Progress in Neurobiology 152 (2017) 200–212

Contents lists available at ScienceDirect

Progress in Neurobiology

journal homepage: www.elsevier.com/locate/pneurobio

Review article

Can data repositories help find effective treatments for complex diseases?

Gregory K. Farber

Office of Technology Development and Coordination, National Institute of Mental Health, National Institutes of Health, 6001 Executive Boulevard, Room 7162, Rockville, MD 20892-9640, USA

ARTICLE INFO

Article history:
Received 9 June 2015
Received in revised form 31 December 2015
Accepted 22 March 2016
Available online 24 March 2016

Keywords:
Data sharing
Data repositories
Data infrastructure
Imaging data
Common data elements
Consents

ABSTRACT

There are many challenges to developing treatments for complex diseases. This review explores the question of whether it is possible to imagine a data repository that would increase the pace of understanding complex diseases sufficiently well to facilitate the development of effective treatments. First, consideration is given to the amount of data that might be needed for such a data repository and whether the existing data storage infrastructure is enough. Several successful data repositories are then examined to see if they have common characteristics. An area of science where unsuccessful attempts to develop a data infrastructure is then described to see what lessons could be learned for a data repository devoted to complex disease. Then, a variety of issues related to sharing data are discussed. In some of these areas, it is reasonably clear how to move forward. In other areas, there are significant open questions that need to be addressed by all data repositories. Using that baseline information, the question of whether data archives can be effective in understanding a complex disease is explored. The major goal of such a data archive is likely to be identifying biomarkers that define sub-populations of the disease.

Published by Elsevier Ltd.

Contents

Abbreviations: CDE, common data element; CRCNS, collaborative research in computational neuroscience; fMRIDC, fMRI data center; IAN, Interactive Autism Network; NDA, NIMH Data Archive; NDAR, National Database for Autism Research; NDCT, National Database of Clinical Trials Related to Mental Illness; PDB, Protein Data Bank; PLM, Patients Like Me; RDoC db, Research Domain Criteria database.

E-mail address: farberg@mail.nih.gov (G.K. Farber).

http://dx.doi.org/10.1016/j.pneurobio.2016.03.008
0301-0082/Published by Elsevier Ltd.

1. Introduction

Over the past few years, big data has been touted as a way to advance many areas of biomedical science (Margolis et al., 2014). This review article will focus on trying to understand whether data repositories can help in the search for treatments for complex diseases. Complex diseases are defined as disorders that do not have a single deeply penetrant genetic cause or a single infectious agent. Generally, these diseases are thought to have multiple genetic and/or environmental contributions. Almost all of those diagnosed with a mental illness have a complex disease.

It is very common for complex diseases to be composed of multiple subpopulations. Each of these subgroups is defined by a unique set of underlying biological causes. However, the subgroups often share common symptoms. The symptoms allow a diagnosis but do not reflect the underlying biological causes and so do not allow us to understand which sub-population a patient belongs to. Type 1 and Type 2 diabetes are a good example. In type 1 diabetes, the body does not produce insulin. In type 2 diabetes, there is some problem with the way the body uses insulin. The biological causes of these two sub-categories of diabetes are quite different, yet those with either type of diabetes share the symptom of elevated blood sugar. The useful treatment options for those with type 1 diabetes are quite different from the treatment options for type 2 diabetes. In this case, testing for the presence of the C-peptide of insulin could provide an effective biomarker to differentiate those individuals who are producing insulin from those who are not (VanBuecken and Greenbaum, 2014). Useful biomarkers help differentiate the subgroups of a disease so that effective treatments can be discovered for each subpopulation.

Finding useful biomarkers for complex diseases is very difficult because of our limited understanding of the number of subpopulations for the disease as well as our limited understanding of the underlying genomic and environmental factors that have caused the disease. The purpose of this paper is to explore whether a data repository can contribute to the discovery of a biomarker that would be useful for identifying distinct subtypes of disorders. Data aggregation also raises questions about the best way to combine data from multiple laboratories and the amount of data that might be needed to uncover subpopulations.

2. Specialized infrastructures?

A data repository for complex diseases will need to deal with large amounts of heterogeneous data. This raises the question of whether specialized infrastructures will be needed to store the data. A number of recent reviews explore various aspects of what is big data (DeMauro et al., 2014; Jagadish et al., 2014). There is no doubt that the amount of data being collected in biomedical research laboratories is rapidly increasing. However, the scale of data collected by some physics experiments, by retailers, by social media providers, and by the government is often much larger in size or in the speed of acquisition than most current biomedical experiments (Schadt et al., 2010), and we already have informatics infrastructures that allow both the storage and analysis of those data. Large biomedical data sets have terabytes to petabytes of data while data sets in those other domains have petabytes or even exabytes (Leung, 2014). The data generated by a biomedical research laboratory can certainly tax the data storage resources in that lab or in the department or even the university, but the same data could be easily stored using the solutions that have been created for big data in other areas at relatively low cost.

Will biomedical data become big data? The change in pace of biomedical data acquisition is hard to measure, but it is likely that genomic data will be the driver for increased storage and computational needs in biomedicine. A recent perspective (Stephens et al., 2015) argues that the amount of sequencing data produced is doubling every seven months. This is roughly consistent with the growth of the sequence read archive (Kodama et al., 2012) seems to be increasing by an order of magnitude roughly every 31 months since January 2009.

The question of the growth of genomic data is important since if biomedical data is growing more quickly than the growth in data storage capacity, infrastructure investments will have to be made specifically to accommodate biomedical data. This genomic data will be only one sort of data that is needed for data mining to understand complex diseases. All of the relevant data will need to be made available in ways that are easily usable by the biomedical research community.

It has been estimated that the unit cost of storage capacity decreases by roughly an order of magnitude every 48 months (Komorowski, 2014). The increase in biomedical data storage is currently a little faster than, but in line with, the performance improvements for storage capacity. However, as Stephens et al. (2015) argue the increase in biomedical data may still be accelerating. For the near future it seems unlikely that biomedical researchers will need to worry about creating special data storage infrastructures or technologies beyond what has been created to deal with existing big data. However, if genomics experiments really begin to outpace the increases in generic storage capacity, specialized infrastructures will be necessary if all of the data are to be preserved. Stephens et al. (2015) correctly argue that the exascale data and computing centers that are in use today are the result of long range planning. The biomedical research community will need to assess this data growth carefully over the next few years and will need to find consensus to build appropriate data sharing and computational infrastructures to be used by the whole community for data mining and analysis of big data. The good news is that it is not necessary today to build a specialized informatics infrastructure to analyze data relevant to complex diseases.

While biomedical researchers might not have the largest datasets today, collectively, they probably have created the most diverse data sets. The heterogeneity in biomedical data arises from both the individual variability of subjects and samples as well as the diversity