British Journal of Clinical Pharmacology

Br J Clin Pharmacol (2016) **81** 804–806 804

EDITORIAL

Big Data: Challenges and opportunities for clinical pharmacology

Received 29 January 2016; **accepted** 29 January 2016

David Flockhart[1], Robert R. Bies[2], Marc R. Gastonguay[3] and Sorell L. Schwartz[4]

[1]Indiana School of Medicine, 1001 West 10th Street, Indianapolis, IN, [2]Department of Pharmaceutical Sciences, School of Pharmacy and Pharmaceutical Sciences, Member, Center for Data Enabled Science and Engineering, State University of New York at Buffalo, Buffalo, [3]Metrum Research Group LLC, 2 Tunxis Rd., Ste 112, Tariffville, CT 06081 and [4]Department of Pharmacology & Physiology, Georgetown University Medical Center, 3900 Reservoir Rd, NW, Washington, DC 20057, USA

> *An eminent physicist has remarked that the future truths of Physical Science are to be looked for in the sixth place of decimals.*
> Albert A. Michelson, 1894 [1]

> *Scientists are trained to recognize that correlation is not causation. … Petabytes allow us to say*: '*Correlation is enough*'.
> Chris Anderson, 2008 [2]

We intend, in this editorial, to highlight some key elements constituting *Big Data*, and to reflect on associated challenges and opportunities for clinical pharmacology. Two of the initial challenges are evidence of perceptual incongruity encumbering Big Data: (1) its definition and (2) the pretence of its emergence as a 21st century zeitgeist in science.

Its definitions vary. Our working description of choice is by the Oxford Dictionaries [3], to wit: *extremely large data sets that may be analyzed computationally to reveal patterns, trends and associations, especially relating to human behaviour and interactions*.

The 'zeitgeist' reference is to the school of thought that Big Data constitutes a philosophical revolution in science, but this seems a far-fetched embroidery of its capacity and, thus, its epistemological place in science. The 'eminent physicist' referred to by Michelson in the epigraph is thought to be Lord Kelvin and, perhaps the source of the quote misattributed to Kelvin: 'There is nothing new to be discovered in physics now. All that remains is more and more precise measurement'. There is no doubt that both statements reflected the view of physics at the close of the 19th century, that there was no longer a need for theory, just data analysis. Enter Einstein, Heisenberg *et al*.

A little over a century after the predicted demise of theory, the Anderson quote in the epigraph referred to Big Data in an article entitled *The End of Theory*: *The Data Deluge Makes the Scientific Method Obsolete*. Notwithstanding that the article appeared in a popular magazine, its impact was and still is significant, having been quoted, with both assent and dissent, in numerous scholarly publications and frequently touted in open lecture presentations [4]. Big Data is now commonly referred to as the Fourth Paradigm [5], wherein the first, second and third are empirical, theoretical and computational (simulation), respectively. Further elaboration of that grouping is not within the scope of this editorial except to say that, if the intended use of the term 'paradigm' is in the sense proffered by Kuhn [6], then the group is a strange admixture, indeed. We are in full agreement with Mazzochi [7]:

> 'The tendency to conflate the undisputed usefulness of Big Data – which above all, is an informational tool – with its presumed ability to provide full scientific understanding, sometime leads Big Data specialists to overstate their claims'.

The pitfall that awaits the expectation that a massive data deluge renders hypothesis and experimentation as remnants of an obsolete logic of scientific discovery is the expectation that Big Data are in some way a reflection of uniform data gathering. As noted by Leonelli [8]:

> 'Big Data that are made available through data bases for future analysis turns out to represent highly selected phenomena, materials and contributions, to the exclusion of the majority of biological work. What is worse, this selection is not the result of scientific choices, which can therefore be taken into account when analyzing the data. Rather, it is the serendipitous result of social, political, economic and technical factors, which determines which data get to travel in ways that are non-transparent and hard to reconstruct by biologists at the receiving end'.

This statement cautions that the vast sources of information can provide a great wealth of insight while simultaneously providing significant challenges to reality. Some of these concerns have been dismissed or considered as

DOI:10.1111/bcp.12896

minimized. Yet they persist as substantial hurdles that must considered when exploiting the Big Data information pool. These challenges include the supposition of an n = all information capture (presumably obviating the need for statistics), validly linking information across disparate sources, whether to implement a supervised *vs*. unsupervised statistical learning approach in analyzing data sources, how to map known physiological or pharmacological interactions as a scaffold for these, assumption of a static system, the assertion that all necessary empirical observations have been made and that the data collection is complete and the assertion that correlation is sufficient and that causation is no longer necessary for valuable information to be extracted from Big Data datasets.

