Special Article

Kidney Res Clin Pract 36:3-11, 2017
pISSN: 2211-9132 • eISSN: 2211-9140
https://doi.org/10.23876/j.krcp.2017.36.1.3

KIDNEY RESEARCH AND CLINICAL PRACTICE

Medical big data: promise and challenges

Choong Ho Lee, Hyung-Jin Yoon

Department of Biomedical Engineering, Seoul National University College of Medicine, Seoul, Korea

The concept of big data, commonly characterized by volume, variety, velocity, and veracity, goes far beyond the data type and includes the aspects of data analysis, such as hypothesis-generating, rather than hypothesis-testing. Big data focuses on temporal stability of the association, rather than on causal relationship and underlying probability distribution assumptions are frequently not required. Medical big data as material to be analyzed has various features that are not only distinct from big data of other disciplines, but also distinct from traditional clinical epidemiology. Big data technology has many areas of application in healthcare, such as predictive modeling and clinical decision support, disease or safety surveillance, public health, and research. Big data analytics frequently exploits analytic methods developed in data mining, including classification, clustering, and regression. Medical big data analyses are complicated by many technical issues, such as missing values, curse of dimensionality, and bias control, and share the inherent limitations of observation study, namely the inability to test causality resulting from residual confounding and reverse causation. Recently, propensity score analysis and instrumental variable analysis have been introduced to overcome these limitations, and they have accomplished a great deal. Many challenges, such as the absence of evidence of practical benefits of big data, methodological issues including legal and ethical issues, and clinical integration and utility issues, must be overcome to realize the promise of medical big data as the fuel of a continuous learning healthcare system that will improve patient outcome and reduce waste in areas including nephrology.

Keywords: Big data, Epidemiology, Data mining, Healthcare, Statistics

Introduction

Recent rapid increase in the generation of digital data and rapid development of computational science enable us to extract new insights from massive data sets, known as big data, in various disciplines, including internet business and finance. In the healthcare area, discovering new actionable insights has not been as common, although several success stories have been published in media and academic journals. This delayed progress of big data technology in the healthcare sector is a little bit odd, considering an earlier prediction that the application of big data technology was inevitable and that the healthcare sector would be one of the sectors expected to be benefited the most from big data technology [1].

The increasing gap between healthcare costs and outcomes is one of the most important issues, and many efforts to fill this gap are under way in many developed countries. The gap between healthcare costs and outcomes was analyzed to be the result of poor management of insights from research, poor usage of available evidence, and poor capture of care experience, all of which led to missed opportunities, wasted resources, and potential harm to patients. It has been suggested the gap could be overcome by the development of a "continuous learning healthcare system (Fig. 1)" in which a virtuous cycle is formed between the research and operational arms of healthcare, and data could be used effectively [2].

Received August 8, 2016; **Accepted** September 26, 2016

Correspondence: Hyung-Jin Yoon
Department of Biomedical Engineering, Seoul National University College of Medicine, 103 Daehak-ro, Jongno-gu, Seoul 03080, Korea.
E-mail: hjyoon@snu.ac.kr

$$\text{Value} = \frac{\text{Patient outcome}}{\text{Healthcare cost}}$$

Figure 1. A continuous learning healthcare system.

Therefore, a pressing need to improve healthcare quality and patient outcomes, increasing data availability, and increasing analytic capabilities are the drivers of the big data era of healthcare [2]. There are many challenges to overcome before big data technology can significantly improve healthcare costs, quality and outcomes.

In this review, we discuss what is big data, what is special about medical big data, what is medical big data for, how medical big data can be analyzed, and what are the challenges for medical big data.

What is big data?

Big data are data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it [3]. As the size of data increases above a critical point, quantitative issues of data are transformed into qualitative issues in the capture, processing, storage, analysis, and visualization of data. Although big data are frequently characterized as the 4 Vs—volume, velocity, variety, and veracity [3], the definition of big data is beyond the scope of the characteristics of data type, such as size or volume. The potential to represent the real world almost without bias, to be linked with other datasets, to be useful and reused, to accumulate value over time, and to innovate a multi-dimensional, systems-level understanding should be considered alongside the 4 Vs of data [4,5]. Although big data have huge datasets, the information they provide may be unsatisfactory for what a particular researcher has in mind, and value creation, which cannot be expected with individual datasets, can be achieved through the potential of linking with other datasets [5].

