are growing in number and in the amount of data they contain. For the purposes of this paper, we focus on secondary data sources. Secondary data is defined as such in terms of its use, and thus one often hears the term "secondary use of data"; these two terms are essentially synonymous. The remainder of this section of the paper is dedicated to discussing specific sources of secondary data.

2.1 Medical Records

Medical records provide an excellent example of data that were collected for a specific purpose (patient care), but used for other purposes such as research or quality assurance. Medical record data have been a veritable treasure trove of clinical data, especially for retrospective research. The widespread adoption of electronic health records (EHRs) offers

the hope of not only improving routine care, but also by providing easier access to clinical data for research purposes.

Medical records contain an extensive array of data ranging from demographics, laboratory values, dispensed medications, imaging and other diagnostic data, clinical interventions, to clinical notes in free-form text. They are highly longitudinal, with repeated observations on patients and they reflect the size of the healthcare organizations whence they come.

An especially appealing vision is to integrate EHR data with genomic, proteomic, and metabolomic information, enabling the discovery and testing of new genotype-phenotype associations. Because EHRs contain large populations with diverse diseases, they have the potential to act as platforms for generating sets of cases and controls for translational research. Ultimately, such integrated datasets can facilitate personalized medicine through advanced decision support technology.

Medical records present a number of challenges to researchers. These include data quality, data heterogeneity, and preservation of confidentiality and privacy of providers as well as patients. As with any secondary data source, it is important to remember that these data were not collected for the purposes of research; rather, they were collected for patient care. Thus it is often the case that medical record data do not provide the information needed for secondary use; in addition, data may be missing or inaccurate. Furthermore, exposure to clinical therapies will be confounded by disease severity, hampering inference on the causal effect of these therapies on health outcomes. Finally, medical record data must be used in a way that preserves the confidentiality and privacy of patients while still allowing users to identify patients over time. This is impossible to accomplish if the very difficult issue of entity resolution is not addressed successfully. Entities, such as individual patients or clinical encounters- in short, anything in the clinical enterprise about which data are recorded- can be difficult to track throughout a medical record database. For example, a given patient might have several (or more) medical record numbers. Or a given medical record number could represent two or more patients. While conducting a retrospective cohort study, where tracking the progress of patients over time is quintessential to the study design, such inconsistencies can pose an insurmountable obstacle. Resolution of entities, through the use of such approaches as a master patient index (MPI), is required to address this problem, yet in many systems the MPI has not been implemented, often because of the cost and effort required to do so, as well as possible concerns about preserving patient confidentiality. Another approach is to establish a system of *ad hoc* proxy identifiers that are implemented for a specific purpose. This is in contrast to the MPI, which is globally applied across all medical record systems in a given institution. However, it is a very time consuming and potentially inaccurate approach, given that such identifiers (typically called Study ID or ID Number) are not usually applied by those with specific training in large clinical databases but by clinical researchers. Ideally, the MPI is the best way to establish entity resolution, but the process of implementing and maintaining an MPI system needs to be monitored constantly to ensure compliance with policies and procedures governing the MPI.

2.2 Administrative Data

The amount of clinical data across the healthcare landscape is equaled only by administrative data. Such data typically focus on insurance or other claims for payment, and may include diagnoses, procedures, and medications and devices. Thus, we use the term "administrative data" to describe those that are not primarily clinical, but which may contain data that are related to the clinical enterprise. These data are generated as a result of patient encounters and consist primarily of billings which are based on diagnosis and procedure codes, typically ICD-9-CM or ICD-10. While EHR data are often available immediately or within 24h, administrative data involve a delay because they build on diagnostic codes that are assigned after discharge. Claims data also have minimal clinical information than EHR data.

While administrative data at first might appear to be not particularly informative, they are the one resource that contains diagnosis and procedures in a form that is easily queried and analyzed. Researchers can use these data to identify and build cohorts of patients for various studies. Access to administrative data is essential not only for the research enterprise but also those working in quality improvement, patient safety, and surveillance.

