Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Chow and Kong, J Biom Biostat 2015, 6:3
DOI: 10.4172/2155-6180.1000236

Prespective Open Access

On Big-Data Analytics in Biomedical Research

Shein-Chung Chow[1,2] and Yuanyuan Kong[2]*

[1]Duke University School of Medicine, Durham, North Carolina, USA
[2]Beijing Friendship Hospital, Capital Medical University; National Clinical Research Center for Digestive Diseases, China

Abstract

In recent years, big data analytics has received much attention in the area of healthcare related biomedical research and development. Big data analytics enables research organizations to analyze a mix of structured and unstructured data for identifying valuable medical information and insights in healthcare related biomedical research and development.

Keywords: Big data analytics; Biomedical research; Unstructured data; Healthcare

Introduction

In healthcare related biomedical research, big data analytics is referred to as the analysis of large data sets which contain a variety of data sets (with similar or different data types) from various data structured, semi-structured or unstructured sources such as registry, randomized or non-randomized studies, published or unpublished studies, and health care databases. The purpose of big data analytics is to detect any possible hidden signals, patterns and/or trends of safety and efficacy of certain test treatments under study. In addition, it is to uncover any possible unknown associations and/or correlations between potential risk factors and clinical outcomes, and other useful biomedical information such as risk/benefit ratio of certain clinical endpoints/outcomes. The finding of big data analytics could lead to more efficient assessment of treatments under study and/or identification of new intervention opportunities, better disease management, other clinical benefits, and improvement of operational efficiency for planning of future biomedical studies.

As indicated in the request for proposal (RFP) at the website of the United States National Institutes of Health (NIH), biomedical research is rapidly becoming data-intensive as investigators are generating and using increasingly large, complex, multi-dimensional, and diverse data sets. However, the ability to release data, to locate, integrates, and analyzes data generated by others, and to utilize the data is often limited by the lack of tools, accessibility, and training. Thus, the NIH has developed the Big Data to Knowledge (BD2K) initiative to solicit development of software tools and statistical methods for data analysis in the four topic areas of data compression and reduction, data visualization, data provenance, and data wrangling as part of the overall BD2K initiative. Details can be found in http://bd2k.nih.gov/about_bd2k.html.

Raghupathi [1] also pointed out that the criteria for platform evaluation include availability, continuity, ease of use, scalability, ability to manipulate at different levels of granularity, privacy and security enablement, standardization of data with incompatible formats and quality assurance [2] typical advantages and limitations of open source platforms, [3] menu-driven, user-friendly and transparent of big data analytics, [4] real-time big data analytics as there is a lag between data collection and data processing, [5] the availability of numerous analytics algorithms, models and methods in a pull-down type of menu, [6] management of data ownership, governance and standards of continuous data acquisition and data cleansing.

In this article, in addition to the challenges outlined by Raghupathi [1] we will focus on some statistical issues regarding the quality, integrity, and validity of big data analytics in biomedical research. The issues include, but are not limited to, representativeness, quality, and integrity of big data, validity of big data analytics, FDA Part 11 compliance for electronic records, and statistical methodology and software development.

Challenging Statistical Issues

Representativeness of big data

In biomedical research, a big data often contains a variety of data sets (with data types) from various data sources including registry, randomized or non-randomized clinical studies, published or unpublished data, and health care databases. As a result, it is a concern whether the big data is a truly representative of the target patient population with the diseases under study because possible selection bias may have occurred when accepting individual data sets into the big data. In addition, heterogeneity is expected within and across individual data sets (studies). The issues of selection bias, heterogeneity, and consequently reproducibility and generalizability are briefly discussed below.