What is special about medical big data?

The complexity of healthcare results from the diversity of health-related ailments and their co-morbidities; the heterogeneity of treatments and outcomes; and the subtle intricacies of study designs, analytical methods and approaches for collecting, processing, and interpreting healthcare data [6]. There are various sources of medical big data, such as administrative claim record, clinical registries, electronic health records, biometric data, patient-reported data, the internet, medical imaging, biomarker data, prospective cohort studies, and large clinical trials [2,7]. Integration of these data sources causes complementary dimensions of data such as large size (smaller than big data from other disciplines, but larger than data of clinical epidemiology), disparate sources, multiple scales (seconds to years), incongruences, incompleteness, and complexity. There is no universal protocol to model, compare, or benchmark the performance of various data analysis strategies [6]. Tanaka et al [8] summarized the characteristics of medical big data compared to traditional clinical epidemiological data and according to the data holder.

Medical big data have several distinctive features that are different from big data from other disciplines. Medical big data are frequently hard to access and most investigators in the medical arena are hesitant to practice open data science for reasons such as the risk of data misuse by other parties and lack of data-sharing incentives [4]. Medical big data are often collected based on protocols (i.e., fixed forms) and are relatively structured, partially due to the extraction process that simplify raw data [9]. Another important feature is that medicine is practiced in a safety critical context in which decision-making activities should be supported by explanations. Medical big data can be costly due to involvement of the personnel, use of expensive instrumentation, and the potential discomfort of the patients involved. Medical big data are relatively small compared to data from other disciplines, and may be collected from a non-reproducible situation. Medical big data can be further affected by several sources of uncertainty, such as measurement errors, missing data, or errors in coding the information buried in textual reports. Therefore, the role of the domain knowledge may be dominant in both analyzing the data and interpreting the results [10]. Other distinctive features of medical big data in analytic aspects includes the different types of patient characteristics, which sometimes may require weighting; the time structure, which may be an additional dimension; and treatment information, time point of treatment decision and change (i.e., time-dependent confounding) [11].

A big data project involves making sense out of all accumulated data on as many variables as possible due to increasing availability and decreasing expense of computing technology [12]. Iwashyna and Liu [13] pointed that there were four ways in which a project might be "big data": material, question, analytic method, and aspiration. Medical big data may include data from new sources as materials for analysis, such as the internet, social media, and so on. Medical big data can give answers to questions focusing on the usefulness of locally stable associations and correlations even in the absence of causal evidence, with analytic methods such as new, often nonlinear, tools for pattern recognition from computer science and other fields, in addition to the conventional statistical tools. Finally, big data technology has been increasingly viewed as the catalyst for a continuously learning health system allowing bidirectional flow between research and operations [13]. As previously discussed big data can be the fuel flowing in a continuously learning healthcare system [2].

Medical big data can be broadly classified into three common forms, such as large n and small p (n = sample numbers, p = parameter numbers); small n and large p; and large n and large p [5]. Data with large n and small p can be dealt with classical statistical methods. One example of this kind of data is administrative claim data. Because this kind of data tend to be incomplete, noisy, and inconsistent, data cleaning such as defining cases to be analyzed is not trivial and understanding the context of data collection is essential. Spurious association can be another problem. Discrimination between statistical

