Review

Big Data and Clinicians: A Review on the State of the Science

Weiqi Wang, PhD; Eswar Krishnan, MD, MPH

School of Medicine, Stanford University, Palo Alto, CA, United States

Corresponding Author:
Eswar Krishnan, MD, MPH
School of Medicine
Stanford University
1000 Welch Road, Suite 203
Palo Alto, CA, 94304
United States
Phone: 1 650 725 8004
Fax: 1 650 723 9656
Email: e.krishnan@stanford.edu

Abstract

Background: In the past few decades, medically related data collection saw a huge increase, referred to as big data. These huge datasets bring challenges in storage, processing, and analysis. In clinical medicine, big data is expected to play an important role in identifying causality of patient symptoms, in predicting hazards of disease incidence or reoccurrence, and in improving primary-care quality.

Objective: The objective of this review was to provide an overview of the features of clinical big data, describe a few commonly employed computational algorithms, statistical methods, and software toolkits for data manipulation and analysis, and discuss the challenges and limitations in this realm.

Methods: We conducted a literature review to identify studies on big data in medicine, especially clinical medicine. We used different combinations of keywords to search PubMed, Science Direct, Web of Knowledge, and Google Scholar for literature of interest from the past 10 years.

Results: This paper reviewed studies that analyzed clinical big data and discussed issues related to storage and analysis of this type of data.

Conclusions: Big data is becoming a common feature of biological and clinical studies. Researchers who use clinical big data face multiple challenges, and the data itself has limitations. It is imperative that methodologies for data analysis keep pace with our ability to collect and store data.

(JMIR Med Inform 2014;2(1):e1) doi:10.2196/medinform.2913

KEYWORDS

big data; database; medical informatics; clinical research; medicine

Introduction

Big data refers to very large datasets with complex structures that are difficult to process using traditional methods and tools. The term process includes, capture, storage, formatting, extraction, curation, integration, analysis, and visualization [1-9]. A popular definition of big data is the "3V" model proposed by Gartner [10], which attributes three fundamental features to big data: high volume of data mass, high velocity of data flow, and high variety of data types. The notion of big data can be traced back to the 1970s [11-13] when scientists realized that they lacked the tools to analyze datasets of large size. In those days, big data was merely several to hundreds of megabytes [14]; now datasets of terabytes are common [15, 16]. Therefore, the "big" in big data reflects the limits of data storage and computational power existing at a given point in time.

Table 1 shows the growth of global big data volume and computer science papers on big data since 2009. This table exemplifies that stored data will be in the tens of zettabytes range by 2020, and research on how to deal with big data will grow exponentially as well.

Big data is gathered in many disciplines and is made possible by ubiquitous information-sensing devices and software [19]. An example is web logs: websites such as Google or Facebook automatically record user information at each visit. Other examples come from the stock market [20], earthquake

surveillance [21], political elections [22], behavioral studies [23], sports [24], pharmaceutical reports [25], health care [26, 27], electronic medical records [28], imaging data [29], genome data [30, 31], and entrepreneur transaction records [32]. Data collection is sometimes interdisciplinary. As an example, a sudden increase in Google search terms such as "flu symptoms" and "flu treatments" can be used to predict an increase in flu patients presenting to hospital emergency rooms [33]. This example also demonstrates that big data has promising predictive power and return on investment. Return on investment of big data has also been suggested for clinical big data [34, 35].

Although arguably valuable, big data is difficult to analyze due to the massive volume of the raw data and its diversity, as shown in Figure 1. Therefore, instead of the raw big data, a large dataset is often extracted from the raw data to generate a secondary storage of data for analysis purposes. This data extraction is applied, for example, when a computer tomography scan is involved in clinical trials and only the physician diagnosis based on the scan is included in data analysis. Similarly, a large volume of descriptive data on various kinds of samplings, tests, or assays can be extracted with only the key parameters kept. As a consequence, the data analyzed in clinical medicine is usually from secondary datasets that contain only data of interest. The secondary datasets, although still large, are not terabytes in size. Additionally, due to the nature of clinical trials, a large dataset in clinical medicine usually does not have an overwhelming number of samples. Kjaergard et al [36] defined clinical trials with 1000 or more participants as large, and the studies in clinical medicine titled big/large, data/dataset generally have thousands of attributes, but only hundreds of samples [37-39].

For this paper, we reviewed the literature to determine the features of clinical big data and determine the methods used for manipulation and analysis of these data. This paper is focused on clinical medicine rather than general health care issues; therefore, we mainly reviewed the studies that appeared relevant to clinicians. We examined the selected studies to extract information on research interests, goals, and achievements, and the implemented methodologies. Our intention was not to conduct an exhaustive systematic review, but instead to enable a literature-based discussion of how the big data issue has been addressed in clinical medicine. Based on our findings, we discuss the challenges and limitations of analysis of large clinical datasets.

Table 1. Global growth of big data and computer science papers on big data.

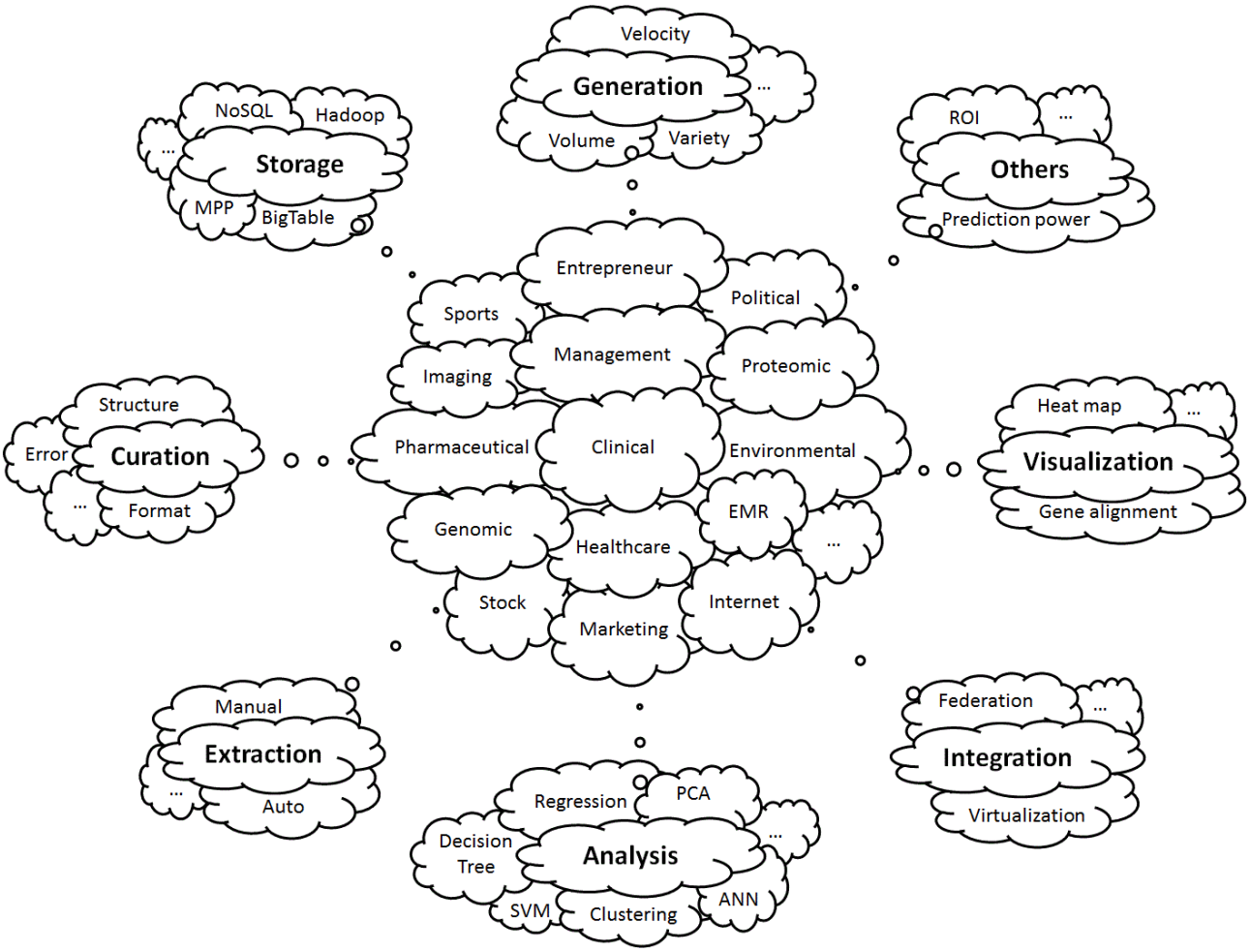
Year	Data volume, ZB[a,c]	Conference papers, CS[b,c]	Journal papers, CS[c]
2009	1.5	12	7
2010	2	26	7
2011	2.5	32	23
2012	3	78	47
2015	8	?	?
2020	44	??	??

