Kanda *Renal Replacement Therapy* (2015) 1:4
DOI 10.1186/s41100-015-0008-9

Renal Replacement Therapy

REVIEW Open Access

Use of big data in medicine

Eiichiro Kanda[1,2]

Abstract

As a result of the advances in computers, data collection has become easy. The huge amount of data now available through computers is called big data. The characteristics of big data are represented by three Vs: (1) volume, (2) variety, and (3) velocity. Volume, variety, and velocity refer to the amount of data, the range of data types and sources, and the speed and frequency of the acquisition of responses from users, respectively. The steps in the analysis of big data include (1) the review of data, (2) the construction of a model, and (3) the interpretation of results. Big data has mainly been analyzed by methods based on correlation analysis. The analytical method mainly based on the correlations found in big data is useful in predicting the prognosis from the analytical results. However, the factors useful for predicting the prognosis are not always useful in the treatment of patients. In medicine, the results of correlating events alone cannot be applied to clinical practice because the causal relationship between causes and the onset of diseases is important. To use big data in medicine, the results obtained from the analysis of big data should be validated by clinical studies.

Keywords: Big data, Medicine, Information technology, Google, Central processing unit, Hadoop, Computer, Clinical trial

Background

The term "big data" has frequently been heard recently. The term has been used in economics and information technology (IT) since the beginning of the 2000s. According to a report of the American economist Diebold published in 2000, big data is defined as "the explosion in the quantity (and sometimes, quality) of available and potentially relevant data, largely the result of recent and unprecedented advancements in data recording and storage technology" [1]. In addition, according to the 2012 White Paper on Information and Communications in Japan (Ministry of Internal Affairs and Communications), big data is defined as the data used to derive information useful for businesses [2]. From these definitions, the term big data not only implies a huge amount of data but also includes concepts related to its contents and utilization methods.

Review

Characteristics of big data

What amount of data can be called big data? There has been no clear definition of the lowest amount of data. According to International Business Machines (IBM) Corporation, 2.5 exabyte (EB) (2.5×10^{18} bytes) of data are created everyday by various transactions including sensors, online payments, and social networks [3]. Assuming that the capacity of a commercially available hard disk recorder is 1 terabyte (TB) (1.0×10^{12} bytes), 2.5 EB corresponds to 2.5 million hard disk recorders. The amount of data used in clinical studies is on the order of megabyte (MB) (1.0×10^{6} bytes) in EXCEL format. However, one unit of big data is on the order of petabyte (PB) (1.0×10^{15} bytes) or EB, which indicates that the amount of big data is exceedingly large.

Each piece of big data originates from various sources. Smartphone sales reached 470 million units and accounted for approximately 30 % of all mobile phones in 2011; the number of smartphones is expected to reach 1.2 billion units, approximately 60 % of the total number of mobile phones, in 2015 [2]. Many smartphone users use social networking services (SNSs) such as Facebook and Twitter, share photos with friends, and play online games. Such action histories are recorded and accumulated as data. The amount of such unstructured data that cannot be compiled in databases has been rapidly increasing. The development of mobile devices such as smartphones is inextricably related to the growth of big data.

Correspondence: tokyo.kyosai.kanda@gmail.com
[1]Department of Nephrology, Tokyo Kyosai Hospital, 2-3-8, Nakameguro, Meguro, Tokyo 153-8934, Japan
[2]Center for Life Science and Bioethics, Tokyo Medical and Dental University, Bunkyo, Tokyo, Japan

Images and documents obtained using smartphones are being stored in certain areas via the Internet. Such an Internet environment is called the cloud, in analogy to clouds in the sky. According to the National Institute of Standards and Technology, cloud computing is a model for enabling convenient on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management or interaction with service providers. With this cloud computing service, we can easily organize and manage data in an integrated fashion. Thus, the disordered information that has been managed by individuals can be easily aggregated to form big data.

The characteristics of big data are represented by three Vs: (1) volume, (2) variety, and (3) velocity (Fig. 1). Volume refers to the amount of data, variety refers to the range of data types and sources, and velocity refers to the speed and frequency of the acquisition of responses from users. In some cases, a fourth V, veracity, the accuracy of data, is included. These four factors are mutually related. For example, as the amount of data increases, the effect of a slight measurement error on the entire data becomes small, increasing the accuracy of the data. In addition, a rapid response plays an important role in improving the velocity. Google reported that the number of searches of particular key words can serve as an indicator of an influenza epidemic. Google Flu Trends can predict the state of influenza epidemics throughout the world almost in real time using Google search data [4]. Although there is no medical relationship between the key words used in searches and influenza, a relationship was found between them by big data analysis. The high speed of data accumulation or real-time information acquisition has thus been demonstrated to supplement information that is medically correct.

Use of big data

There are three indispensable factors when using big data: (1) databases where a huge amount of data can be stored, (2) skills in analysis including statistics and programming, and (3) the ability to apply the analytical results to create something new and valuable. In the IT industry, the advances in hardware are rapid, the capacity of hard disks and servers has become huge, and the analysis speed of central processing units (CPUs) has continuously increased.