Table 1. Medical big data analysis vs. classical statistical analysis

	Medical big data analysis	Classical statistical analysis
Application	Hypothesis-generating	Hypothesis-testing
Questions of interest	Overcoming the limitation of locally or temporally stable association with continually updating the data and algorithm	Trying to prove causal relationships
Domain knowledge	More important in interpretation of the results	Important both in collection of data and interpretation of the results
Sources of data	Any kind of sources; frequently multiple sources	Carefully specified collection of data; usually single source
Data collection	Recording without the direct supervision of a human	Human-based measurement recording
Coverage of data to be analyzed	Substantial fraction of entire population	Small data samples from a specific population with some assumptions of their distribution
Data size	Frequently huge	Relatively small
Nature of data	Unstructured and structured	Mainly structured
Data quality	Rarely clean	Quality controlled
Research questions of data analysis	May be different from those of data collection	Same as those of data collection
Underlying assumption of the model	Frequently absent	Based on various underlying probability distribution function
Analytic tools	Frequently automated with data mining algorithm	Manually by expert with classical statistics
Main outputs of analysis	Prediction, models, patterns identified	Statistical score contrasted against random chance
Privacy & ethics	Concerns about privacy and ethical issues	Data collection according to the pre-approved protocol; informed consent from the participants

and scientific significance of domain expertise may be crucial. Second form is data with small *n* and large *p*. Microarray analysis datasets are typical examples and classical statistical tests may not be able to deal with this type of data efficiently. Curse of dimensionality and multiple testing issues are the main problems with this type of data. Last, some data has large *n* and large *p*, where the issues of the first and second types may be raised according to circumstances. Although medical big data analysis and clinical epidemiology share many features, there are several differences between these two, some of which are summarized in Table 1.

What is medical big data for?

It has been pointed out that the pressing need to improve healthcare quality and patient outcomes, increasing data availability, and increasing analytic capabilities are three drivers of the big data era in healthcare, and that the potential of big data analytics application is improving the values of healthcare by improving outcomes and reducing waste in resources [2]. The strength of big data is finding associations, not showing whether these associations have meaning, and finding a signal is only the first step [14]. Big data analytics are generally not focused on causal inference, but rather on correlation or on identifying patterns amid complex data [2].

The potential value of medical big data has been demonstrated in: 1) the delivery of personalized medicine; 2) the use of clinical decision support systems such as automated analysis of medical images and the mining of medical literature; 3) tailoring diagnostic and treatment decisions and educational messages to support desired patient behaviors using mobile devices; 4) big data-driven population health analyses revealing patterns that might have been missed if smaller batches of uniformly formatted data had been analyzed instead; and 5) fraud detection and prevention [15]. Diagnosis on the basis of high-resolution measurement such as microarray or next-generation sequencing, the monitoring of molecular characteristics in the course of treatment to use for prediction and treatment decisions, and continuous monitoring of individuals' health are among the potential uses of medical big data [11]. Rumsfeld et al [2] summarized eight areas of application of big data analytics to improve healthcare: 1) predictive modelling for risk and resource use; 2) population management; 3) drug and medical device safety surveillance; 4) disease and treatment heterogeneity; 5) precision medicine and clinical decision support; 6) quality of care and performance measurement; 7) public health; and 8) research applications. Predictive analytics using big data technology is technology that learns from experience (data) to predict the future behavior of individuals in order to drive better decisions, i.e., future insights, based on a full picture of associations, for example, across time or a wide geographic area, or observed in a substantial fraction of entire population [16,17]. Big data are necessary but not sufficient, and simply accumulating a large dataset is of no value if the data cannot be analyzed in a way that generates future insights that we can act upon [17]. The value of medical big data should be evaluated with this prospective.

How can medical big data be analyzed?

Big data analysis exploits various algorithms of data mining, which can be defined as the automatic extraction of useful, often previously unknown information from large databases or datasets using advanced search techniques and algorithms to discover patterns and correlations in large pre-existing databases [18]. The tasks of data mining can be summarized as description, finding human-interpretable patterns and associations, and prediction, foretelling some response of interest [10]. Clinical data mining can be defined as the application of data mining to a clinical problem [19].

The algorithms of data mining are categorized as supervised, unsupervised, and semi-supervised learning. Supervised learning means to predict a known output of target, using a training set that includes already classified data to draw inference or classify prospective, testing data. In unsupervised learning, there is no output to predict, so analyzers try to find naturally occurring patterns or grouping within unlabeled data. Semi-supervised learning means to balance performance and precision using small sets of labeled or annotated data and a much larger unlabeled data collection [6,20].