[a]Data from *oracle* [17].

[b]Data from *Research Trends* [18].

[c]CS, computer science; ZB, zettabytes (1 zettabyte = 1000 terabytes = 10^6 petabytes = 10^{18} gigabytes, GB).

Figure 1. A schematic of the issues surrounding storage and use of big data. Clinical big data, as well as big data in other disciplines, have been surrounded by a number of issues and challenges, including (but not limited to): generation, storage, curation, extraction, integration, analysis, visualization, etc. ANN: artificial neuron network; EMR: electronic medical record; MPP: massively parallel-processing; PCA: principle component analysis; ROI: return of investment; SVM: support vector machine.



Methods

We conducted a literature review to identify studies on big data in medicine, especially clinical medicine. We used different combinations of keywords to search PubMed, Science Direct, Web of Knowledge, and Google Scholar for literature of interest, mainly from the last 10 years. The key words were: "big data medicine", "large dataset medicine", “clinical big data”, “clinical large dataset”, “clinical data warehouse”, “clinical database”, “clinical data mining”, “biomedical big data”, “biomedical database”, “biomedical data warehouse”, “healthcare big data”, “healthcare database”, and “healthcare data warehouse”.

Results

Big Data in Clinical Medicine

Big data plays an important role in medical and clinical research and has been leveraged in clinically relevant studies. Major research institute centers and funding agencies have made large investments in the arena. For example, the National Institutes of Health recently committed US $100 million for the big data to Knowledge (BD2K) initiative [40]. The BD2K defines “biomedical” big data as large datasets generated by research groups or individual investigators and as large datasets generated by aggregation of smaller datasets. The most well-known examples of medical big data are databases maintained by the Medicare and Healthcare Cost and Utilization Project (with over 100 million observations). One of the differences between medical big data and large datasets from other disciplines is that clinical big data are often collected based on protocols (ie, fixed forms) and therefore are relatively structured, partially due to the extraction process that simplify raw data as mentioned above. This feature can be traced back to the Framingham Heart Study [41], which has followed a cohort in the town of Framingham, Massachusetts since 1948. Vast amounts of data have been collected through the Framingham Heart Study, and the analysis has informed our understanding of heart diseases, including the effects of diet, exercise, medications, and obesity on risk [42]. There are many other clinical databases with different scopes, including but not limited to, prevalence and trend studies, risk factor studies, and genotype-phenotype studies.

Prevalence and Trend Studies

One of the major uses for clinical big data is in analysis of the prevalence or trends of a disease or phenotype among different populations. An early big data study evaluated a cohort consisting of 890,394 US veterans with diabetes followed from 2002 through 2006 [43]. Bermejo-Sanchez et al [44] observed 326 of the birth defect Amelia among 23 million live births,

stillbirths, and fetal anomalies from 23 countries and 4 continents, and found the trend of higher prevalence of Amelia among younger mothers. Histological features that differ between chronic idiopathic inflammatory bowel disease and normality and between Crohn's disease and ulcerative colitis were identified in 809 large bowel endoscopic biopsies [45]. Kelly et al [46] studied the prevalence of hip abnormalities of 8192 subjects with hemophilia A and B. Siregar et al [47] performed a population-based study on patients after cardiac surgery in all 16 cardiothoracic surgery centers in the Netherlands. Elshazly et al [48] examined 1.3 million US adults for patient-level discordance of non-high-density lipoprotein cholesterol and low-density lipoprotein cholesterol. Chan and McGarey [49] summarize how large datasets can be analyzed to achieve population-based conclusions, specifically for determination of secular trends, health disparities, geographic variation, and evaluation of specific diseases and treatments. This paper also summarized the strengths and limitations of large-sized datasets and addressed issues such as missing data and bias. These issues will also be discussed in brief below.

Risk Factor Studies

Clinical big data can also be used to determine causality, effect, or association between risk factors and the disease of interest. Ursum et al [50] examined the relationships between seroconversion and patient age with inflammatory effects of autoantibodies in 18,658 rheumatoid arthritis patients and controls, and showed that citrullinated proteins and peptides were more reliable markers for rheumatoid arthritis than was Immunoglobulin M rheumatoid factor. Ajdacic-Gross et al [51] examined the data on 11,905 Swiss conscripts from 2003 for stuttering and found that there was no single overwhelming risk factor for stuttering, although premature birth and parental alcohol abuse appeared influential. Data collected on 14,433 patients from the 155 Veterans Administration medical centers in all 50 US states, Puerto Rico, and the District of Columbia were used to identify the alcohol dependence of medications [52]. By analysis of 53,177 cases of contrast administration in 35,922 patients from the Radiology and Cardiac Catheterization Laboratory databases, an increase in contrast nephropathy was associated with use of sodium bicarbonate [53]. Echocardiography and electrocardiogram-gated single-photon emission computed tomography traces for the evaluation of left ventricular ejection fraction were compared in 534 patients [54]. Zhang et al [55] examined clinical data of 16,135 adult patients and elucidated the relationships between glycemic, blood glucose level, and intake of insulin with mortality. Mitchel et al [56] studied the effect of 2 types of insulin on 7720 patients selected from 8 million in UK. Kobayashi et al [57] analyzed 19,070 records on right hemicolectomy from 3500 Japanese hospitals and successfully developed a risk model. It should be noted that in these studies, the terms of "association" and "causality" must be rigorously distinguished; most of the studies claimed association, whereas causality was rarely asserted.