The challenge of working with administrative data is threefold. First, there is a lot of it, even for a single patient. Consider US Medicare claims data, which contains a record for each and every transaction with the healthcare system. A second challenge is that creating a simple analytic file from such data can be very difficult when there are many multivalued attributes and repeating groups, and even composite attributes that must be decomposed to be useful in an analysis. Finally, because of the lack in clinical detail, administrative data are seldom useful by themselves for answering most research questions. A common solution is to link them to clinical and other healthcare data so that meaningful analyses can be performed. However, record linkage can be difficult if linking fields such as identifiers are not readily available in the data, or if there are errors in the identifiers, or if they don't match identifiers in other resources that need to be linked to the administrative data. In addition, linking administrative and clinical or other data may be forbidden by data use agreements or general use policies.

2.3 Web Search Logs

An increasing number of people use the Internet to seek health information before they visit their doctor [5]. Systematically collecting and analyzing these health-related web searches has been shown to have considerable potential for syndromic surveillance, the analysis of health-related data to forecast a disease case or outbreak that warrants public health response [6]. Traditional systems for syndromic surveillance rely on data from clinical encounters with health professionals or pharmacy data [7]. For conditions where consumers consult the Internet before they visit a physician, systematic mining of web searches may be a valuable addition to traditional approaches. Recently, web search logs were also shown to be useful for the detections adverse events associated with pharmaceutical products [8].

2.4 Social Media

Researchers and others are increasingly turning to social media as a source of data.

Discussion boards, blogs, Twitter, and other social networking resources provide extremely rich sources of data that can be mined for identifying previously unreported drug side effects, monitoring health beliefs and behaviors, and disease outbreaks. With as many as 500 million tweets per day, the Web provides a virtual flood of data [9]. These sources pose unique challenges. For example, the language that is used on social media is not standard, being replete with abbreviations, graphics, emoticons, and typographical errors. This poses a substantial challenge for extracting meaningful information from these data. Second, user agreements often prohibit the secondary use of these data, although some sources such as Twitter allow sampling of substantial numbers of records. Third, de-identification of the data poses a very substantial challenge. Stripping the names of people or places from Web communications can be done using controlled vocabularies, but this becomes much more difficult when nicknames or abbreviations are used. Finally, even if text in a Web communication is easily parsed and analyzed, there is no guarantee that the communication is truthful or accurately expresses what the poster wanted to say.

3 Infrastructure

The primary computational challenge related to big datasets concern their size (volume). However, additional important challenges relate to the speed with which they are gathered and should be processed (velocity), and the diversity of the data itself (variety). Efficient processing of big datasets is typically achieved by distributing computational tasks over a cluster of computers. In order to achieve this, the data itself must be stored in a distributed fashion. The Hadoop Distributed File System (HDFS) and the associated MapReduce algorithm are well-known examples. While Hadoop is typically aimed at processing the data in batch, more recently real-time distributed processing systems have emerged. Below, we introduce these technologies in more detail and describe existing and potential applications in the biomedical domain.

For clinical researchers, all these infrastructures provide new opportunities to conduct large-scale research in a speedy fashion. Many research have started using such infrastructure in various biomedical applications such as bioinformatics and genomic analysis [10], image informatics [11], and clinical informatics [12,13]. In particular, researchers can consider moving large research datasets into NOSQL paradigm instead managing traditional file system on a single machine. There are software tools such as Sqoop (http://sqoop.apache.org/) for converting traditional database into NOSQL infrastructure like HDFS.

3.1 Storage

3.1.1 File Systems

HDFS is a distributed file system designed to run on commodity hardware, where a data file is replicated on different servers for reliability purposes. Another benefit is locality because Hadoop prefers data and computation to be co-located on the same machine. Typically, HDFS stores data in raw format such as text or image files, but accessing files within HDFS requires special tools. HDFS has been used for storing biomedical data such as electrophysiological data [13].