Analytic goals of medical big data are prediction, modeling, and inference; classification, clustering, and regression are common methods exploited in these contexts [5]. Classification is a kind of supervised learning and can be thought as predictive modeling in which the output vector or predicting variable is categorical. Classification means to construct a rule to assign objects to one of a pre-specified set of classes (predicting variable) based solely on a vector of measurements taken on these objects. Classification techniques include logistic regression, naïve Bayesian methods, decision trees, neural networks, Bayesian networks, and support vector machine. The classification performance can be evaluated by various performance metrics tested in a test set or an independent validation set. These techniques can be used to develop a decision support system assigning a diagnosis among several possible diagnoses or to build models to predict a prognosis based on data from analysis of many biomarkers. Clustering is unsupervised learning used to find groupings in the data through the use of distance metrics. Clustering techniques include k-means clustering, principle components-based clustering, and self-organizing maps. Clustering performance can be evaluated by its performance in a subsequent supervised learning task. Clustering is frequently used in microarray data analysis or phylogenetic analysis, and also can be used in redefining of disease according to pathophysiologic mechanisms providing more specific therapeutic options. Regression is supervised learning where output variable is continuous and is a statistical analysis tool that quantifies the relationship between a dependent variable and one or more independent variables to depict trends in the data. Linear regression is the most commonly used technique in this category. Examples of its applications include a longitudinal analysis of patients' data or decision support system [5,20].

Iavindrasana et al [19] summarized nine steps in the data mining process: 1) learning of the application domain, such as determining the relevant prior knowledge of the domain and the goal of the data mining application; 2) dataset selection; 3) data cleaning and preprocessing; 4) data reduction and projection; 5) matching of the objective defined in step 1 to a data mining method; 6) choice of the algorithm and search for data patterns; 7) pattern extraction; 8) evaluation and interpretation; and 9) use of the discovered knowledge. The issues in data cleaning and pre-processing step includes data type issues such as binary, nominal, ordinal or numerical; variable domination issues in case of numerical data; redundancies among several variables; temporality issues; missing value issues; and outlier issues. Data reduction and projection step include reducing the number of variables for computation efficiency and overcoming the curse of dimensionality. During pattern extraction, the dataset can be divided into training and testing sets and the model developed in the training set is then tested in the testing set. There are many methods to split the dataset, such as cross validation, stratified cross validation, leave-one-out, and bootstrapping. The most commonly used performance metrics for evaluation are accuracy, sensitivity, specificity, receiver operating characteristic curve, precision, recall, f-measure, number of positive predictions, and number of false positives [19]. During the pattern extraction step, various algorithms can be tried and the algorithm showing the best performance can be chosen (so called "bake-off"); but in medical domain, transparency (or understandability) is another critical issue other than performance to be considered, as well as performance.

Medical big data have several issues related to the data themselves which although not specific to big data, needed to be considered during analyses. The issue of multiple comparison will not be discussed in this review.

Missing value

Medical big data analytics deal with data collected for other purposes, such as patient care in the case of electronic medical records, and these data inherently have many variables with missing values.

Although the simplest and most overused way to handle missing values is to remove the cases with missing values, or complete-case analysis, it is valid only when missing values are assumed to be independent of both observed and unobserved data (see below). This assumption is not realistic in most situations. Therefore, complete-case analysis in these cases may bias the conclusion. Another major drawback of the complete-case analysis is that reducing the number of data points available for analysis generally is very inefficient [21].

Missingness may exhibit various relationships with data already observed or unobserved data. Missing data are classified into three types: 1) missing completely at random (MCAR), 2) missing at random (MAR), and 3) not missing at random (NMAR). MCAR is missingness of which probability does not depend on either observed or unobserved data. If data are MCAR, the probability of a missing observation is the same for all entities. In these situations, complete-case analysis does not bias the scientific inference. This is rarely met in practice. MAR is missingness of which probability does not depend on unobserved data but depend on observed data. In these cases, the process of missingness should be adjusted for all the variables that affect the probability of missingness. NMAR is missingness of which probability depends on unobserved data. There are many tool kits to handle these types of missingness including NMAR, such as in SAS, R, Stata, and WinBUGS [21]. There is no unique way to analyze NMAR data, nor will there ever be a program that will work well for all NMAR datasets [21,22]. It has been reported that if fewer than 10% of values were missing, many of the commonly used methods would result in similar conclusions. If between 10% and 60% of values were missing, multiple imputation was recommended. If missingness for more than 60% of the values, no method was found to give satisfactory results [21]. More details on incomplete data are reviewed in Wong et al [21].