Genotype–Phenotype Studies

With the advancement of genotyping technology, an increasing amount of risk-factor studies have attempted to assess association on the genetic level through evaluation of gene expression and/or genomic data obtained from patients and controls. For example, clinical and genetic data from 5700 patients who had been treated with warfarin were used to create an algorithm to estimate the appropriate dose [58]. Causality of autism spectrum disorders has been investigated by analysis of 31,516 clinical cases on copy number variation in patients versus 13,696 controls [59]. Koefoed et al [60] made efforts to assess the effects of signal transmission and calculated all combinations of three genotypes from 803 single-nucleotide polymorphism (SNP) genotypes (2.3 billion combinations) for 1355 controls and 607 patients with bipolar disorder. These studies are similar to risk-factor studies, yet often the big data is significantly larger in volume due in genetic analyses than in risk-factor studies.

Method Development Studies

A number of studies have taken advantage of clinical big data to establish new methods or techniques, or to develop new tools to enable analysis of data and decision making. In a typical example, Hill et al [61] designed an interface to use clinical data to evaluate risk ratios for various diseases to aid in evaluation of treatment options. Liu et al [62, 63] have used large-scale data analysis to optimize diagnosis of breast cancer from full-field digital mammography images. Lin et al [64] made efforts to formalize the phenotype variable in the database Genotypes and Phenotypes. Stephen et al [65] developed an algorithm to categorize pediatric patients presenting with respiratory distress into different subtypes using clinical variables from a clinical data warehouse. Clinical data warehouses or databases have been created from radiotherapy clinical trial data [66], gene mutations [67], cancer patient data [68, 69], kidney disease patient data [70], and gastrointestinal surgery patient data [71]. Additionally, studies have focused on personalized big data [72], citizen-centric health care versus patient-centric health care [72, 73], medication orders [74, 75], and decision making and information management/retrieval in general [75-80]. The dramatic increase in the number of studies with large scope in the past few years indicates an increasing desire of researchers to manipulate clinical big data; "big data-assisted clinics" may be expected in the near future.

Discussion

Diversity of Data in Clinical Medicine

The huge body of medical research that has been performed using large datasets demonstrates the broad spectrum of data resources used and shows that the structure of the medical dataset depends on the research question. Data from different subareas of medical research have broad diversity in terms of numbers of entries, types of data stored (or levels), dimensionality, and sample size [81]. Datasets obviously differ greatly in size: gene expression datasets derived from high-throughput microarray and next-generation sequencing technologies, such as those that analyze SNPs and copy number variations, tend to be massive, whereas clinical trial dataset are not as big. Phan et al [82] suggested that data in medicine be divided into four different levels: the molecular level (eg, genome data), cellular and tissue level (eg, stem cell differentiation data), clinical and patient level (eg, clinical trial data), and biomedical knowledge base level (ie, a comprehensive

data collection). Additionally, data tend to have different levels of dimensionality (ie, number of attributes or parameters, *p*) and sample sizes (ie, number of records/entries, n). Typical datasets fall into one of three categories, as summarized by Sinha et al [83]: large n, small *p*; small n, large *p*; and large n, large *p*. Thanks to advancements in computational technology, most algorithms can handle low-dimensional data (ie, large n, small *p*) without encountering significant difficulty.

Most clinical data, however, is high-dimensional (ie, small n, large *p* or large n, large *p*) due to a limited number of patients. One typical example comes from a study of 69 Broca's aphasic patients (ie, n=69) who were tested with nearly 6000 stimulus sentences (ie, *p*~6000) [84]. With similar dimensionality, Mitchell et al [39] studied bipolar disorder where the sample consisted of only 217 patients. For high-dimensional data, each point, sample, or element is described by many attributes [83] with the involvement of the "curse of dimensionality" [85]. Because high-dimensional data are sparse in dimensions, most classification or clustering approaches do not work well because the increase in problem space reduces the overall density of data samples. To solve this problem, compression methods and significance testing are usually used to either reduce the dimensionality or select relevant features before data analysis by some sort of data preprocessing [83].

Methods for Manipulation of Clinical Big Data

Technologies for Data Storage and Handling

Due to the massiveness and complexity of big data, nonrelational and distributed databases such as Apache Hadoop [86], Google BigTable [87], NoSQL [88], and massively parallel-processing databases are used rather than traditional relational databases to store data. A large number of biostatistics software packages have been used to handle large clinical datasets, some of which enabled the features of cloud-based or distributed computing. Popular software packages include, but are not limited to, SAS [36, 51-53], Mplus [51], SPSS [36, 39, 45], PP-VLAM [89], Stata [90], and R [91]. These technologies and tools greatly facilitate the handling of big data.

Methodologies for Data Preprocessing

Clinical raw big data can be highly diverse and uninformative without preprocessing. Extraction of a diagnosis from raw computer tomography data is an example of one of the predominant manners in which clinical big data are preprocessed. This type of processes relies on a specialist's personal expertise and can be a source of bias. Most early analyses of big data, including that collected by the Framingham Heart Study adopted some form of preprocessing; therefore, challenges exist in curation [6]. As an alternative to expert preprocessing, computational algorithms or statistical approaches, including compression methods, significance testing, or normalization [92] can be implemented to preprocess raw big data. This methodology may also introduce bias and can cause uncertainty problems during data integration.

In some scenarios, visualization can be a part of data preprocessing (as well as result exhibition). Typical examples in this regard include the use of heat maps [93], gene alignments [94], protein structure visualization [95], scatterplot matrix, tree visualization, network visualization, parallel coordinates, stacked graphs, etc. When the big data of interest are scattered or stored at different resources, data integration [96, 97] and federation [98] is an important phase during data preprocessing. Approaches such as the Information Manifold [97], which allows browsing and querying of multiple networked information sources, can provide solutions to uncertainty problems after data integration and mapping [99].

Statistical Approaches to Data Analysis

A number of popular statistical methods have been implemented in clinical data analysis. The most common include linear regression and logistic regression [30], latent class analysis [100], principle component analysis [101], and classification and regression trees [100]. Additionally, logarithmic and square-root transformations [58], naive Bayes methods [102], decision trees [103], neural networks [104], support vector machines [105], and hidden Markov models [83] are also used to study problems in medical data.

When a dataset is not overly complicated, a single test (eg, a simple Student's *t* test) should be powerful enough to reject a null hypothesis, and single hypothesis testing is the methodology to adopt [106]. Sometimes one cannot establish the significance of a hypothesis until different statistical tests have been applied to the same dataset. Multiple testing is often used to identify correlations that deserve further investigation [107]. Algorithms for false discovery rate [108] and family-wise error rate [109] calculation have been implemented for multiple testing in studies on gene expression data and datasets with similar levels of complexity.

