PharmacoEconomics (2016) 34:91–93
DOI 10.1007/s40273-015-0378-4

EDITORIAL

Big Data and Its Role in Health Economics and Outcomes Research: A Collection of Perspectives on Data Sources, Measurement, and Analysis

Eberechukwu Onukwugha[1]

Published online: 25 January 2016

Health economists and outcomes researchers have watched the term 'big data' increase in prominence over the last several years. However, to date, the use of big data in medicine has not been concretely illustrated across a variety of health economics and outcomes research (HEOR). At the same time, many of the same observers agree that fundamental questions remain unanswered and include (1) "What does the term 'big data' mean?" and (2) "What does the availability of big data mean for individuals who produce and use findings from HEOR?" This editorial tackles the first question and leaves contributors to this issue of *PharmacoEconomics* to discuss the promises, possibilities and potential pitfalls of using big data in HEOR.

Big data refers to large amounts of information that require new technologies for capture, storage or analysis, due in part to the amount of information, the speed at which it is generated, and its content. Two approaches have been used in formally defining big data [1]: the 3V definition [2, 3] and the 4V definition [4]. The 3V definition includes volume, variety and velocity [2, 3]. According to Berman [2], for data to be properly characterized as big data, we must be able to talk about "the size, complexity, and restlessness" of the data. The 4V definition adds 'value' to the list [4]. In their description of big data, Gantz and Reinsel [4] note that big data "is not only about the original content stored or being consumed but also about the information around its consumption." Thus, big data involves technology and system architecture, not only content. Smartphones are often used to illustrate the complexity of big data. In addition to providing information that is easily captured as flat files or simple-format records, they can also provide more complex data like geographic location, motion and direction information [2].

In this section, the 4V definition is used to describe big data in terms of its volume, variety, velocity and value. The large amounts of information from health records, social media, server logs, Web clickstream, machine/sensor and geolocation data illustrate the volume of big data [2, 5]. The variety of big data represents the different potential sources (e.g. administrative data, electronic health records, sensors, smartphones, social networks) and formats (e.g. video, audio, text or image) [1, 2]. The velocity of big data relates to the speed at which data transfers occur as well as to the rapidly changing nature of the data due to the sources, formats and categories of the contributing data [1, 2]. Value, "the most important aspect of big data" [1], relates to the untapped potential for drawing important, unique and transformative insights from big data. This fourth characteristic of big data has important implications for clinical research and population health research [6]. This special issue focuses on the application to population health research and particularly HEOR.

The full-length articles in this special issue draw on one or more of the components of the 4V definition and contribute to our understanding of the role of big data in HEOR. Contributions were sought that addressed important aspects of HEOR, including data sources, measurement, regression modelling and simulation. Additionally, the goal was to include a geographically diverse set of

✉ Eberechukwu Onukwugha
eonukwug@rx.umaryland.edu

[1] Department of Pharmaceutical Health Services Research, University of Maryland School of Pharmacy, 220 Arch Street, 12th floor, Baltimore, MD, USA

applications to highlight perspectives across institutional, government and health system settings. The articles illustrate innovative, linked data sources and discuss practical considerations for their development, reliability and use in HEOR (e.g. Lorgelly et al. [7] and Thorn et al. [8]). The articles discuss the practical challenges and opportunities with regards to measurement of healthcare cost and utilization using large, complex datasets (e.g. Canavan et al. [9], Payakachat et al. [10], Asaria et al. [11]). The articles utilize analytic methods and tools that are particularly suited for developing evidence from large-volume datasets. These articles illustrate the use of classification and regression trees for analysing prescribing patterns [12], clustering algorithms for cost prediction [13] and data visualization tools for examining prescription drug fill patterns [14]. Last but not least, the full-length articles in this special issue describe innovative opportunities for linking dynamic simulation modelling (DSM) with electronic health records [15] and integrating DSM with big data for evidence generation [16].

Together, these articles offer a much-needed snapshot of current data sources, analytic methods, opportunities and challenges. The hope is that future work will offer additional insights and lessons learned to increase our knowledge of the role of big data in HEOR. This knowledge base is important given that observational data, whether used for regression or simulation modelling, are critical to evidence generation in HEOR. We will need to be sure that big data provide more value and not more noise. As we consider the availability of more complex data, we cannot forget what we already know about the importance of study design or the appropriate interpretation of study findings. We cannot assume that more data necessarily means more information. Indeed, as the volume of data increases, it will be important to pay continued (or *more*) attention to established concerns regarding measurement, bias, and fallacies relevant to empirical analysis and interpretation. We should keep in mind points offered in William Crown's commentary [17] on model specification and big data, including his point that more data is not an automatic source of bias reduction. It is also important to note the strengths, weaknesses, opportunities and threats discussed in Brendan Collins' commentary [18].

With regards to the role of big data in HEOR, we will need thoughtful data linkages, model specifications, and interpretation to leverage the potential of big data. We will also need richer measures, including environmental measures (e.g. physical and social environmental measures of place and space), economic measures (e.g. income measured at an individual or small-area level), clinical measures ('health record'-type detail with 'experimental study'-type accuracy) and utilization measures (including accurate, validated measures where needed). In addition, we should have a clear sense of *what* may be missing (e.g. preferences, opportunity costs, direct nonmedical and intangible costs) and *who* may be missing (e.g. the uninsured, the homeless, the medically underserved, the insured individual who does not utilize health services) from the data to identify creative ways to leverage the breadth and depth of big data.