Curse of dimensionality

High dimensional data are data with too many attributes compared to the number of observational units. Microarray data or next generation sequencing data are typically high dimensional datasets. In high-dimension datasets, many numerical analyses, data sampling protocols, combinatorial inference, machine learning methods, and data managing processes are susceptible to the "curse of dimensionality" [6]. The term, "the curse of dimensionality," was coined by Richard Bellman in the 1950s to describe the difficulty of optimization in high dimensional datasets [5].

Sparsity, multicollinearity, model complexity, computational cost to fit model, and model overfitting are the issues accompanied by high dimensional datasets [5]. The space volume increases rapidly as data dimension increases; thus, the distance between data points increases accordingly. The stability of distance metrics is critical in statistical inference; therefore, this sparsity between data points affects most quantitative analyses, even for big data [6]. Multicollinearity is a phenomenon in which two or more predictor variables in a model, such as the multivariate regression model, are not independent. It violates the common regression technique assumption that requires the predictor variables to be independent of the error term (model residuals). Multicollinearity makes a model unreliable or underpowered. Although in traditional statistical analyses with standard datasets, multicollinearity is exceptional, it may be ubiquitous in big data analyses [6]. Model overfitting may cause the problem of generalizability. High dimensional data can be handled with dimension reduction [23] or feature selection [24]. It is important to recognize that reducing dimensionality or feature selection may cause loss of key mechanistic information. There is an overall tradeoff between a false positive rate and the benefit of identifying novel insights [25].

Bias control

Randomized controlled trials minimize bias and control confounding and are therefore considered the gold standard of design validity [26]. Every dataset, however, has limitations. Randomized controlled trials are frequently showing the lack of generalizability because randomized controlled trials generally are conducted under ideal conditions, among highly selected patients followed by highly qualified physicians. Randomization is not always possible due to practical or ethical reasons [26]. It is practically impossible to perform a randomized intervention for a novel biomarker without specific measures to control its *in vivo* levels in human. It also is frequently lengthy and costly to obtain an answer for its question. Randomized controlled trials often produce heterogeneous results and a single randomized trial cannot be expected to provide a gold-standard result that applies to all clinical studies [27]. Well-designed observational studies may be less prone to heterogeneous results than randomized controlled trials, possibly due to a broad representation of the population at risk and less opportunity for differences in the management of subjects among observational studies, which already are diverse with respect to disease severity, treatment protocols, and coexisting illnesses. In contrast, each randomized controlled trial may have a distinct group of patients according to its specific inclusion and exclusion criteria, and the experimental protocol for therapy may not be representative of clinical practice [27]. Clinical studies, both observational and interventional, frequently lack the ability to provide reliable answers to their research questions because of inadequate sample sizes. Underpowered studies are subject to multiple sources of bias, may not represent the larger population, and are regularly unable to detect differences between treatment groups. Most importantly, underpowered studies can, moreover, lead to incorrect conclusions [7]. Big data analyses on various data from administrative claim database or national registries can be used to overcome these limitations. Big data studies provide real-world healthcare information from a broader, population-based perspective. Administrative claim data have broad generalizability, large numbers of patient records, and less attrition than clinical trials; they are faster and less costly than primary data collection, and can often be linked with other datasets [7].