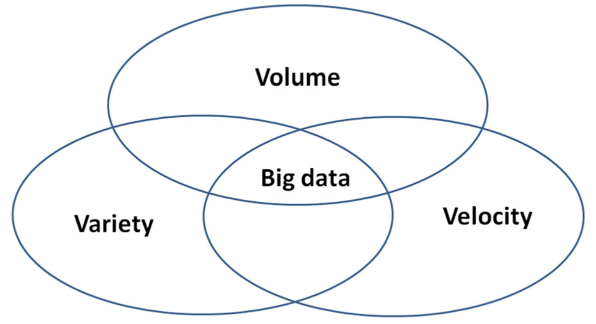

Fig. 1 Characteristics of big data: the three Vs

There are several types of database. A relational database is the most commonly used database and stores structured data. For example, medical data obtained through clinical studies is tabulated in EXCEL files. In the columns, variables such as ID, age, gender, and serum creatinine level are listed, whereas in the rows, the data for each patient are listed. Software such as EXCEL and ACCESS is used for relational databases. However, the amount of unstructured data, such as images and text on homepages, has been increasing. To handle these unstructured data, columnar databases, in-memory-type databases, and Hadoop have been developed. Conventional databases have been managed by a high-performance server, whereas Hadoop, consisting of several servers, realizes distributed processing and data are copied and stored on multiple servers (Fig. 2).

Analysis of big data

The steps in the analysis of big data include (1) the review of data, (2) the construction of a model, and (3) the interpretation of results. Big data does not necessarily consist of the required data in forms that are easy to analyze. Significant results cannot always be derived from a huge amount of data; rather, we must examine specific relationships that could be derived from the data. To do this, missing values and outliers that may mask the relationship should be properly processed and appropriate items should be selected.

Then, a model to represent the relationship among data is constructed. First, the required data are visualized to examine the trends of the entire data. Histograms and

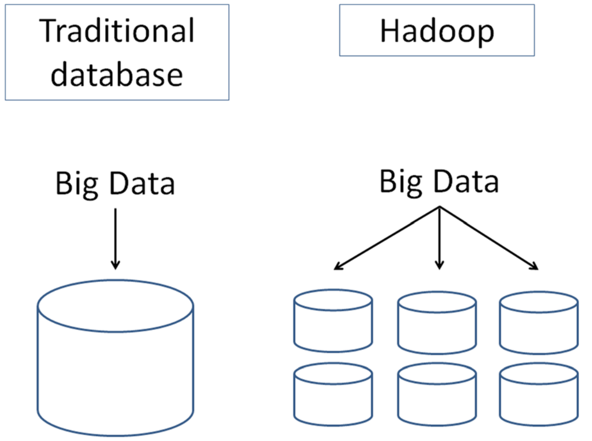

Fig. 2 Difference between Hadoop and conventional databases

scatter diagrams are effective tools for examining the change in item A of interest with time and frequency. Next, other items related to item A are searched for. Here, the linearity of a relationship between two items is examined using the correlation coefficient and by linear regression analysis. If the linearity of a relationship is evident, the analysis of the data will be easier and the conclusion will be clear. However, in practice, a linear relationship is not often found, and categorization and logarithms of the items are then examined. If a linear relationship is not found, nonlinear relationships are also examined. Cluster analysis is carried out to search for unknown groups. In addition, decision trees and support vector machines are sometimes used to classify items into groups. Appropriate analysis should be carried out as needed because data can change depending on the region of origin and changes in the environment.

Once a model is constructed, the volume of data required for statistical analysis is determined. Hypothesis testing cannot be appropriately applied to big data in some cases. For example, for a certain parameter θ, testing of the hypothesis that $\theta = 0$ is meaningful only when this hypothesis has scientific significance. However, in practice, θ is often nonzero. Therefore, the confidence interval of θ is obtained and whether or not 0 falls within the confidence interval is judged instead of hypothesis testing.

The obtained results are interpreted to be applicable to actual facts. For example, assuming that the mortality risk ratio of an exposure group to a control group is 1.5, we can conclude that the mortality risk increases 1.5-fold or by 50 % when a person undergoes the exposure of interest.

Significance of big data

In conventional statistics, because the number of samples was small, data characteristics have been inferred on the basis of the law of large numbers and the central limit theorem under the assumption of random sampling from a general population. In contrast, for big data, all data (N = all) are used as the sample. For N = all, the interpretation of data significantly changes. When the number of samples is small, the analysis should be carried out considering errors and the distribution used to model the data. Therefore, the data of average patients are analyzed and the data of patients outside the 95 % confidence interval are not analyzed. However, as the size of the data approaches all, the necessity of considering the errors decreases, enabling a detailed analysis for individual patients. In other words, personalized medicine is possible because all the data of each patient are stored in databases. The accuracy required for each item of big data differs from that of conventional data. For example, a one-order-of-magnitude difference in accuracy can be ignored when comparing populations of various countries. In research dealing with a small number of samples, the effect of outliers on the results is significant, but it is negligible in the case of big data.