Selection bias: In practice, it is likely that most data sets with positive results will enter the big data, in which selection bias may have occurred. If we let μ and μ_b be the true means of the target patient population and big data, respectively. Also, let μ_P and μ_N be the true means of data sets with positive and negative results, respectively and r is the true proportion of data with positive results. In this case, $\mu=r\mu_{P+}(1-r)\mu_N$, where r is often unknown. Thus, selection bias for accepting individual data sets could have a significant impact on the finding of big data analytics. In other words, the assessment of μ through the big data analytics $\hat{\mu}_b$ could be biased because $bias(\hat{\mu}_b) = E(\hat{\mu}_b) - \mu = \mu_b - \mu = \in$. If the big data only contains data sets with positive results, then $\mu_{b=}\mu_P$. Consequently, the bias $\in = (1-r)(\mu_P - \mu_N)$, which could be substantial if μ_P is far away from μ_N. As a result, the findings of big data analytics could be biased and hence misleading due to selection bias.

Heterogeneity: In addition to the representativeness and selection

***Corresponding author:** Yuanyuan Kong, PhD, Clinical Epidemiology and EBM Unit, Beijing Friendship Hospital, Capital Medical University; National Clinical Research Center for Digestive Diseases, Beijing 100050, China. Tel: 86-10-63139163; E-mail: kongyuanyuan2000@163.com

Received June 16, 2015; **Accepted** June 23, 2015; **Published** June 30, 2015

Citation: Chow SC, Kong Y (2015) On Big-Data Analytics in Biomedical Research. J Biom Biostat 6: 236. doi:10.4172/2155-6180.1000236

bias of data sets, heterogeneity within and between individual data sets from different sources is also a great concern. In practice, although individual data sets may come from clinical studies conducted with the same patient population, data from these studies may be collected under similar but different study protocols with similar but different doses or dose regimens at different study sites with local laboratories. These differences will cause heterogeneity within and between individual data sets. In other words, these data sets may follow similar distributions with different means and different variances σ_i^2, where i=1,2,..,k (possible data sets in the big data), $\mu_i \neq \mu_j$ and $\sigma_i \neq \sigma_j$ for $i \neq j$ The heterogeneity could decrease the reliability of the assessment of the treatment effect.

Reproducibility and generalizability: As indicated above, the heterogeneity within and across individual data sets (studies) in the big data center could have an impact on the reliability of the assessment of the treatment effect. In addition, as the big data continues growing, it is a concern whether the findings from the big data analytics is reproducible and generalizable from one big data center (database) to another big data center (database) of similar patient population with the same diseases or conditions under study. For evaluation of reproducibility and generalizability, the concept using a sensitivity index proposed by Shao and Chow [7] is useful. Let (μ_0, σ_0) and (μ_1, σ_1) denote the population of the original database (big data center) and another database (another big data center). Thus, since the two databases are for similar patient populations with the same diseases and/or conditions, it is reasonable to assume that $\mu_1=\mu_0+\varepsilon$, and $\sigma_1 = C\sigma_0$, where ε and C are shift parameters in location and scale, respectively. After some algebra, it can be verified that

$$E_1 = \left|\frac{\mu_1}{\sigma_1}\right| = |\Delta|\left|\frac{\mu_0}{\sigma_0}\right| \text{, where } \Delta\left[1+\frac{\varepsilon}{\mu_0}\right]/\mathrm{C}$$

is the sensitivity index for generalizability. In other words, if $|\Delta| \leq 1-\delta$, where δ is a pre-specified small number, we then claim that the results from the original big data center are generalizable to another big data center with data obtained from similar patient population with the same diseases and/or conditions. In practice, since ε and C are random, statistical methodology for assessment of Δ is necessarily developed.

Data quality, integrity, and validity

In biomedical research, data management is an integral part of the clinical trial process, which ensures the quality, integrity and validity of data collected from trial subjects to a database system. A typical example for data management process is illustrated in Figure 1. Data management delivers a clean and high-quality database for statistical analysis and consequently enables clinical scientists to draw conclusions regarding the effectiveness, safety, and clinical benefit/risk of the test treatment under investigation. An invalid and/or poor quality database may result in wrong and/or misleading conclusions regarding the drug product under investigation. Thus, the objective of the data management process in clinical trials is not only to capture the information that the intended clinical trials are designed to capture, but also to ensure the quality, integrity and validity of the collected data. These data sets are then formed a big data through a database system. Since the big data center contains electronic data records from a variety of sources, some regulatory requirements must be met for assurance of data quality, integrity and validity of the electronic data in the big data center.