Challenges and Limitations of Use of Clinical Big Data

Overview

Big data itself has many limitations. These limitations include "adequacy, accuracy, completeness, nature of the reporting sources, and other measures of the quality of the data", as summarized previously [110]. The consequences of these limitations are succinctly summarized in the book titled "Models. Behaving. Badly." [111]. Modeling can often lead to a biased statistical correlation or inference, sometimes known as a "false discovery". Clinical big data users face a large spectrum of challenges, including but not limited to sample size, selection bias, interpretation problem, missing values, dependence problems, and data handling methodologies.

Sample Size

One of the counterintuitive challenges in analysis of big data clinical datasets is that sometimes the sample size is not as big, compared with the number of attributes to allow statistically significant analysis. Population survey methods are sometimes adopted because these methods can provide larger datasets. However, the authenticity and accuracy of this type of data are arguably limited; hence, survey methods cannot be reliably used to produce an adequate description or prediction [39].

Selection Bias

Any dataset is a selection of data rather than the whole data world; therefore, selection bias is a very real limitation [112]

even if the sample size is big. In that sense, all studies of clinical data have this limitation to some degree [39].

Interpretation Problem

Gebregziabher et al [43] stated that the datasets generated through many translational research projects to answer questions of public health interest are not self-explanatory due to complexity and inadequate description/documentation of the dataset's parameters and associated metadata. The methodologies for interpreting the data can therefore be subject to all sorts of philosophical debate. For example, the data may not be totally naïve or objective and interpretation may be biased by subjective assumptions and/or manipulations by individual analysts.

Missing Values

It is common problem that large datasets have missing values, and in many cases the problem can be significant [44]. A typical example is the Framingham Heart Study where data on serum uric acid are largely missing. Additionally, the covariates (ie, attributes) may not fully capture the degree of risk for patients and may cause uncertainty in results [53].

Dependence Problems

One issue that has been often neglected is the dependence of data. Dependence between either attributes or samples in datasets can cause the degrees of freedom to decrease and/or some statistical principles to no longer apply. Examples of this are found when the same patients are evaluated multiple times through follow-up and when correlations in gene expression are drawn based on samples from different patients treated with similar medications [83]. As many statistical methods do not account for dependence, results from these tests may be unreliable if this issue is not properly addressed before the data analysis.

Data Handling Methodologies

Effective processing of big data has always been a challenge. One must consider all the aspects of the dataset, including collection, curation, extraction, integration, interpretation, imputation, and selection of appropriate statistical methods, during processing and analysis. It has been claimed that analyses of large datasets are often suboptimal due to the researcher's lack of knowledge of the available tools and methodologies [83]. On the other hand, algorithms to handle big data are also underdeveloped to some extent and deserve more attention [113].

Conclusions

This paper reviewed studies that analyzed clinical big data and that discuss issues related to data storage and analysis. Big data is becoming a common feature of biological and clinical studies. Today, a single biophysical researcher can generate terabytes of data in hours. Over the last decade, clinical datasets have grown incredibly in size, mostly due to use of modern technologies for collection and recording of data. Researchers who use clinical big data face multiple challenges, and the data itself has limitations. It is imperative that methodologies for data analysis keep pace with our ability to collect and store data.

Authors' Contributions

WW did the search and primary review of the literature cited in this article and wrote the manuscript. EK guided the research and critically revised the manuscript.

Conflicts of Interest

None declared.

References

1. Wenkebach U, Pollwein B, Finsterer U. Visualization of large datasets in intensive care. Proc Annu Symp Comput Appl Med Care 1992:18-22 [FREE Full text] [Medline: 1482864]
2. Wang J, Chen Y, Hua R, Wang P, Fu J. A distributed big data storage and data mining framework for solar-generated electricity quantity forecasting. Proc. SPIE 8333, Photonics and Optoelectronics Meetings (POEM) 2011 [FREE Full text] [doi: 10.1117/12.919640]
3. Wang JZ, Chen YJ, Hua R, Wang P, Fu J. A distributed big data storage and data mining framework for solar-generated electricity quantity forecasting. Proc. SPIE 8333, Photonics and Optoelectronics Meetings (POEM) 2011 2011 [FREE Full text] [doi: 10.1117/12.919640]
4. Fu J, Chen ZH, Wang JC, He MQ, Wang JZ. Distributed storage system big data mining based on HPC application-A solar photovoltaic forecasting system practice. Information-Tokyo 2012;15(9):3749-3755.
5. Brinkmann BH, Bower MR, Stengel KA, Worrell GA, Stead M. Large-scale electrophysiology: acquisition, compression, encryption, and storage of big data. J Neurosci Methods 2009;180(1):185-192 [FREE Full text] [doi: 10.1016/j.jneumeth.2009.03.022] [Medline: 19427545]
6. Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, et al. Big data: The future of biocuration. Nature 2008;455(7209):47-50 [FREE Full text] [doi: 10.1038/455047a] [Medline: 18769432]
7. Calatroni A, Roggen D, Troster G. Collection and curation of a large reference dataset for activity recognition. Ieee Sys Man Cybern 2011 2011:30-35. [doi: 10.1109/ICSMC.2011.6083638]
8. O'Driscoll A, Daugelaite J, Sleator RD. 'Big data', Hadoop and cloud computing in genomics. J Biomed Inform 2013;46(5):774-781. [doi: 10.1016/j.jbi.2013.07.001] [Medline: 23872175]