Amazon's Simple Storage Service (S3) is used to store data for services in Amazon Elastic Compute Cloud (EC2), which include

- Content Storage and Distribution: i.e. store media files for a website;
- Storage for Data Analysis: i.e. Elastic MapReduce load data from S3;
- Backup, Archiving: store a copy of files in the cloud;
- Static Website Hosting: similar to content distribution, upload the html, js files to the cloud.

The purpose of S3 is similar to HDFS but aimed at processing tasks that are run on the Amazon cloud. Researchers have used Amazon S3 for storing genome sequence data [10,14].

3.1.2 Database Systems

Besides simple file storage, users require efficient query and retrieval capabilities. Database systems provide such capabilities. For big data, relational databases are not preferred because 1) the data volume is often too big for most relational databases; 2) the data are too heterogeneous and complex to fit into a predefined schema. Therefore NoSQL databases are preferred, which aim at handling big data without rigid definition of schema. We cover three different NoSQL databases: Dynamo (a key-value store), MongoDB (a document store), and HBase (a column store). Each of them provides flexible and efficient mechanism for querying data.

DynamoDB is a proprietary NoSQL database service developed by Amazon. In DynamoDB, a database is a collection of tables. And a table is a collection of items and each item is a collection of attributes. Unlike in traditional, relational databases, individual items in DynamoDB can have an arbitrary number of attributes. In healthcare applications, a patient can have multiple encounter records with various attributes, which can be easily captured in a DynamoDB. For example, one record can have only one diagnosis and one medication, while the other have multiple diagnoses, medications, lab results and imaging results. DynamoDB integrates well with Amazon Elastic MapReduce, where computation is automatically scaled up and down, based on the workload requirements from the analytics.

MongoDB is an open-source document database, which is tailored for storing JavaScript Object Notation (JSON) objects. Since JSON is widely used on the web, MongoDB is popular as the backend for large-scale web applications. MongoDB also has well-supported integration with Hadoop. One may use MongoDB as an input source and/or an output destination for Hadoop jobs. Like DynamoDB, MongoDB does not require a fixed schema. MongoDB is suitable for storing both unstructured data like clinical notes as well as structured data, like diagnostic and medication information of patients. For example, lociNGS uses MongoDB for handling genomic data [15].

HBase is an Open Source version of Bigtable [16] and is compatible with Hadoop and HDFS. Data in HBase (or Bigtable) are organized as rows, columns, column families and timestamps. HBase is suitable for random, real-time read/write access to big data sets. HBase also provides an easy way to manage data with multiple versions. For example, since each patient may have multiple encoun-

ters, the analytics can track the last three visits of each patient easily using HBase thanks to the timestamp associated with each row.

3.1.2 Privacy and Security Considerations

Cloud computing and storage have tremendous impact in many industries including healthcare. Many of such technologies enable faster and more efficient data sharing. Data in healthcare are sensitive where patient privacy needs to be protected. Laws have been passed including Health Insurance Portability and Accountability Act (HIPAA) and Health Information Technology for Economic and Clinical Health Act (HITECH), which both provide privacy regulations on Protected Health Information (PHI). Rosenthal and et al. [17] have systematically compared cloud computing technology and traditional internal environment for supporting biomedical applications. Besides all the benefits and flexibility that a cloud provides, they found also that there are more security risks associated with traditional environments as opposed to cloud environments. Nevertheless, security and privacy concerns still require better protection schemes in the cloud for biomedical data, especially on technologies around encryption and access control. Major cloud vendors such as Amazon have made strong commitment and progress in this direction for supporting HIPAA compliant applications [18].

3.2 Processing

The Hadoop ecosystem is the most popular and widely used for processing big datasets, but it is primarily intended for batch processing of large datasets. Alternative technologies aim at use cases other than batch processing such as real-time stream processing and in-memory computation.

