Biostatistics and Biometrics
Open Access Journal
ISSN: 2573-2633

Opinion
Volume 1 Issue 3 - April 2017
DOI: 10.19080/BBOAJ.2017.01.555563

Biostat Biometrics Open Acc J
Copyright © All rights are reserved by Sergio Davalos

# Big Data has a Big Role in Biostatistics with Big Challenges and Big Expectations

Sergio Davalos*

*Center for Business Analytics, University of Washington, USA*

**Submission:** April 25, 2017; **Published:** May 04, 2017

*Corresponding author: Sergio Davalos, Center for Business Analytics, University of Washington, USA, Tel: +1-253-692-4658; Fax: +1-253-692-4523; Email: sergiod@uw.edu

## Abstract

Big data has now gone beyond a conceptual construct to a viable working approach for addressing and making use of huge volumes of data. The expectations of big data applications and outcomes in healthcare such as: a 20% decrease in patient mortality, better information regarding patient health and symptoms, reducing readmission, better point of care decision making, integration of smart devices and sensors with data bases, everyday genome sequencing, developing a treatment approach for cancer, and assessing the risk of readmission [1-3]. There are six key characteristics of big data that impact its use in healthcare: volume, velocity, variety, veracity, validity, and volatility. These impact the use of big data methods. There are several types of challenges in the use of big data in biostatistics: statistical verifiability, computational load, and technical learning curve.

Keywords: Big data; Challenges; Expectations; Statistics; Machine learning

## Introduction

Big data has now gone beyond a conceptual construct to a viable working approach for addressing and making use of huge volumes of data. Big data was first a term for massive or complex data sets where traditional computational methods are inadequate. It has now grown to include an approaches and technologies. As successful use of big data has grown, so have the expectations of its applications and outcomes in healthcare such as: a 20% decrease in patient mortality, better information regarding patient health and symptoms, reducing readmission, better point of care decision making, integration of smart devices and sensors with data bases, everyday genome sequencing, developing a treatment approach for cancer, and risk of readmission [1-3].

While the use of statistics in healthcare excels its use in other industries, the use of big data techniques has lagged in healthcare [3]. There are six key characteristics of big data that impact its use in healthcare. Volume is one. The massive amounts of data needs to be stored somewhere and to be effectively used, it needs to be accessible. Additionally, the data needs to be processed before it can be stored. Velocity is another. This refers to how quickly the data is coming which puts a burden on subsequent computing. This is particularly true when the data is captured in real time. Variety is a third. Big data can be textual or numeric and structured or unstructured. In addition to typical information such as patient statistics or epidemic data, big data can include geo spatial data, 3 D data, audio, video, blog files, and social media. A fourth characteristic and, perhaps, most key is veracity. This refers to the quality of data. Big data can be imprecise, noisy, uncertain, full of biases, and abnormalities. A fifth characteristic is validity. This refers to whether the data is correct and accurate for its intended use. This also goes toward the generalizability of the model produced when applied to different populations [4]. The sixth characteristic is volatility. This refers to how long the data is valid and how long it needs to be stored. As more understanding and knowledge is gained as well as conditions change, the data may need to be collected again or models regenerated. A clear example of this is concept drift where the statistical properties of the target variable, which the model is trying to predict, change over time in unforeseen ways [5].

There are several key challenges in using big data in general that apply to biostatistics. One key challenge is how to determine the validity of the model and the significance of the outcomes. When a statistical model is used, standard statistical measures can be used. When machine learning methodologies such as neural networks or genetic algorithms are used, this becomes problematic because there are no such tests. As a consequence, exogeneous assumptions in most statistical methods for Big

Data cannot be validated due to incidental endogeneity [6]. As a result, big data analysis relies heavily on data validation, confusion matrix, and AUC analysis. Another key challenge is the sheer volume of the data. Effective processing and analysis may require in memory processing. This can impose a limit on how much data can be processed based on available computer memory. For instance, a matrix may have 10,000,000 rows and 30,000 columns. This alone would require close to 300GBytes of memory. Carrying out matrix calculations with matrices of this size is not feasible with standard processing methods. One common use of matrix calculations is in singular value decomposition (SVD) to reduce the dimensionality of data [7]. A third challenge is the dimensionality of the big data. With the extensive methods for data capture, the number of dimensions can be very large. This can be challenging when applying variable selection methods. The number of combinations of variables can be computationally unfeasible. This may also result in sparse matrices and can be problematic for analysis techniques.

There are two other challenges that are of a technical nature. One is the what platform or software to use. The solutions vary from open source such as R, MysSql, and Knime to proprietary provides by IBM, Microsoft, and Provalis. In addition, legacy software such as SPSS and SAS has been extended to address big data. Since different researchers may work on different systems, this can lead to problems in sharing data, processing, and analysis results. A second challenge is the learning curve associated with adopting big data methods and the computational platform used. For example, Hadoop, a framework for the distributed processing of large data sets across clusters of computers, is built on simple programming models and requires several layers of computing systems to be put in place. Another example is the use of cloud computing. Amazon provides AWS for the large scale processing needed. Microsoft provides its own product, Azure. In both cases, it requires an investment of time and resources which later on may prove to be a dead end. In summary, use of big data in biostatistics has methodological and technical consequences. Rather than providing easy solutions, things are more complex.

## References

1. Ajit Kumar Roy (2016) Impact of Big Data Analytics on Healthcare and Society. J Biom Biostat 7: 300.
2. Zolfaghar K, Meadem N, Teredesai A, Roy SB, Chin SC, et al. (2013) Big data solutions for predicting risk-of-readmission for congestive heart failure patients. 2013 IEEE International Conference on pp. 64-71.
3. Murdoch TB, Detsky AS (2013) The inevitable application of big data to health care. JAMA 309(13): 1351-1352.
4. Suman K (2017) Characteristics of Biomarkers on Predictive Ability of Risk Models in Development and Validation Populations. Biostat Biometrics Open Acc J 1(1): 1-2.
5. Stiglic G, Kokol P (2011) Interpretability of sudden concept drift in medical informatics domain. In Data Mining Workshops (ICDMW), pp. 609-613.
6. Fan J, Han F, Liu H (2014) Challenges of big data analysis. Natl Sci Rev 1(2): 293-314.
7. Liu W, Park EK (2014) Big data as an e-health service. In Computing, Networking and Communications (ICNC), pp. 982-988.