9. Lee KK, Tang WC, Choi KS. Alternatives to relational database: comparison of NoSQL and XML approaches for clinical data storage. Comput Methods Programs Biomed 2013;110(1):99-109. [doi: 10.1016/j.cmpb.2012.10.018] [Medline: 23177219]
10. Beyer MA, Douglas L. The Importance of 'Big Data': A Definition. 2012 URL: http://www.gartner.com/it-glossary/big-data/ [accessed 2013-08-25] [WebCite Cache ID 6J7x1gswQ]
11. Olbers D, Müller P, Willebrand J. Inverse technique analysis of a large data set. Physics of the Earth and Planetary Interiors 1976;12(2-3):248-252. [doi: 10.1016/0031-9201(76)90054-6]
12. Byth DE, Eisemann RL, de Lacy IH. Two-way pattern analysis of a large data set to evaluate genotypic adaptation. Heredity 1976;37(2):215-230. [doi: 10.1038/Hdy.1976.84]
13. Chaudron J, Assenlineau L, Renon H. A new modification of the Redlich—Kwong equation of state based on the analysis of a large set of pure component data. Chemical Engineering Science 1973;28(3):839-846. [doi: 10.1016/0009-2509(77)80018-3]
14. Graefe JF, Wood RW. Dealing with large data sets. Neurotoxicol Teratol 1990;12(5):449-454. [Medline: 2247031]
15. Ackerman MJ. Big data. J Med Pract Manage 2012;28(2):153-154. [Medline: 23167038]
16. Trelles O, Prins P, Snir M, Jansen RC. Big data, but are we ready? Nat Rev Genet 2011 Mar;12(3):224. [doi: 10.1038/nrg2857-c1] [Medline: 21301471]
17. analysis ATK. Big Data and the Creative Destruction of Today's Business Models. 2013 URL: http://www.atkearney.com/strategic-it/ideas-insights/article/-/asset_publisher/LCcgOeS4t85g/content/big-data-and-the-creative-destruction-of-today-s-business-models/10192 [accessed 2013-11-25] [WebCite Cache ID 6LOmyDbvg]
18. Halevi G, Moed HF. The Evolution of Big Data as a Research and Scientific Topic: Overview of the Literature. 2012 URL: http://www.researchtrends.com/issue-30-september-2012/the-evolution-of-big-data-as-a-research-and-scientific-topic-overview-of-the-literature/ [accessed 2013-11-25] [WebCite Cache ID 6LOn9qSGu]
19. Hutchins J, Ihler A, Smyth P. Probabilistic analysis of a large-scale urban traffic sensor data set. Knowledge Discovery from Sensor Data 2010;5840:94-114. [doi: 10.1007/978-3-642-12519-5_6]
20. Sleutel S, De Neve S, Beheydt D, Li C, Hofman G. Regional simulation of long-term organic carbon stock changes in cropland soils using the DNDC model: 1. Large-scale model validation against a spatially explicit data set. Soil Use Manage 2006;22(4):342-351. [doi: 10.1111/j.1475-2743.2006.00045.x]
21. Cianchini G, De Santis A, Balasis G, Mandea M, Qamili E. Entropy based analysis of satellite magnetic data for searching possible electromagnetic signatures due to big earthquakes. Ma Comput Sci Eng 2009:29-35 [FREE Full text]
22. Issenberg S. How President Obama's campaign used big data to rally indicidual voters. Technol Rev 2013;116(1):38-49 [FREE Full text]
23. Kessler RC, Brown RL, Broman CL. Sex differences in psychiatric help-seeking: evidence from four large-scale surveys. J Health Soc Behav 1981 Mar;22(1):49-64. [Medline: 7240706]
24. Lewis M. Moneyball: The Art of Winning an Unfair Game: W. In: Moneyball: The Art of Winning an Unfair Game. New York, New York: W. W. Norton & Company; 2003.
25. Ekins S, Williams AJ. When pharmaceutical companies publish large datasets: an abundance of riches or fool's gold? Drug Discov Today 2010;15(19-20):812-815. [doi: 10.1016/j.drudis.2010.08.010] [Medline: 20732447]
26. Grimley Evans J, Tallis RC. A new beginning for care for elderly people? BMJ 2001;322(7290):807-808 [FREE Full text] [Medline: 11290619]
27. Jee K, Kim GH. Potentiality of big data in the medical sector: focus on how to reshape the healthcare system. Healthc Inform Res 2013;19(2):79-85 [FREE Full text] [doi: 10.4258/hir.2013.19.2.79] [Medline: 23882412]
28. Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA 2013;309(13):1351-1352. [doi: 10.1001/jama.2013.393] [Medline: 23549579]
29. Toga AW. The clinical value of large neuroimaging data sets in Alzheimer's disease. Neuroimaging Clin N Am 2012;22(1):107-118, ix [FREE Full text] [doi: 10.1016/j.nic.2011.11.008] [Medline: 22284737]
30. Bakke PS, Zhu G, Gulsvik A, Kong X, Agusti AG, Calverley PM, et al. Candidate genes for COPD in two large data sets. Eur Respir J 2011;37(2):255-263 [FREE Full text] [doi: 10.1183/09031936.00091709] [Medline: 20562129]
31. Solomon BD, Nguyen AD, Bear KA, Wolfsberg TG. Clinical genomic database. Proc Natl Acad Sci U S A 2013;110(24):9851-9855 [FREE Full text] [doi: 10.1073/pnas.1302575110] [Medline: 23696674]
32. Liu XL, Du JP, Li WZ, Zuo M, Han ZM. Data warehousing for data mining based on Olap technology. In: Ciict 2008: Proceedings of China-Ireland International Conference on Information and Communications Technologies. 2008 Presented at: China-Ireland International Conference on Information and Communications Technologies; 26-28 Sept. 2008; Beijing, China p. 176-179. [doi: 10.1049/cp:20080786]
33. Lohr S. The New York Times. 2012. The Age of Big Data URL: http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?pagewanted=1&_r=0 [WebCite Cache ID 6J7xJAwId]
34. McCann E. EHR Boosts ROI. 2013 URL: http://www.healthcareitnews.com/news/ehr-boost-roi-revenue-medical-group [accessed 2013-11-25] [WebCite Cache ID 6LOnNYQHV]