3.2.1 Batch Processing with the Hadoop Ecosystem

Hadoop uses the MapReduce programming model, described in the seminal paper by Dean and Ghemawat [19]. It is designed for processing huge data sets such as web logs, and is tightly coupled to HDFS. MapReduce provides a simple but powerful abstraction for many data processing tasks such as web crawling and indexing. Beyond the original search engine related applications, MapReduce has been used in a large number of other domains. Also many systems have been developed on top of Hadoop to simplify implementation of complex tasks (e.g., Oozie, Cascade, described below) and to provide additional functionalities (e.g., Mahout, a machine learning library).

Oozie is a workflow engine specialized in dealing with Hadoop jobs. A workflow is a collection of actions (i.e. Hadoop MapReduce jobs) arranged in a directed acyclic graph that specifies the tasks dependency. Oozie workflow actions start jobs in remote systems, which notify Oozie when they are finished. At this point Oozie proceeds to the next action in the workflow. Oozie can thus function as the workflow engine in big data analysis pipelines. Similar to Oozie, Cascading is another popular high-level abstraction on Hadoop that handles dependency among tasks. For example, the PARAMO system [20] provides a scalable system for computing a large number of clinical predictive modeling pipelines using electronic health records, which can be implemented in either Oozie or Cascading.

Pig is a high-level data processing tool that also runs on top of Hadoop. Pig commands are written in a language called Pig Latin. The Pig system converts those commands into Hadoop jobs. Pig was originally developed at Yahoo but later published as open source. Pig is used primarily as an ETL tool for processing big data, and has therefore potential applications in DRNs. Because the heterogeneity of source data in DRNs can be challenging, Pig can be used to efficiently convert the source data into the DRN schema.

3.2.2 Non-Hadoop Alternatives

Storm is a distributed real-time computation system. Its topology is a DAG with so-called spout and bolts, where spouts generate tuples from input streams, and bolts process those tuples in real time. For streaming analytic applications, Storm can be extremely important. Storm has been used heavily in Twitter, for which real-time analytics are essential.

Spark is a very fast, distributed computing framework, which supports in-memory computing. Spark was initially developed for two applications where placing data in memory helps: iterative algorithms, which are common in machine learning, and interactive data mining. Spark provides interfaces to HDFS and MapReduce but also supports streaming data. On top of Spark, Shark is a data warehouse similar to Hive with friendly integration of machine learning algorithms [21].

GraphLab [22] provides a high-level graph-parallel abstraction that efficiently and intuitively expresses computational dependencies. GraphLab provides three phases of abstraction: Gather, Apply, and Scatter (GAS). With these abstractions, many machine learning algorithms can be easily implemented, especially for graph analysis, graphical models etc.

4 Analytics

When we analyze big datasets, their high volumes are not just inducing computational challenges but also create analytical pitfalls. Furthermore, the velocity of big data urges the need to avoid manual steps in the analytical pipeline. The variety that is typical found in big datasets, finally, creates the need to further integrate statistical, machine learning, NLP and semantic methods. Below, we illustrate these issues by describing three common analytical challenges that are associated with the use of big data in the biomedical and health domain.

4.1 Patient Selection from EHR and Administrative Data

A principal analytical challenge that is associated with the secondary use of clinical and administrative data for research purposes is the selection of relevant patients [23]. There rarely exists a single data item that can be used to identify all patients that satisfy a given diagnostic criterion. Instead, one must often combine queries on coded, numerical, and free-text data using NLP techniques. Furthermore, there are often risks of selection bias involved. For instance, when certain subgroups of the underlying population systematically undergo more complete medical evaluation, these subgroups tend to be over-represented in diagnostic incidence rates -- a phenomenon known as diagnostic sensitivity bias [24].

Typically, algorithms for selecting relevant patient records are constructed by combining clinical expertise and knowledge of the data sources at hand. However, Tessier-Sherman and colleagues found that a carefully constructed algorithm for identifying hypertensive patients from claims data only found 43-61% of patients with elevated blood pressure values in their medical charts [25]. Ritchie et al. describe a solution for this problem, which consists of a sophisticated, iterative approach to algorithm construction [26]. In their approach, clinical experts are consulted to develop an algorithm that selects cases via disease-specific combinations of billing codes, patient encounters, laboratory data, and NLP techniques on unstructured patient records. Subsequently, the results of applying the algorithm are reviewed by physicians who were not involved in algorithm development, and their feedback is used to improve the algorithm. This is repeated until the accuracy of the algorithm is considered satisfactory.

