BIOPRESERVATION AND BIOBANKING
Volume 00, Number 00, 2016

DOI: 10.1089/bio.2015.0106

COMMENTARY

Biobanking with Big Data: A Need for Developing "Big Data Metrics"

Kozlakidis Zisis

The term ''big data'' has often been used as an all-encompassing phrase for research that involves the use of large-scale data sets. However, the use of the term does little to signify the underlying complexity of definitions, of data sets, and of the requirements that need to be taken into consideration for sustainable research and the estimation of downstream impact. In particular, ''big data'' is frequently connected with biobanks and biobank networks as the institutions involved in tissue preservation are increasingly and perhaps unavoidably linked to the *de facto* preservation of information.

''Big data'' is commonly defined as collections of data sets so large and complex that its manipulation and management present significant logistical challenges (Oxford English Dictionary, 2013). Within the sphere of clinically oriented research, the term is often synonymous to electronic patient records from large hospitals, clinical trials, and/or -omics-based consortia and their associated banked samples. These data can be structured or unstructured, generated from diverse sources (sometimes in real time), and in very large volumes.

"Big Data" and Biobanking

Despite the much anticipated revolution in healthcare, the path from big data to clinical impact for specific research questions remains unclear. Clinically oriented biobanks are important sources for the provision of research-ready tissue as well as associated data under best practices, as an added-value proposition to their users. However, they do face a dual harmonization bottleneck of analytical laboratory and of digital techniques (data collection, curation, and storage). Both of these aspects can affect directly the biobank utilization rates and long-term sustainability.

The data contained within biobanks can be generated upon sample collection and/or through the retrospective linkage to medical/research records within their respective healthcare/academic institutions, resulting in a transfer and storage of the additional information. This connection is usually achieved through the deterministic linkage of samples and data from within clinical records using a unique identifier, such as hospital number, and a ''key'' that connects it to a banked sample in a secure manner. In a similar process, -omics data, for example, can be added to the accompanying sample information. However, the lack of reliable metrics to characterize the outcome of this data accrual can potentially confound the impact of such efforts.

Time for "Big Data Metrics"

The continuous collection of information for banked samples from different sources can create a heterogeneous information repository, especially as different subsets from a biobank's collections are used for different purposes—producing and sometimes returning different data. Indeed such a situation can occur at a high rate and the most practical approach might be that of characterizing the data according to biobank ''big data metrics.'' The European Nucleotide Archive (ENA) provides a useful example, where the experience of the last 30 years has resulted in three tiers of data, divided simply into unprocessed, processed, and interpreted information. The existence of data norms and recording of parameters relate to each one of these tiers.

In biobanking, three central questions can form the core of such metrics: (i) the type of information (raw/processed), (ii) the depth of detail (number of set parameters, structured/unstructured), and (iii) the completion status of the data (using perhaps the definitions of Accuracy, Timeliness, Comparability, Usability, and Relevance as adopted by the Canadian Institute for Health Information, 2009, and the European Statistical System, 2003). These three parameters are universally understood, provide a comparative measure, and could be potentially implemented as metrics or at least investigated by biobanks as part of the extant diverse range of Laboratory Information Management Systems.

The adoption of such descriptive ''big data'' metrics in biobanking could ease the difficulties of managing the expanding sets of deposited information alongside physically banked samples. In addition as biobanking networks continue to develop shared platform infrastructures (e.g., the EuPA Biobank Initiative), the characterization of the information already contained within their individual collections becomes an imperative prerequisite to a potential combination and subsequent reconciliation of data from different sources.

It Is the Tool, Not the Solution

''Big data metrics''—when developed—will become a tool for the management of biobanking growth, not necessarily a

Division of Infection and Immunity, University College London, London, United Kingdom.

solution. The speed with which biobanks have to deal with and assimilate technological change is considerable and frequently with significant associated costs. The addition of another tool, however useful, might also conceal additional costs and might need supporting evidence before any wide adoption considerations. These costs could be addressed by major research initiatives, such as the recent Big Data to Knowledge program established by the National Institutes of Health (2012), as well as the community-wide interactions fostered by organizations such as the International Society for Biological and Environmental Repositories, the European, Middle Eastern & African Society for Biopreservation & Biobanking, and others. The inclusion of relevant expertise from the pharmaceutical industry in handling large data sets, from their clinical trials, for example, is likely to be critical in the development, standardization, and adoption of big data metrics in biobanking.

The consistent aim of any such metrics development remains the more efficient characterization and management of increasing and different amounts of data being deposited in biobanks globally. As such, any ''big data metrics'' would need to adhere to the customary requirements for reproducibility, transferability, scalability, and perhaps flexibility—allowing some customization according to local needs. The outcome is hoped to be the ability of biobanks to carry on describing themselves well in the future and able to cope with future demands, in terms of the information they contain, and from a user perspective, the ability to locate the relevant information alongside appropriate banked samples for research purposes.

Address correspondence to:
Kozlakidis Zisis, PhD
Division of Infection and Immunity
University College London
Farr Institute of Health Informatics Research
222 Euston Road
London NW1 2DA
United Kingdom

E-mail: z.kozlakidis@ucl.ac.uk