35. group CIW. EMR Benefits and Return on Investment Categories. 2008 URL: http://www.informatics-review.com/wiki/index.php/EMR_Benefits_and_Return_on_Investment_Categories [accessed 2013-11-25] [WebCite Cache ID 6LOndM19u]
36. Kjaergard LL, Villumsen J, Gluud C. Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. Ann Intern Med 2001 Dec 4;135(11):982-989. [Medline: 11730399]
37. Mancia G, Omboni S, Ravogli A, Parati G, Zanchetti A. Ambulatory blood pressure monitoring in the evaluation of antihypertensive treatment: additional information from a large data base. Blood Press 1995;4(3):148-156. [Medline: 7670648]
38. Reichelt JG, Heimdal K, Møller P, Dahl AA. BRCA1 testing with definitive results: a prospective study of psychological distress in a large clinic-based sample. Familial Cancer 2002;3(1):21-28. [doi: 10.1023/B:FAME.0000026820.32469.4a]
39. Mitchell PB, Johnston AK, Corry J, Ball JR, Malhi GS. Characteristics of bipolar disorder in an Australian specialist outpatient clinic: comparison across large datasets. Aust N Z J Psychiatry 2009;43(2):109-117. [doi: 10.1080/00048670802607220] [Medline: 19153918]
40. NHGRI. Request for Information (RFI): Input on Development of Analysis MethodsSoftware for Big Data. 2013 URL: http://grants.nih.gov/grants/guide/notice-files/NOT-HG-13-014.html [accessed 2013-08-27] [WebCite Cache ID 6JB6Y33gd]
41. Wolf PA, Abbott RD, Kannel WB. Atrial fibrillation as an independent risk factor for stroke: the Framingham Study. Stroke 1991;22(8):983-988 [FREE Full text] [Medline: 1866765]
42. Hubert HB, Feinleib M, McNamara PM, Castelli WP. Obesity as an independent risk factor for cardiovascular disease: a 26-year follow-up of participants in the Framingham Heart Study. Circulation 1983;67(5):968-977. [Medline: 6219830]
43. Gebregziabher M, Egede L, Gilbert GE, Hunt K, Nietert PJ, Mauldin P. Fitting parametric random effects models in very large data sets with application to VHA national data. BMC Med Res Methodol 2012;12:163 [FREE Full text] [doi: 10.1186/1471-2288-12-163] [Medline: 23095325]
44. Bermejo-Sánchez E, Cuevas L, Amar E, Bakker MK, Bianca S, Bianchi F, et al. Amelia: a multi-center descriptive epidemiologic study in a large dataset from the International Clearinghouse for Birth Defects Surveillance and Research, and overview of the literature. Am J Med Genet C Semin Med Genet 2011;157C(4):288-304. [doi: 10.1002/ajmg.c.30319] [Medline: 22002956]
45. Cross SS, Harrison RF. Discriminant histological features in the diagnosis of chronic idiopathic inflammatory bowel disease: analysis of a large dataset by a novel data visualisation technique. J Clin Pathol 2002;55(1):51-57 [FREE Full text] [Medline: 11825925]
46. Kelly D, C Zhang Q, M Soucie J, Manco-Johnson M, Dimichele D, Joint Outcome Subcommittee of the Coordinating Committee for the Universal Data Collection Databasethe Hemophilia Treatment Center Network Investigators. Prevalence of clinical hip abnormalities in haemophilia A and B: an analysis of the UDC database. Haemophilia 2013;19(3):426-431. [doi: 10.1111/hae.12073] [Medline: 23252621]
47. Siregar S, Roes KC, van Straten AH, Bots ML, van der Graaf Y, van Herwerden LA, et al. Statistical methods to monitor risk factors in a clinical database: example of a national cardiac surgery registry. Circ Cardiovasc Qual Outcomes 2013;6(1):110-118. [doi: 10.1161/CIRCOUTCOMES.112.968800] [Medline: 23322806]
48. Elshazly MB, Martin SS, Blaha MJ, Joshi PH, Toth PP, McEvoy JW, et al. Non-high-density lipoprotein cholesterol, guideline targets, and population percentiles for secondary prevention in 1.3 million adults: the VLDL-2 Study (very large database of lipids). J Am Coll Cardiol 2013 Nov 19;62(21):1960-1965. [doi: 10.1016/j.jacc.2013.07.045] [Medline: 23973689]
49. Chan L, McGarey P. Using large datasets for population-based health research. In: Gallin JI, Ognibene FP. eds. Principles and Practice of Clinical Research. 3rd ed. Maryland Heights, MO: Elsevier, Inc; 2012:371-381.
50. Ursum J, Bos WH, van de Stadt RJ, Dijkmans BA, van Schaardenburg D. Different properties of ACPA and IgM-RF derived from a large dataset: further evidence of two distinct autoantibody systems. Arthritis Res Ther 2009;11(3):R75 [FREE Full text] [doi: 10.1186/ar2704] [Medline: 19460147]
51. Ajdacic-Gross V, Vetter S, Müller M, Kawohl W, Frey F, Lupi G, et al. Risk factors for stuttering: a secondary analysis of a large data base. Eur Arch Psychiatry Clin Neurosci 2010;260(4):279-286. [doi: 10.1007/s00406-009-0075-4] [Medline: 19826856]
52. Monnelly EP, Locastro JS, Gagnon D, Young M, Fiore LD. Quetiapine versus trazodone in reducing rehospitalization for alcohol dependence: a large data-base study. J Addict Med 2008;2(3):128-134. [doi: 10.1097/ADM.0b013e318165cb56] [Medline: 21768982]
53. From AM, Bartholmai BJ, Williams AW, Cha SS, Pflueger A, McDonald FS. Sodium bicarbonate is associated with an increased incidence of contrast nephropathy: a retrospective cohort study of 7977 patients at mayo clinic. Clin J Am Soc Nephrol 2008;3(1):10-18 [FREE Full text] [doi: 10.2215/CJN.03100707] [Medline: 18057306]
54. Habash-Bseiso DE, Rokey R, Berger CJ, Weier AW, Chyou PH. Accuracy of noninvasive ejection fraction measurement in a large community-based clinic. Clin Med Res 2005;3(2):75-82 [FREE Full text] [Medline: 16012124]
55. Zhang Y, Hemond MS. Uncovering the predictive value of minimum blood glucose through statistical analysis of a large clinical dataset. AMIA Annu Symp Proc 2009;2009:725-729 [FREE Full text] [Medline: 20351948]