4.2 Bias and Confounding in Observational Data

Another challenge concerns the analysis of data that were gathered during routine care, and not under experimental conditions such as a randomized controlled trial. This means that all diagnostic and therapeutic procedures were only conducted when they were deemed necessary for the patient in question [27]. For instance, when respiratory failure is observed in hospitalized pneumonia patients, they are admitted to intensive care unit (ICU) for mechanical ventilation. But the mortality among these patients is higher than among patients without respiratory failure − who are not sent to the ICU. A naive approach to analyze data from hospitalized pneumonia patients may lead to the biased conclusion that ICU admission reduces the chances of survival.

The traditional approach to avoid this type of bias is to identify known confounders (e.g. demographic variables, diagnosis, and illness severity) and adjust for them either by conditioning or by propensity scoring [28]. However, the identification of confounders is a manual step, which is preferably avoided in big data applications. Furthermore, some of these confounders may be lacking in data (especially in claims data), and there may also exist unknown confounders.

Novel, data-driven approaches therefore use search techniques to identify variables that are both associated with the intervention and the outcome, but could not have occurred later in time than the intervention. For instance, De Vries et al. [28] investigated the effect of cardiac rehabilitation on survival using a large Dutch insurance claims database. Cardiac function is a well-known confounder of survival in this context, but was missing from the data. Instead the authors used all available information, comprising hospital diagnoses-treatment combinations, outpatient prescriptions, medical devices, the occurrence of lab tests, GP visits, ICU days and other services, to construct a large set of proxies for cardiac function and other potential confounders. Subsequently, generalized boosted regression [29] was used to estimate a propensity function, i.e. the probability of receiving cardiac rehabilitation as a function of 99 selected variables. The treatment effect was estimated by weighting all observations for patients who did not receive cardiac rehabilitation by the inverse of the propensity score.

4.3 Finding Associations in High-dimensional Data

As data sets grow in the number of variables that they contain, the task of finding associations with diseases or health outcomes becomes more difficult because there are higher risks of chance findings. For instance, with the conventional threshold for statistical significance of 0.05, a genome-wide association study that involves 500,000 single-nucleotide polymorphisms (SNPs) will yield 25,000 false-positive associations, within which are buried a few genuine causal alleles. From one perspective, this is a classical "multiple testing" problem, and the traditional solution is to adjust the threshold for statistical significance. For instance, Risch and Merikangas proposed to use a significance threshold of 5×10^{-8} (equivalent to a p-value of 0.05 after a Bonferroni correction for 1 million independent tests) in genome-wide association studies [30]. However, to detect an allele with a frequency of 15% and associated odds ratio of 1.25, we would need nearly 6,000 cases and 6,000 controls with this threshold. For 500,000 independent SNPs, the required sample size would be 6 billion [31].

An interesting alternative to Bonferroni correction that is increasingly applied in such situations, is permutation testing. Bonferroni correction is not informed by the data at hand, and therefore assumes a "worst case" scenario. Permutation testing can empirically assess the probability of having observed a particular result by chance, in the dataset that is being analyzed. Permutation tests shuffle case and control labels in the data, and therefore no meaningful association can be observed in permuted data. The lowest p-values observed in the permuted datasets, which represent the strongest apparent chance finding, represent a null distribution with which p-values from the original data can be compared. Interestingly, permutation testing can easily be implemented in a distributed processing system such as MapReduce.

Another alternative to Bonferroni correction was used in the Google Flu Trends project [7,32]. Google designed an automated method for selecting influenza-related search queries, requiring no previous knowledge about influenza. Fifty million search queries were separately tested for their correlation with data from the CDC's US Influenza Sentinel Provider Network, independently in nine different geographical regions. The risk of false-positive findings was low because the chance that a random search query can fit the influenza percentages in nine regions is considerably less than the chance that a random search query can fit a single location. The 45 highest-scoring queries were combined to create the final model that had a mean correlation of 0.90 over the nine regions.