In more traditional sample size experiments, challenges such as imbalanced experimental designs, missing data, lack of randomization or adequate controls, multiplicity of comparisons, etc. are difficult enough to manage. Increasing the size of the dataset does not necessarily eliminate these issues. In Big Data analyses, n = all is a condition where it is assumed that one has collected all of the data (or perhaps more apt, all of the information that exists in a particular area). Thus, sampling bias is of no concern because one has all possible samples. This is fanciful when measured against reality, and absurd as a default assumption. As a matter of concern, accepting this supposition raises the likelihood that sampling issues may actually be magnified in a Big Data analysis. David Spiegelhalter notes:

> 'There are a lot of small data problems that occur in Big Data…They don't disappear because you've got lots of the stuff. They get worse' [9]

Collection of all available data does not necessarily ensure that all possible data points have been captured. Important sampling bias may still exist. For example, scientists at Rutgers University considered the possible scenario where government agencies might make decisions about disaster relief based on social media data. Their analysis of Twitter usage during Hurricane Sandy led to the conclusion that the borough of Manhattan suffered the worst damage and that partying peaked after the storm subsided. Of course, this was the entirely inaccurate result of biased data collection. The Big Data dataset contained relatively few tweets from the most severely affected coastal areas because of power outages, failing cell phone batteries and a general trend toward lower social media usage in those communities compared with Manhattan [10] When linking information across disparate sources, consideration must be given to the type of data, the data collection process, the level of detail of the data and the time epoch when the data were collected. All of these factors can potentially confound the analysis.

The choice of a supervised *vs*. an unsupervised learning approach hinges on whether a target outcome that is desired to be predicted. A supervised learning approach requires having system output (target outcome) and then evaluation of possible predictors. These approaches range from LASSO, linear regression through to more complex approaches such as support vector machines and classification algorithms for the selection of optimal predictors [11]. Many of these more sophisticated supervised learning approaches are automated, leveraging the power of machine learning to facilitate the exploration of the system.

Unsupervised learning approaches, on the other hand, set out to evaluate the relationships among a set of variables in a dataset without attempting to predict anything in particular. Clustering algorithms or modularity analyses are those approaches that reveal interrelationships among variables in a dataset and can be useful in understanding how information groups together. However, these analyses are limited to correlational assessments and do not consider the underlying mechanistic constraints of the system.

Mapping physiological or pharmacological information as a scaffold can be a very powerful means of organizing information in datasets. In particular, it can be a framework related to cellular functioning or, perhaps, to a pharmacologic pathway where data can be applied in a structured sense to organize the information gleaned on the basis of what is known about the system.

These models also assume the system observed is static. By creating a large network of correlations, there is little information on mechanism or the fundamentals of the system that would allow for it to be utilized for prediction. This makes Big Data predictions and models vulnerable to changes in the system. One may develop a predictive model in a particular set of data that arose under very specific conditions. If those conditions change, or if the system changes dynamically, the model will no longer be able to predict outcomes adequately. This is what was observed with the Google Flu predictor where search terms were analyzed to determine and compared with flu outbreaks [12]. It was a robust predictor at the outset. Publicizing the predictor was followed by a subsequent two-fold over-prediction of influenza-like along with complete misses of flu outbreaks (including the non-seasonal A-H1N1 pandemic in 2009). Plausibly, knowledge of the predictor changed what individuals were searching for and therefore confused the classifier/predictive algorithm. Lazer *et al*. point out that 'quantity of data does not mean that one can ignore foundational issues of measurement and construct validity and reliability and dependencies among data' attributing this issue as part of the problem with the Google Flu predictor [12].

This is not to distract from the fact that Big Data usher in an information age that is and will continue to be will no doubt be transformative by the power it affords the scientific enterprise. For example:

Lang Li at the Center for Computational Biology and Bioinformatics, Indiana University School of Medicine focuses on identification of drug interactions in the context of personalized medicine. The first step in his work was to establish an ontology and corpus of information using text-mining techniques that identify likely drug–drug interactions by searching abstracts of publications in PUBMED. These hits were assessed for both the intensity of the interaction reported as well as the significance level. Once pruned, the most interesting interactions predicted using these Big Data text mining approaches were investigated further [13]. In the case of a loratadine and simvastatin interaction identified using this technique, evaluation of interaction at various CYP enzymes and transporters was undertaken. When these assessments did not result in a mechanistic explanation of the possible interaction, Han *et al*. went on to examine interactions between these compounds at the level of the muscle cell

[14]. What was identified was a pharmacodynamic interaction at the level of the muscle tubule that otherwise would not have been known. This work contributed to knowledge of mechanisms that may help to understand the high rates of myopathy observed with statin treatment.