56. Morgan CL, Evans M, Toft AD, Jenkins-Jones S, Poole CD, Currie CJ. Clinical effectiveness of biphasic insulin aspart 30:70 versus biphasic human insulin 30 in UK general clinical practice: a retrospective database study. Clin Ther 2011;33(1):27-35. [doi: 10.1016/j.clinthera.2011.01.023] [Medline: 21397771]
57. Kobayashi H, Miyata H, Gotoh M, Baba H, Kimura W, Kitagawa Y, et al. Risk model for right hemicolectomy based on 19,070 Japanese patients in the National Clinical Database. J Gastroenterol 2013 Jul 27. [doi: 10.1007/s00535-013-0860-8] [Medline: 23892987]
58. International Warfarin Pharmacogenetics Consortium, Klein TE, Altman RB, Eriksson N, Gage BF, Kimmel SE, et al. Estimation of the warfarin dose with clinical and pharmacogenetic data. N Engl J Med 2009;360(8):753-764 [FREE Full text] [doi: 10.1056/NEJMoa0809329] [Medline: 19228618]
59. Moreno-De-Luca D, Sanders SJ, Willsey AJ, Mulle JG, Lowe JK, Geschwind DH, et al. Using large clinical data sets to infer pathogenicity for rare copy number variants in autism cohorts. Mol Psychiatry 2013;18(10):1090-1095 [FREE Full text] [doi: 10.1038/mp.2012.138] [Medline: 23044707]
60. Koefoed P, Andreassen OA, Bennike B, Dam H, Djurovic S, Hansen T, et al. Combinations of SNPs related to signal transduction in bipolar disorder. PLoS One 2011;6(8):e23812 [FREE Full text] [doi: 10.1371/journal.pone.0023812] [Medline: 21897858]
61. Hill B, Proulx J, Zeng-Treitler Q. Exploring the use of large clinical data to inform patients for shared decision making. Stud Health Technol Inform 2013;192:851-855. [Medline: 23920678]
62. Li H, Giger ML, Yuan Y, Chen W, Horsch K, Lan L, et al. Evaluation of computer-aided diagnosis on a large clinical full-field digital mammographic dataset. Acad Radiol 2008;15(11):1437-1445 [FREE Full text] [doi: 10.1016/j.acra.2008.05.004] [Medline: 18995194]
63. Li H, Giger ML, Lan L, Bancroft Brown J, MacMahon A, Mussman M, et al. Computerized analysis of mammographic parenchymal patterns on a large clinical dataset of full-field digital mammograms: robustness study with two high-risk datasets. J Digit Imaging 2012;25(5):591-598 [FREE Full text] [doi: 10.1007/s10278-012-9452-z] [Medline: 22246204]
64. Lin KW, Tharp M, Conway M, Hsieh A, Ross M, Kim J, et al. Feasibility of using Clinical Element Models (CEM) to standardize phenotype variables in the database of genotypes and phenotypes (dbGaP). PLoS One 2013;8(9):e76384 [FREE Full text] [doi: 10.1371/journal.pone.0076384] [Medline: 24058713]
65. Stephen R, Boxwala A, Gertman P. Feasibility of using a large Clinical Data Warehouse to automate the selection of diagnostic cohorts. AMIA Annu Symp Proc 2003:1019 [FREE Full text] [Medline: 14728522]
66. Roelofs E, Persoon L, Nijsten S, Wiessler W, Dekker A, Lambin P. Benefits of a clinical data warehouse with data mining tools to collect data for a radiotherapy trial. Radiother Oncol 2013;108(1):174-179. [doi: 10.1016/j.radonc.2012.09.019] [Medline: 23394741]
67. Stenson PD, Mort M, Ball EV, Shaw K, Phillips AD, Cooper DN. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet 2013 Sep 28. [doi: 10.1007/s00439-013-1358-4] [Medline: 24077912]
68. Hu H, Brzeski H, Hutchins J, Ramaraj M, Qu L, Xiong R, et al. Biomedical informatics: development of a comprehensive data warehouse for clinical and genomic breast cancer research. Pharmacogenomics 2004;5(7):933-941. [doi: 10.1517/14622416.5.7.933] [Medline: 15469411]
69. Savas S. A curated database of genetic markers from the angiogenesis/VEGF pathway and their relation to clinical outcome in human cancers. Acta Oncol 2012;51(2):243-246. [doi: 10.3109/0284186X.2011.636758] [Medline: 22150118]
70. Singh SK, Malik A, Firoz A, Jha V. CDKD: a clinical database of kidney diseases. BMC Nephrol 2012;13:23 [FREE Full text] [doi: 10.1186/1471-2369-13-23] [Medline: 22540288]
71. Suzuki H, Gotoh M, Sugihara K, Kitagawa Y, Kimura W, Kondo S, et al. Nationwide survey and establishment of a clinical database for gastrointestinal surgery in Japan: Targeting integration of a cancer registration system and improving the outcome of cancer treatment. Cancer Sci 2011;102(1):226-230. [doi: 10.1111/j.1349-7006.2010.01749.x] [Medline: 20961361]
72. Chawla NV, Davis DA. Bringing big data to personalized healthcare: a patient-centered framework. J Gen Intern Med 2013;28(suppl 3):S660-S665. [doi: 10.1007/s11606-013-2455-8] [Medline: 23797912]
73. Han Y, Itälä T, Hämäläinen M. Citizen Centric Architecture approach - taking e-health forward by integrating citizens and service providers. Stud Health Technol Inform 2010;160(Pt 2):907-911. [Medline: 20841816]
74. Boussadi A, Caruba T, Zapletal E, Sabatier B, Durieux P, Degoulet P. A clinical data warehouse-based process for refining medication orders alerts. J Am Med Inform Assoc 2012;19(5):782-785 [FREE Full text] [doi: 10.1136/amiajnl-2012-000850] [Medline: 22523345]
75. Hernandez P, Podchiyska T, Weber S, Ferris T, Lowe H. AMIA Annu Symp Proc. 2009. Automated mapping of pharmacy orders from two electronic health record systems to RxNorm within the STRIDE clinical data warehouse URL: http://europepmc.org/abstract/MED/20351858/reload=0;jsessionid=4aLVRuWM5ugKTsb134ug.0 [accessed 2014-01-09] [WebCite Cache ID 6MVbvtUxj]
76. Cuggia M, Garcelon N, Campillo-Gimenez B, Bernicot T, Laurent JF, Garin E, et al. Roogle: an information retrieval engine for clinical data warehouse. Stud Health Technol Inform 2011;169:584-588. [Medline: 21893816]