FDA Part 11 compliance: The FDA Part 11 compliance is referred to as requirements or criteria as described in 21 Codes of Federal

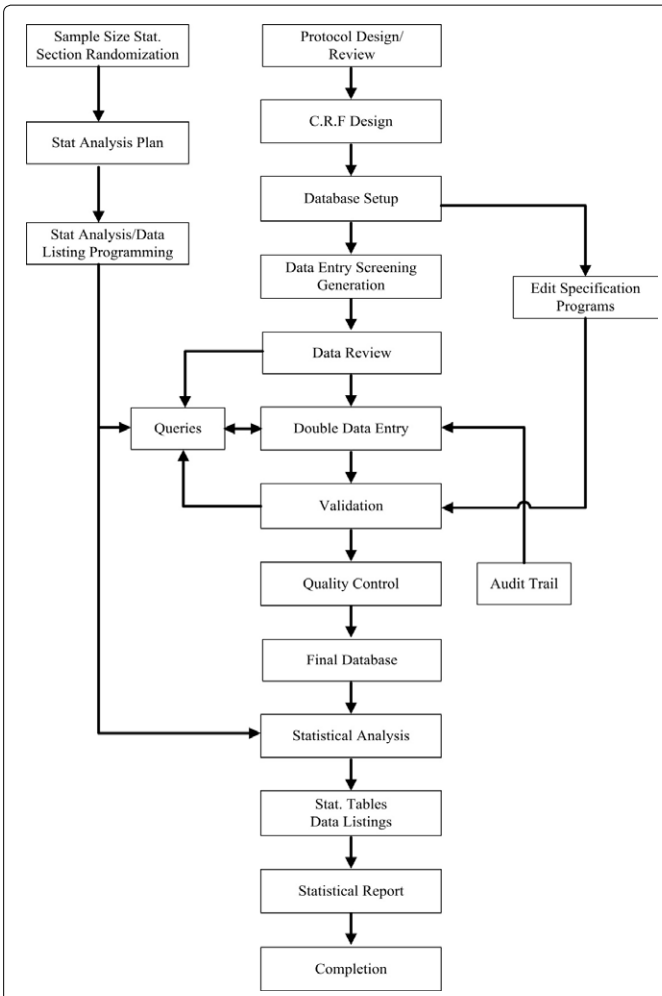

Figure 1: Typical example of data management process.

Registration (CFR) Part 11 under which the FDA will consider electronic records and signatures to be generally equivalent to paper records and handwritten signatures. It applies to any records required by the FDA or submitted to the FDA under agency regulations. To reinforce Part 11 compliance, FDA has published a compliance policy guide–CPG 7153.17, Enforcement Policy: 21 CFR Part 11 Electronic Records, Electronic Signatures. In addition, the FDA also published numerous draft guidance documents to assist the sponsors for Part 11 compliance. FDA Part 11 compliance has a significant impact on the process of clinical data management and consequently the big data management, which has recently become the focus for Good Data Management Practice (GDMP) in compliance with Good Statistics Practice (GSP) and Good Clinical Practice (GCP) for data quality, integrity, and validity. For example, 21 CFR Part 11 requires that procedures regarding creation, modification, maintenance, and transmission of records must be in place to ensure the authenticity and integrity of the records. In addition, the adopted systems must ensure that electronic records are accurately and reliably retained. 21 CFR Part 11 has specific requirements for audit trail systems to discern invalid or altered records. For electronic signatures, they must be linked to their respective electronic records to ensure that signatures cannot be transferred to falsify an electronic record. The FDA requires that systems must have the ability to generate documentation suitable for FDA inspection to verify that the requirements set forth by the 21 CFR Part 11 are met.