5 Discussion

The electronic capture of biomedical and health data is quickly growing in many areas, providing new sources of information and new opportunities for answering health-related research questions. The resulting datasets are "big" in size, in diversity, and in the speed with which they are gathered. This creates unprecedented challenges for storing, processing, and analyzing these data.

In this paper we have reviewed the main technical and methodological challenges for big data researchers in biomedicine and health, focusing on the various sources of big datasets, different infrastructures for data storage and data processing, and analytical challenges. Computer scientists have developed various technological solutions for storing and processing big datasets over the last decade. To date, the Hadoop ecosystem constitutes the most powerful and mature framework for handling big data, but is restricted to batch processing. To support real-time processing, new technologies such as Storm and Spark are emerging. A major challenge for the next years is to create efficient and robust implementations of analytical methods that are required for the biomedical and health domain, within these frameworks. This will require that existing analytical algorithms be transformed to fit the distributed processing model, while ensuring confidentiality of the data being analyzed.

Another recurring challenge when working with big data is the need to circumvent manual steps during any operation on the data, be it for storage, preprocessing, or analysis. This is particularly true when the number of variables (i.e., dimensions) of the dataset is large, and the step has to be carried out for each variable separately. Complete automation of the entire storage, preprocessing and analysis pipeline is then imperative, purely for reasons of feasibility. At the same time, there must be guarantees that the results from such analyses are accurate and reliable. This is another major issue for the big data research agenda in the next years.

Acknowledgements

Niels Peek and John Holmes chair the IMIA working group on Data Mining and Big Data Analytics. The members of the working group are gratefully acknowledged for their collaboration and fruitful discussions.