77. Zhou X, Chen S, Liu B, Zhang R, Wang Y, Li P, et al. Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support. Artif Intell Med 2010;48(2-3):139-152. [doi: 10.1016/j.artmed.2009.07.012] [Medline: 20122820]
78. Zapletal E, Rodon N, Grabar N, Degoulet P. Methodology of integration of a clinical data warehouse with a clinical information system: the HEGP case. Stud Health Technol Inform 2010;160(Pt 1):193-197. [Medline: 20841676]
79. Hanss S, Schaaf T, Wetzel T, Hahn C, Schrader T, Tolxdorff T. Integration of decentralized clinical data in a data warehouse: a service-oriented design and realization. Methods Inf Med 2009;48(5):414-418. [doi: 10.3414/ME9240] [Medline: 19657544]
80. Evans RS, Lloyd JF, Pierce LA. AMIA Annu Symp Proc. 2012. Clinical use of an enterprise data warehouse URL: http://europepmc.org/abstract/MED/23304288/reload=0;jsessionid=P5NVyPK8BYs9OV0Ra8hg.0 [accessed 2014-01-09] [WebCite Cache ID 6MVc6BqLh]
81. Harrison JH. Introduction to the mining of clinical data. Clin Lab Med 2008;28(1):1-7. [doi: 10.1016/j.cll.2007.10.001] [Medline: 18194715]
82. Phan JH, Quo CF, Cheng C, Wang MD. Multiscale integration of -omic, imaging, and clinical data in biomedical informatics. IEEE Rev Biomed Eng 2012;5:74-87. [doi: 10.1109/RBME.2012.2212427] [Medline: 23231990]
83. Sinha A, Hripcsak G, Markatou M. Large datasets in biomedicine: a discussion of salient analytic issues. J Am Med Inform Assoc 2009;16(6):759-767 [FREE Full text] [doi: 10.1197/jamia.M2780] [Medline: 19717808]
84. Drai D, Grodzinsky Y. A new empirical angle on the variability debate: quantitative neurosyntactic analyses of a large data set from Broca's aphasia. Brain Lang 2006;96(2):117-128. [doi: 10.1016/j.bandl.2004.10.016] [Medline: 16115671]
85. Bellman R, Kalaba R. DYNAMIC PROGRAMMING AND STATISTICAL COMMUNICATION THEORY. Proc Natl Acad Sci U S A 1957;43(8):749-751 [FREE Full text] [Medline: 16590080]
86. Borthakur D. The Hadoop Distributed File System: Architecture and Design. 2007 URL: https://hadoop.apache.org/docs/r0.18.0/hdfs_design.pdf [accessed 2013-11-25] [WebCite Cache ID 6LOoNyUXR]
87. Chang F, Dean J, Ghemawat S. Bigtable: A Distributed Storage System for Structured Data. 2006 Presented at: the 7th USENIX Symposium on Operating Systems Design and Implementation (OSDI'06); 2006; Seattle, WA, US p. 205-218.
88. Litt S. NoSQL: The Unix Database (With awk). 2007 URL: http://www.troubleshooters.com/lpm/200704/200704.htm [accessed 2013-11-25] [WebCite Cache ID 6LOp7PTH6]
89. van der Burgt YE, Taban IM, Konijnenburg M, Biskup M, Duursma MC, Heeren RM, et al. Parallel processing of large datasets from NanoLC-FTICR-MS measurements. J Am Soc Mass Spectrom 2007;18(1):152-161. [doi: 10.1016/j.jasms.2006.09.005] [Medline: 17055738]
90. Stata Corporation. Stata reference manual : release 6. In: Stata Reference Manual Set, 4vol: Release 6. College Station, TX: Stata Corp; 1999.
91. R Development Core Team. official website for R. Vienna, Austria The R Project for Statistical Computing URL: http://www.r-project.org/ [accessed 2014-01-08] [WebCite Cache ID 6MU3hBSqT]
92. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003;19(2):185-193 [FREE Full text] [Medline: 12538238]
93. Pleil JD, Stiegel MA, Madden MC, Sobus JR. Heat map visualization of complex environmental and biomarker measurements. Chemosphere 2011;84(5):716-723. [doi: 10.1016/j.chemosphere.2011.03.017] [Medline: 21492901]
94. Garcia-Betancur JC, Menendez MC, Del Portillo P, Garcia MJ. Alignment of multiple complete genomes suggests that gene rearrangements may contribute towards the speciation of Mycobacteria. Infect Genet Evol 2012;12(4):819-826. [doi: 10.1016/j.meegid.2011.09.024] [Medline: 22008279]
95. Aita T, Nishigaki K. A visualization of 3D proteome universe: mapping of a proteome ensemble into 3D space based on the protein-structure composition. Mol Phylogenet Evol 2011;61(2):484-494. [doi: 10.1016/j.ympev.2011.06.020] [Medline: 21762784]
96. Lenzerini M. Data Integration: A Theoretical Perspective. 2002 Presented at: Proceedings of the ACM Symposium on Principles of Database Systems (PODS); 2002; Roma, Italy p. 233-246. [doi: 10.1145/543613.543644]
97. Halevy A, Rajaraman A, Ordille J. Data integration: the teenage years. 2006 Presented at: VLDB '06 Proceedings of the 32nd international conference on Very large data bases; 2006; Seoul, Korea p. 9-16.
98. Haas LM, Lin ET, Roth MA. Data integration through database federation. IBM Syst. J 2002;41(4):578-596. [doi: 10.1147/sj.414.0578]
99. Shyu C, Ytreberg FM. Reducing the bias and uncertainty of free energy estimates by using regression to fit thermodynamic integration data. J Comput Chem 2009;30(14):2297-2304. [doi: 10.1002/jcc.21231] [Medline: 19266482]
100. Taylor W, Gladman D, Helliwell P, Marchesoni A, Mease P, Mielants H, CASPAR Study Group. Classification criteria for psoriatic arthritis: development of new criteria from a large international study. Arthritis Rheum 2006;54(8):2665-2673 [FREE Full text] [doi: 10.1002/art.21972] [Medline: 16871531]
101. Gschwind R, Robert Y. Analyse der zeitlichen Veränderungen der Papillen-Reflexion mittels Hauptkomponentenanalyse. Klin Monatsbl Augenheilkd 2008;190(04):249. [doi: 10.1055/s-2008-1050370]
102. Santafé G, Lozano JA, Larrañaga P. Bayesian model averaging of naive Bayes for clustering. IEEE Trans Syst Man Cybern B Cybern 2006;36(5):1149-1161. [Medline: 17036820]

103. Farré J, Cabrera JA, Romero J, Rubio JM. Therapeutic decision tree for patients with sustained ventricular tachyarrhythmias or aborted cardiac arrest: a critical review of the Antiarrhythmics Versus Implantable Defibrillator trial and the Canadian Implantable Defibrillator Study. Am J Cardiol 2000;86(9A):44K-51K. [Medline: 11084100]
104. Lisboa PJ. A review of evidence of health benefit from artificial neural networks in medical intervention. Neural Netw 2002;15(1):11-39. [Medline: 11958484]
105. Chen HF. In silico log P prediction for a large data set with support vector machines, radial basis neural networks and multiple linear regression. Chem Biol Drug Des 2009;74(2):142-147. [doi: 10.1111/j.1747-0285.2009.00840.x] [Medline: 19549084]
106. Aickin M, Gensler H. Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods. Am J Public Health 1996;86(5):726-728. [doi: 10.2105/Ajph.86.5.726]
107. Bender R, Lange S. Adjusting for multiple testing—when and how? Journal of Clinical Epidemiology 2001;54(4):343-349. [doi: 10.1016/S0895-4356(00)00314-0]
108. Broberg P. A comparative review of estimates of the proportion unchanged genes and the false discovery rate. BMC Bioinformatics 2005;6:199 [FREE Full text] [doi: 10.1186/1471-2105-6-199] [Medline: 16086831]
109. van der Laan MJ, Dudoit S, Pollard KS. Multiple testing. Part II. Step-down procedures for control of the family-wise error rate. Stat Appl Genet Mol Biol 2004;3. [doi: 10.2202/1544-6115.1041]
110. Sanders CM, Saltzstein SL, Schultzel MM, Nguyen DH, Stafford HS, Sadler GR. Understanding the limits of large datasets. J Cancer Educ 2012;27(4):664-669. [doi: 10.1007/s13187-012-0383-7] [Medline: 22729362]
111. Derman E. Models. In: Models.Behaving.Badly. Why Confusing Illusion with Reality Can Lead to Disaster, on Wall Street and in Life. New York, New York: Free Press; 2012.
112. Kobayashi T, Kishimoto M, Swearingen CJ, Filopoulos MT, Ohara Y, Tokuda Y, et al. Differences in clinical manifestations, treatment, and concordance rates with two major sets of criteria for Behçet's syndrome for patients in the US and Japan: data from a large, three-center cohort study. Mod Rheumatol 2013;23(3):547-553. [doi: 10.1007/s10165-012-0696-8] [Medline: 22752504]
113. Jacobs A. The pathologies of big data. Commun ACM 2009;52(8):36. [doi: 10.1145/1536616.1536632]

Abbreviations

BD2K: Big Data to Knowledge
CS: computer science
SNP: single-nucleotide polymorphism
ZB: zettabytes

Edited by G Eysenbach; submitted 27.08.13; peer-reviewed by L Toldo, J Gao; comments to author 27.10.13; revised version received 25.11.13; accepted 08.12.13; published 17.01.14

Please cite as:
Wang W, Krishnan E
Big Data and Clinicians: A Review on the State of the Science
JMIR Med Inform 2014;2(1):e1
URL: http://www.medinform.jmir.org/2014/1/e1/
doi:10.2196/medinform.2913
PMID:25600256