In practice, data management of big data is the top priority in the plan for 21 CFR Part 11 compliance for assurance of data quality, integrity and validity. A typical plan for Part 11 compliance for data management process usually includes (1) gap assessment, (2) user requirements specification, (3) validation master plan, and (4) tactical implementation plan. The task is implemented through a team consisting of senior experienced personnel from multiple disciplinary areas such as information technology (IT), programming, and data managers.

Missing Data – Missing values or incomplete data are commonly encountered in biomedical research and hence it has become a major issue for big data analytics. One of the primary causes of missing data is the dropouts. Reasons for dropouts include, but are limited to, refusal to continue in the study (e.g., withdrawal of informed consent), perceived lack of efficacy, relocation, adverse events, unpleasant study procedures, worsening of disease, unrelated disease, non-compliance with the study, need to use prohibited medication, and death [4]. How to handle the incomplete data is always a challenge to the statisticians in practice. Imputation is a very popular methodology to compensate for the missing data and is widely used in biomedical research. As compared to its popularity, however, its theoretical properties are far from well understood. As indicated by Soon [8], addressing missing data in clinical trials involves missing data prevention and missing data analysis. Missing data prevention is usually done through the enforcement of good clinical practices (GCP) during protocol development and clinical operations personnel training for data collection. This will lead to reduced biases, increased efficiency, less reliance on modeling assumption and less need for sensitivity analysis. However, in practice, missing data cannot be totally avoided. Missing data often occur due to factors beyond the control of patients, investigators, and clinical project team.

Statistical methodology and software development

Confounding factors: In big data analytics, there are many sources of variation that have an impact on the assessment of treatment effect relating to a certain new regimen or intervention. If some of these variations are not identified and properly controlled, they can become mixed with the treatment effect. In this case, the treatment effect is confounded by effects due to these variations. In biomedical research, there are many subtle, unrecognizable, and seemingly innocent confounding factors that can cause ruinous results of big data analytics. Moses [9] gave the example of the devastating result in the confounder being the personal choice of a patient. The example concerns a polio-vaccine trial that was conducted on two million children worldwide to investigate the effect of Salk poliomyelitis vaccine. This trial reported that the incidence rate of polio was lower in the children whose parents refused injection than those who received placebo after their parent gave permission [6]. After an exhaustive examination of the data, it was found that susceptibility to poliomyelitis was related to the differences between the families who gave the permission and those who did not.

Sometimes, confounding factors are inherent in the designs of individual studies in the big data. For example, dose titration studies in escalating levels are often used to investigate the dose-response relationship of the anti-hypertensive agents during the phase 2 stage of clinical development. For a typical dose titration study, after a washout period during which previous medication stops and the placebo is prescribed, N subjects start at the lowest dose for a pre-specified time interval. At the end of the interval, each patient is evaluated as a responder to the treatment or a non-responder according to some criteria pre-specified in the protocol. In a titration study, a subject will continue to receive the next higher dose if he or she fails, at the current level, to meet some objective physiological criteria such as reduction of diastolic blood pressure by a pre-specified amount and has not experienced any unacceptable adverse experience. Dose titration studies are quite popular among clinicians because they mimic real clinical practice in the care of patients [10]. The major problem with this typical design for a dose titration study is that the dose-response relationship is often confounded with time course and the unavoidable carryover effects from the previous dose levels which cannot be estimated and eliminated. Thus, in big data analytics, appropriate statistical methodology must be developed in order to address the issue of possible confounding factors for a valid assessment of the treatment effect under investigation.