References

1. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform 2008:128-44.
2. Friedman C, Elhadad N. Natural language processing in health care and biomedicine. In: Shortliffe EH, Cimino JJ (eds.). Biomedical Informatics. Computer Applications in Health Care and Biomedicine (4th ed.), London: Springer, 2014. p. 255-84.
3. Deserno TM. Biomedical Image Processing. Berlin: Springer; 2011.
4, Rubin DL, Greenspan H, Brinkley JF. Biomedical Imaging Informatics. In: Shortliffe EH, Cimino JJ, editors. Biomedical Informatics. Computer Applications in Health Care and Biomedicine (4th ed.), London: Springer; 2014, p. 285-327.
5. Eysenbach G, Köhler C. Health-Related Searches on the Internet. JAMA 2004;291:2946.
6. Carneiro HA, Mylonakis E. Google trends: a web-based tool for real-time surveillance of disease outbreaks. Clin Infect Dis 2009;49(10):1557-64.
7. Mandl KD, Overhage JM, Wagner MM, Lober WB, Sebastiani P, Mostashari F, et al. Implementing syndromic surveillance: a practical guide informed by the early experience. J Am Med Inform Assoc 2004;11:141-50.
8. 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):404-8.
9. New Tweets per second record, and how! Twitter, Inc 2014 [cited 2014 Jan 15]; Available from: URL: https://blog.twitter.com/2013/new-tweets-per-second-record-and-how
10. Langmead B, Schatz MC, Lin J, et al. Searching for SNPs with cloud computing. Genome Biol 2009;10:R134.
11. Wang Y, Goh W, Wong L, Montana G; Alzheimer's Disease Neuroimaging Initiative. Random forests on Hadoop for genome-wide association studies of multivariate neuroimaging phenotypes. BMC Bioinformatics 2013;14 Suppl 16:S6.
12. Ng K, Ghoting A, Steinhubl SR, Stewart WF, Malin B, Sun J. PARAMO: A PARAllel predictive MOdeling platform for healthcare analytic research using electronic health records. J Biomed Inform 2014;48:160-70.
13. Sahoo SS, Jayapandian C, Garg G, Kaffashi F, Chung S, Bozorgi A, et al. Heart beats in the cloud: distributed analysis of electrophysiological ‚Big Data' using cloud computing for epilepsy clinical research. J Am Med Inform Assoc 2014;21(2):263-71.
14. Zhao S, Prenger K, Smith L, Messina T, Fan H, Jaeger E, et al. Rainbow: A tool for large-scale whole-genome sequencing data analysis using cloud computing. BMC Genomics 2013;14:425.
15. Hird SM. LociNGS: A Lightweight Alternative for Assessing Suitability of next-Generation Loci for Evolutionary Analysis. PloS One 2012;7(10):e46847.
16. Chang F, Dean J, Ghemawat S, Hsieh WC, Wallach DA, Burrows M, et al. Bigtable: A distributed storage system for structured data. ACM Trans Comput Syst 2008;26(2):4.
17. Rosenthal A, Mork P, Li MH, Stanford J, Koester D, Reynolds P. Cloud computing: a new business paradigm for biomedical information sharing. J Biomed Inform 2010;43(2):342-53.
18. Amazon Web Services. Creating Healthcare Data Applications to Promote HIPAA and HITECH Compliance. White paper, Amazon, August 2012. http://media.amazonwebservices.com/AWS_HIPAA_Whitepaper_Final.pdf (last accessed 20 May 2014)
19. Jeffrey D, Ghemawat S. MapReduce: Simplified data processing on large clusters. Sixth Symposium on Operating Systems Design & Implementation (OSDI); 2004:137–50.
20. Ng K, Ghoting A, Steinhubl SR, Stewart WF, Malin B, Sun J. PARAMO: A PARAllel Predictive MOdeling Platform for Healthcare Analytic Research Using Electronic Health Records. J Biomed Inform 2013, doi:10.1016/j.jbi.2013.12.012.
21. Xin RS, Rosen J, Zaharia M, Franklin MJ, Shenker S, Stoica I. Shark: SQL and rich analytics at scale. ACM SIGMOD Conference, 2013. 1145/2463676.2465288.
22. Low Y, Gonzalez J, Kyrola A, Bickson D, Guestrin C, Hellerstein JM. Graphlab: A new parallel framework for machine learning. In: Grünwald P, Spirtes P, editors. Proc 26th Conference on Uncertainty in Artificial Intelligence. AUAI Press; 2010. p. 340-9.
23. McPheeters ML, Sathe NA, Jerome RN, Carnahan RM. Methods for systematic reviews of administrative database studies capturing health outcomes of interest. Vaccine 2013 ;31(Suppl 10):K2-6.
24. Greenwald P, Friedlander BR, Lawrence CE, Hearne T, Earle K. Diagnostic sensitivity bias -- an epidemiologic explanation for an apparent brain tumor excess. J Occup Med 1981;23(10):690-4.
25. Tessier-Sherman B, Galusha D, Taiwo OA, Cantley L, Slade MD, Kirsche SR, et al. Further validation that claims data are a useful tool for epidemiologic research on hypertension. BMC Public Health 2013;13:51.
26. Ritchie MD, Denny JC, Crawford DC, Ramirez AH, Weiner JB, Pulley JM, et al. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. Am J Hum Genet 2010;86(4):560-72.
27. D'Agostino RB Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med 1998;17(19):2265-81.
28. De Vries H, Kemps HMC, Van Engen-Verheul MM, Kraaijenhagen RA, Peek N. Cardiac abilitation and survival in a large representative community cohort of Dutch patients. Submitted for publication.
29. Friedman JH. Greedy function approximation: A gradient boosting machine. Annals of Statistics 2001;29(5):1189-232.
30. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science 1996;273:1516–7.
31. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 2005 Feb;6(2):95-108.
32. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2009 Feb 19;457(7232):1012-4.

Correspondence to:
Niels Peek
Centre for Health Informatics
The University of Manchester
Vaughan House
Portsmouth Street
Manchester M13 9GB, United Kingdom
E-mail: niels.peek@manchester.ac.uk