Correlation plays a more significant role than a causal relationship in the analysis of big data [5]. On Internet shopping sites such as Amazon, a recommendation system is often adopted. In such a system, products are recommended to a user on the basis of the correlation among products in which the user had previously shown interest and those actually purchased, i.e., association analysis. The research using Google Flu Trends explained above also uses such correlation.

Application of big data analysis to medicine

Can analytical methods mainly based on the correlations found in big data be directly applied to medical research? As an example, we consider the analysis of big data in medicine by focusing on correlation alone. Patients with chronic kidney disease (CKD) are administered various drugs: an angiotensin receptor blocker (ARB) for hypertension, an antihyperuricemic for hyperuricemia, a lipid-lowering drug for dyslipidemia, a phosphorus-lowering drug for hyperphosphatemia, and an erythropoiesis-stimulating agent (ESA) for renal anemia. The result of analysis may reveal a strong correlation between the drugs and the symptoms. However, the existence of a correlation does not mean that we can recommend an ESA for those who are administered an ARB. Each patient has a different clinical condition; not only the correlations of superficial events but also the clinical conditions behind them should be taken into consideration in the treatment of patients.

Assume that a new antihypertensive drug A is developed. The blood pressure of patients decreases when they are administered drug A. However, we cannot conclude that blood pressure decreases owing to the efficacy of drug A alone because improvements in lifestyle, such as loss of weight, quitting smoking, regular exercise, and a low-salt diet, may affect the result. There exist confounders that affect both drug A and the effect of decreasing blood pressure. Therefore, the effect of confounders should be removed in analyzing the relationship between drug A and the effect of decreasing blood pressure. Hence, a randomized controlled trial (RCT) is carried out to compare the efficacy of drug A in a group receiving drug A (drug A group) relative to a group receiving a placebo (placebo group). In the RCT, the subjects are allocated to the drug A and placebo groups such that the background factors of the two groups are identical, enabling the comparison of the relationship between drug A and the effect of decreasing blood pressure without the effect of confounders. Thus, RCTs are carried out in the clinical trial of new drugs to remove the effect of confounders.

The analytical method mainly based on the correlations found in big data is useful in predicting the prognosis from the analytical results. However, the factors useful for predicting the prognosis are not always useful in the treatment of patients. The correlation between events should be carefully evaluated because erroneous judgment can lead to the death of patients, particularly in medicine. To apply big data to medicine, related events should be found first by data mining using a technique based on correlation and then the causal relationship among the events should be demonstrated considering pathophysiological backgrounds.

Steps toward application of big data to medicine

To meaningfully apply big data to medicine, the results of analysis should be appropriately interpreted. There are several steps toward achieving this end: (1) understanding the issues and data in medicine, (2) data processing and analysis by modeling, and (3) the application of the obtained results to clinical practice. In step (1), the issues currently in question are clarified by gaining an understanding of the medical background to establish the context and significance of the data. In addition, whether the issues are worth the time, effort, and financial cost of analysis is also judged. It is important to consider the balance between the significance and cost of analysis. In step (2), data are collected and processed. For data collection, data are extracted from several databases or a data collection program is used. After that, data are processed for use in analysis. This process is comparable to precooking. In data analysis, mathematical modeling is carried out to formulate events using data. Here, the accuracy with which the model simulates the events is evaluated statistically. Much more time is required for data processing than for data analysis. In step (3), the medical significance, applicability, and progressiveness of the obtained analytical results are evaluated. Here, medical knowledge is required. Medically significant results can be obtained when experts with different skills work as a team because different areas of expertise and skills are required for each step.

Conclusions

As a result of the advances in computers, big data has been increasingly utilized. With big data, the correlations among various events are clarified. Finding new correlations is expected to greatly contribute to the development of medicine. In medicine, correlations suggested by the analysis of big data should be validated through clinical study because a causal relationship is the basis of medical application. Medically significant results can be obtained from big data when experts with different skills, such as medicine and data analysis, work as a team.

Competing interests

There is no conflict of interest.

Received: 10 September 2015 Accepted: 12 October 2015
Published online: 15 December 2015

References

1. Diebold F. "Big data" dynamic factor models for macroeconomic measurement and forecasting. Seattle: Eighth World Congress of the Econometric Society; 2000.
2. Japan Moiaaco. White paper. Ministry of Internal Affairs and Comunications of Japan: Japan. 2012
3. Corporation IBM. IBM study: digital era transforming CMO's agenda, revealing gap in readiness. 2011.
4. 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.
5. Mayer-Schonberger V, Cukier K. Big data: a revolution that will transform how we live, work, and think: John Murray Publishers Ltd; 2013