Statistical methodology and software development: As indicated earlier, NIH has launched the bd2K initiative to focus on the following areas: data compression/reduction, data visualization, data provenance, and data wrangling, which require innovative analytical methods and software tools with the objective of addressing critical current and emerging needs of the biomedical research community for using, managing, and analyzing the larger and more complex data sets inherent to biomedical big data. Data compression is referred to as the algorithm-based conversion of large data sets into alternative representations that require less space in memory, while data reduction is the reduction of data volume via the systematic removal of unnecessary data bulk. Data visualization refers to human-centric data representation that aids information presentation, exploration, and manipulation. Data provenance, on the other hand, is referred to as the chronology or record of transfer, use, and alteration of data that document the reverse path from a particular set of data back to the initial creation of a source dataset. Finally, data wrangling is a term that is applied to activities that make data more usable by changing their form but not their meaning, which may involve reformatting data, mapping data from one data model to another, and/or converting data into more consumable forms.

Contraversial Issues

One of the most controversial issues in big data analytics occurs when the finding of the big data analytics (with a large scale) is inconsistent with that from a relatively small scale of adequate well-controlled randomized clinical trial which was conducted under the similar target patient population. In this case, the representativeness of the big data is questionable which may be due to the possible selection bias of accepting *poor* data sets into the big data. The inconsistency may indicate that there are major dissimilarities among individual data sets (studies) in the big data. Thus, it is suggested that similarities/dissimilarities, possible interactions, and poolability be carefully assessed for identifying the possible causes of inconsistencies.

The other controversial issue that is commonly seen is related to reproducibility of an established predictive model from big data analytics using similar but slightly different statistical methods. For example, in a case-control study utilizing the technique of propensity score matching with respect to some selected variables for matching, the use of (logistic) regression analysis with forward or backward stepwise approach often arrive similar but different predictive models with different sets of risk factors (predictors).

Another controversial issue in big data analytics is related to the possible time effect. In practice, it is likely that the findings from big data analytics at different time periods are different. This may be due to the availability of advanced technology, genetic changes in patient

population, and health care over time. As a result, it is suggested that these factors be taken into consideration for a more accurate and reliable assessment of treatment effect (or clinically meaningful difference) under study.

Concluding Remarks

As big data include data sets from a variety of sources including registries, randomized or non-randomized clinical studies, published or unpublished data, positive or negative clinical results (data), and healthcare database, heterogeneity within and across these data sets will have an impact on the assessment of treatment effects of interest. Big data analytics provides opportunities for uncovering hidden important medical information, determining possible associations or correlations between possible risk factors and clinical outcomes, predictive model building, validation, and generalization, critical information for planning of future studies. Statistical methodology and software development are necessary for achieving these ultimate goals. Although there are benefits for big data analytics, statistical issues regarding representativeness of the big data and its quality, integrity, and validity as described in this article must be addressed to ensure the success of the big data analytics.

References

1. Raghupathi W, Raghupathi V (2014) Big data analytics in healthcare promise and potential. Health Inf Sci Syst 2: 2-3.
2. Bollier D (2010) The promise and peril of big data. The Aspen Institute, Washington DC.
3. Chow SC, Liu JP (2013) Design and analysis of clinical trials (3rdedn.) John Wiley and Sons, New York.
4. DeSouza CM, Legedza TR, Sankoh AJ (2009) An overview of practical approaches for handling missing data in clinical trials. J Biopharm Stat 19: 1055-1073.
5. IHTT (2013) Transforming health care through big data strategies for leveraging big data in the health care industry.
6. Meier P (1972) The biggest public health experiment ever, the 1954 field trial of the salk poliomyeitis vaccine (3rdedn.) Wadsworth, Belmont.
7. Shao J, Chow SC (2002) Reproducibility probability in clinical trials. Stat Med 21: 1727-1742.
8. Soon G (2009) Editorial Missing data prevention and analysis. Journal of Biopharmaceutical Statistics 19: 941-944.
9. Moses LE (1985) Statistical concepts fundamental to investigations. N Engl J Med 312: 890-897.
10. Ohlhorst FJ (2012) Big data analytic turning big data into big money. John Wiley and Sons, New York.