Big data analyses are basically observational studies, and thus share the limitations of observational studies in addition to the limitations inherent to the big data. Big data analytics that are based on observational data will be subject to the inherent limitations of such data [2]. Observational studies cannot test causality and should be considered hypothesis-generating. It is recommended that the results of observational studies should not influence clinical practice until these hypotheses are tested in adequately powered randomized controlled trials [28]. Although Benson and Hartz [29] found little evidence that estimates of treatment effects in observational studies reported after 1984 were either consistently larger than or qualitatively different from those obtained in randomized controlled trials, Tai et al [28] compared the results of both Nurses' Health Study publications and randomized controlled trials for breast cancer, ischemic heart disease, and osteoporosis, and reported that the concordant (the difference in effect size between studies 0.15 or less) rate was less than 25%. The effect size of observational studies is frequently inflated due to selection bias, confounding, and methodological weaknesses such as measurement error. In addition, large observational studies can produce implausibly precise estimates of effect size that are highly statistically significant but clinically unimportant [7,28]. To minimize the impediments to drawing valid inferences, specific scientific best practices should be adopted, such as generation of a priori hypotheses in a written protocol, detailed analytical plans noting specific methods and safeguards against bias, and transparent reporting with justification of any changes in plans. Potential clinically important effects should be defined a priori and the results discussed accordingly [7]. There are two analytic techniques to address the problem of confounding in observational studies; propensity score analysis and instrumental variable analysis [26]. Propensity score is the likelihood of a patient being assigned to an intervention on the basis of his or her pre-intervention characteristics, and propensity score analysis is performed by creating pseudo-randomization of all possible measured confounders using the propensity score. The limitation of propensity score analysis is that, even if the propensity score method is able to reduce bias due to all measured confounders, it fails to limit bias due to unmeasured or unknown confounders. Instrumental variable analysis is comparing patient groups according to an instrumental variable which is randomly distributed, rather than comparing patients with respect to the actual intervention received. A critical step in instrumental variable analysis is to find an appropriate instrument. An instrumental variable should meet three requirements: 1) to be associated with the intervention or exposure (relevancy assumption); 2) not to directly affect the outcome of interest, but to only indirectly affect the outcome through the intervention assignment (exclusion restriction); and 3) to be independent of confounders [26]. Theoretically, this technique aims to control for unmeasured or unknown confounders [26]. Recently, there have been an increasing number of Mendelian randomization studies, a variant of instrumental variable analysis, which uses genetic variants as instrumental variables to circumvent the issues of both unmeasured confounding and reverse causation in observational studies [30].

What are the challenges for medical big data?

Although the potential of big data analytics is promising, assessing the "state of science" and recognizing that, at present, the application of big data analytics is largely promissory is important [2]. Therefore, it is critical to delineate some of challenges for big data applications in healthcare. First, the evidence of practical benefits of big data analytics is scarce. Second, there are many methodological issues, such as data quality, data inconsistency and instability, limitations of observational studies, validation, analytical issues, and legal issues, some of which are discussed in previous sections. An effort to improve the data quality of electronic health records is necessary.

In the nephrology area, although chronic kidney disease is one of the hottest area of research, its codes are not assigned in many of administration claim databases; most cases of acute kidney injury not requiring dialysis therapy are not coded in claim databases. Therefore, these practices need to be corrected. Many of these technical issues are remained to be solved. Last, clinical integration and utility is an issue. Big data analytics need to be integrated into clinical practice to reap the substantial benefits, and clinical integration requires the validation of clinical utility of big data analytics. The issues of clinical integration and utility have been largely overlooked [2]. It is critical to solve these challenges to fasten the application of big data technology in medical sector and thus to improve patient outcome and to reduce waste of resources in healthcare, which should be the real value of big data studies.

Conflicts of interest

All authors have no conflicts of interest to declare.

Acknowledgments

This work was funded by the Korea Meteorological Administration Research and Development Program (grant number KMIPA 2015-5120).

References

[1] Murdoch TB, Detsky AS: The inevitable application of big data to health care. *JAMA* 309:1351-1352, 2013

[2] Rumsfeld JS, Joynt KE, Maddox TM: Big data analytics to improve cardiovascular care: promise and challenges. *Nat Rev Cardiol* 13:350-359, 2016

[3] Bellazzi R: Big data and biomedical informatics: a challenging opportunity. *Yearb Med Inform* 9:8-13, 2014

[4] Scruggs SB, Watson K, Su AI, Hermjakob H, Yates JR 3rd, Lindsey ML, Ping P: Harnessing the heart of big data. *Circ Res* 116:1115-1119, 2015