Another example is the studies by Bernhard Palsson and coworkers at UCSD. They apply systems frameworks to -omics data sources. Their work has facilitated an understanding of cellular dynamics with such detail that they have predicted across scales using this approach 'from bacteria to human'. This was accomplished using non-linear optimization in conjunction with the -omics datasets that provide the constraints necessary to focus these large quantities of information into useful scaffolds that further increase the insight that can be discovered. Recent accomplishments include publications reporting 'systems biology guided identification of synthetic lethal gene pairs and its potential use to discover antibiotic combinations' [15] and the determination of a metabolic signature that facilitates the evaluation of quality of a unit of red blood cells [16]. The red blood cell work utilized an approach that combined endothelial markers and clinical outcomes to develop an algorithm that determines the red blood cell unit quality. These are significant challenges, the opportunity to identify novel antibiotic combinations that may have activity in resistant organisms and the means of evaluating the likely utility of a unit of red blood cells with respect to clinical outcomes, and highlight how Big Data can be leveraged to address critical public health problems.

Both of these examples illustrate the utility of Big Data techniques for generating hypotheses or identifying gaps in current scientific knowledge, but they also demonstrate the importance of going beyond signal detection and exploring potential explanatory and mechanistic underpinnings for their findings. These findings are often extended with specific experiments to refine findings and provide new insights that make important contributions to their respective fields.

Big Data provide enormous opportunities for clinical pharmacology as a nexus to help guide and provide frameworks that leverage this resource, but with full cognizance of limitations. Accordingly, while 'cheerleading' from the popular press is one thing, we find similarly conveyed acclamation from the scientific community to be disquieting. Even the 'BD2K' label the NIH attached to its data science initiative has this flavour. We feel compelled to caution: 'Be careful, there are impressionable students listening. Do you really want to introduce them to the end of theory?'

The enthusiasm for the remarkable instrumental utility Big Data is justified. Nonetheless, the logic of scientific discovery, manifest within hypothesis and theory, lives, waiting, perhaps, for another century when its demise is again predicted.

References

1 Michelson AA. Speech at the dedication of Ryerson Physics Lab, U. of Chicago 1894 as cited printed in the Annual Register, University of Chicago Press, 1896; 159.

2 Anderson C. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete, June 2008, Accessed online at http://www.wired.com/2008/06/pb-theory/, January, 2016.

3 Oxford Dictionaries, accessed online at www.oxforddictionaries.com/us, January 2016.

4 Atul Butte, TedMed, April 2012, accessed at http://www.tedmed.com/talks/show?id=7340, January 2016.

5 Hey T, Tansley S, Tolle K. The Fourth Paradigm: Data-Intensive Scientific Discovery. WA: Microsoft Research Redmond, 2009.

6 Kuhn T. The Structure of Scientific Revolutions. Chicago: The University of Chicago Press, 1962.

7 Mazzochi F. Could Big Data be the end of theory in science? EMBO Rep 2015; 16: 1250–5.

8 Leonelli S. What difference does quantity make? On the epistemology of Big Data in biology. Big Data & Society April–June 2014; 1–11.

9 Harford T. Big Data: are we making a big mistake? FT Magazine March 28, 2014 (http://www.ft.com/intl/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html).

10 Zoldan A. More Data More Problems: Is Big Data Always Right? WIRED (http://www.wired.com/insights/2013/05/more-data-more-problems-is-big-data-always-right/).

11 James G, Witten D, Hastie T, Tsibshirani R. An introduction to statistical learning. Heidelberg, Dordrecht, London: Springer New York; 5, 6, 24–29. ISSN 1431-875X, Library of Congress Control Number: 2013936251.

12 Lazer D, Kennedy R, King G, Vespignani A. The parable of Google flu: traps in Big Data analysis. Science 2014; 343: 1203–5.

13 Wu HY, Karnik S, Subhadarshini A, Wang Z, Philips S, Han X, Chiang C, Liu L, Boustani M, Rocha LM, Quinney SK, Flockhart D, Li L. An integrated pharmacokinetics ontology and corpus for text mining. BMC Bioinformatics 2013; 14: 35. doi:10.1186/1471-2105-14-35.

14 Han X, Quinney SK, Wang Z, Zhang P, Duke J, Desta Z, Elmendorf JS, Flockhart DA, Li L. Identification and mechanistic investigation of drug–drug interactions associated with myopathy: a translational approach. Clin Pharmacol Ther 2015; 98: 321–7. doi:10.1002/cpt.150.

15 Aziz RK, Monk JM, Lewis RM, Loh SI, Mishra A, Nagle AA, Satyanarayana C, Dhakshinamoorthy S, Luche M, Kitchen DB, Andrews KA, Fong NL, Li HJ, Palsson BO, Charusanti P. Systems biology-guided identification of synthetic lethal gene pairs and its potential use to discover antibiotic combinations. Sci Rep 2015; 5: 16025. doi:10.1038/srep16025.

16 Bordbar A, Johansson PI, Paglia G, Harrison SJ, Wichuk K, Magnúsdóttir M, Valgeirsdottir S, Gybel-Brask M, Ostrowski SR, Palsson S, Rolfsson O, Sigurjónsson OE, Hansen MB, Gudmundsson S, Palsson BO. Identified metabolic signature for assessing red blood cell unit quality is associated with endothelial damage markers and clinical outcomes. Transfusion 2016; Pubmed date 01/2016, epub ahead of print.