The papers in this special issue provide practical, provocative discussions regarding the use of large, complex data in HEOR. We hope that these papers spur continued discussions because the availability of big data is neither a silver bullet nor a temporary distraction. Developments in information technology will support its continued relevance into the foreseeable future. The opportunities for linking clinical, cost and contextual data, as well as the challenges that arise in this undertaking are welcome developments. They challenge us to continue efforts to improve the conduct and translation of HEOR for real-world impact.

Compliance with Ethical Standards

Conflicts of interest Dr. Onukwugha declares grant funding or consulting income from Bayer HealthCare Pharmaceuticals, Novartis, Janssen Analytics (division of Johnson and Johnson), IMS Health, Pfizer and Sanofi-Aventis.

References

1. Hashema IAT, Yaqooba I, Anuara NB, et al. The rise of "big data" on cloud computing: review and open research issues. Inf Syst. 2015;47:98–115. doi:10.1016/j.is.2014.07.006. http://www.sciencedirect.com/science/article/pii/S0306437914001288. Accessed 23 Dec 2015.
2. Berman JJ. Introduction, in: principles of big data. 2013. Boston: Morgan Kaufmann. p. xix–xxvi. doi:10.1016/B978-0-12-404576-7.09980-9. http://www.sciencedirect.com/science/article/pii/B9780124045767099809. Accessed 23 Dec 2015.
3. Douglas Laney. 3D Data Management: Controlling Data Volume, Velocity and Variety. Gartner. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf. Accessed 7 Jan 2016.
4. Gantz J, Reinsel D. Extracting value from chaos. IDC iView (2011), p. 1–12. http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf. Accessed 16 Dec 2015.
5. HortonWorks. Data Sheet: 5 Types of Hadoop Data. http://hortonworks.com/wp-content/uploads/downloads/2013/08/Hortonworks.5TypesOfData.v1.0.pdf. Accessed 28 Mar 2015.
6. Krumholz HM. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. Health Aff. 2014;33(7):1163–70. doi:10.1377/hlthaff.2014.0053.

7. Lorgelly PK, Doble B, Knott RJ, et al. Realising the value of linked data to health economic analyses of cancer care: a case study of cancer 2015. Pharmacoeconomics. doi:10.1007/s40273-015-0343-2. http://link.springer.com/article/10.1007/s40273-015-0343-2. Accessed 16 Dec 2015.
8. Thorn JC, Turner E, Hounsome L, et al. Validation of the hospital episode statistics outpatient dataset in England. doi:10.1007/s40273-015-0326-3. http://link.springer.com/article/10.1007/s40273-015-0326-3. Accessed 16 Dec 2015.
9. Canavan C, West J, Card T. Calculating total health service utilisation and costs from routinely collected electronic health records using the example of patients with irritable bowel syndrome before and after their first gastroenterology appointment. doi:10.1007/s40273-015-0339-y. http://link.springer.com/article/10.1007/s40273-015-0339-y. Accessed 16 Dec 2015.
10. Payakachat N, Tilford JM, Ungar WJ. National Database for Autism Research (NDAR): big data opportunities for health services research and health technology assessment. doi:10.1007/s40273-015-0331-6. http://link.springer.com/article/10.1007/s40273-015-0331-6. Accessed 16 Dec 2015.
11. Asaria M, Grasic K, Walker S. Using linked electronic health records to estimate healthcare costs: key challenges and opportunities. doi:10.1007/s40273-015-0358-8. http://link.springer.com/article/10.1007/s40273-015-0358-8. Accessed 16 Dec 2015.
12. Schilling C, Mortimer D, Dalziel K, et al. Using Classification and Regression Trees (CART) to identify prescribing thresholds for cardiovascular disease. doi:10.1007/s40273-015-0342-3. http://link.springer.com/article/10.1007/s40273-015-0342-3. Accessed 16 Dec 2015.
13. Onukwugha E, Qi R, Jayasekera J, et al. Cost prediction using a survival grouping algorithm: an application to incident prostate cancer cases. doi:10.1007/s40273-015-0368-6. http://link.springer.com/article/10.1007/s40273-015-0368-6.
14. Bjarnadottir MV, Malik S, Onukwugha E, et al. Understanding adherence and prescription patterns using large-scale claims data. doi:10.1007/s40273-015-0333-4. http://link.springer.com/article/10.1007/s40273-015-0333-4. Accessed 16 Dec 2015.
15. Johnson O, et al. NETIMIS: Dynamic simulation of health economics outcomes using big data.
16. Marshall DA, Burgos-Liz L, Pasupathy KS, et al. Transforming healthcare delivery: integrating dynamic simulation modelling and big data in health economics and outcomes research. doi:10.1007/s40273-015-0330-7. http://link.springer.com/article/10.1007/s40273-015-0330-7. Accessed 16 Dec 2015.
17. Crown WH. Specification issues in a big data context: controlling for the endogeneity of consumer and provider behaviours in healthcare treatment effects models. doi:10.1007/s40273-015-0362-z. http://link.springer.com/article/10.1007/s40273-015-0362-z. Accessed 16 Dec 2015.
18. Collins B. Big data and health economics: strengths, weaknesses, opportunities and threats. doi:10.1007/s40273-015-0306-7. http://link.springer.com/article/10.1007/s40273-015-0306-7. Accessed 16 Dec 2015.