[5] Sinha A, Hripcsak G, Markatou M: Large datasets in biomedicine: a discussion of salient analytic issues. *J Am Med Inform Assoc* 16:759-767, 2009

[6] Dinov ID: Methodological challenges and analytic opportunities for modeling and interpreting Big Healthcare Data. *Gigascience* 5:12, 2016

[7] Slobogean GP, Giannoudis PV, Frihagen F, Forte ML, Morshed S, Bhandari M: Bigger data, bigger problems. *J Orthop Trauma* 29 Suppl 12:S43-S46, 2015

[8] Tanaka S, Tanaka S, Kawakami K: Methodological issues in observational studies and non-randomized controlled trials in oncology in the era of big data. *Jpn J Clin Oncol* 45:323-327, 2015

[9] Wang W, Krishnan E: Big data and clinicians: a review on the state of the science. *JMIR Med Inform* 2:e1, 2014

[10] Bellazzi R, Zupan B: Predictive data mining in clinical medicine: current issues and guidelines. *Int J Med Inform* 77:81-97, 2008

[11] Binder H, Blettner M: Big data in medical science--a biostatistical view. *Dtsch Arztebl Int* 112:137-142, 2015

[12] DeRouen TA: Promises and pitfalls in the use of “Big Data” for clinical research. *J Dent Res* 94(9 Suppl):107S-109S, 2015

[13] Iwashyna TJ, Liu V: What’s so different about big data?. A primer for clinicians trained to think epidemiologically. *Ann Am Thorac Soc* 11:1130-1135, 2014

[14] Khoury MJ, Ioannidis JP: Medicine. Big data meets public health. *Science* 346:1054-1055, 2014

[15] Roski J, Bo-Linn GW, Andrews TA: Creating value in health care through big data: opportunities and policy implications. *Health Aff (Millwood)* 33:1115-1122, 2014

[16] Meltzer AC, Pines JM: What big data can and cannot tell us about emergency department quality for urolithiasis. *Acad Emerg Med* 22:481-482, 2015

[17] Ketchersid T: Big data in nephrology: friend or foe? *Blood Purif* 36:160-164, 2013

[18] Lavecchia A: Machine-learning approaches in drug discovery: methods and applications. *Drug Discov Today* 20:318-331, 2015

[19] Iavindrasana J, Cohen G, Depeursinge A, Müller H, Meyer R, Geissbuhler A: Clinical data mining: a review. *Yearb Med Inform* 121-133, 2009

[20] Deo RC: Machine learning in medicine. *Circulation* 132:1920-1930, 2015

[21] Wong WK, Boscardin WJ, Postlethwaite AE, Furst DE: Handling missing data issues in clinical trials for rheumatic diseases. *Contemp Clin Trials* 32:1-9, 2011

[22] Dinov ID: Volume and value of big healthcare data. *J Med Stat Inform* 4: 3, 2016

[23] Li L: Dimension reduction for high-dimensional data. *Methods Mol Biol* 620:417-434, 2010

[24] Saeys Y, Inza I, Larrañaga P: A review of feature selection techniques in bioinformatics. *Bioinformatics* 23:2507-2517,

2007

[25] Alyass A, Turcotte M, Meyre D: From big data analysis to personalized medicine for all: challenges and opportunities. *BMC Med Genomics* 8:33, 2015

[26] Laborde-Castérot H, Agrinier N, Thilly N: Performing both propensity score and instrumental variable analyses in observational studies often leads to discrepant results: a systematic review. *J Clin Epidemiol* 68:1232-1240, 2015

[27] Concato J, Shah N, Horwitz RI: Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med* 342:1887-1892, 2000

[28] Tai V, Grey A, Bolland MJ: Results of observational studies: analysis of findings from the Nurses' Health Study. *PLoS One* 9:e110403, 2014

[29] Benson K, Hartz AJ: A comparison of observational studies and randomized, controlled trials. *N Engl J Med* 342:1878-1886, 2000

[30] Boef AG, Dekkers OM, le Cessie S: Mendelian randomization studies: a review of the approaches used and the quality of reporting. *Int J Epidemiol* 44:496-511, 2015