International Journal of
Environmental Research and Public Health

MDPI

Editorial

Real World—Big Data Analytics in Healthcare

Daniele Piovani [1,2] and Stefanos Bonovas [1,2,*]

[1] Department of Biomedical Sciences, Humanitas University, Pieve Emanuele, 20090 Milan, Italy
[2] IRCCS Humanitas Research Hospital, Rozzano, 20089 Milan, Italy
* Correspondence: stefanos.bonovas@hunimed.eu

The term Big Data is used to describe extremely large datasets that are complex, multi-dimensional, unstructured, and heterogeneous and that are accumulating rapidly and may be analyzed with appropriate informatic and statistical methodologies to reveal patterns, trends, and associations [1]. In medical and healthcare research, Big Data sources include electronic health records (EHRs), administrative or claims databases, product and disease registries, smart/wearable/self-monitoring devices, and large-scale collaborations for the collection and storage of health data and biospecimens in biobanks.

The definition of what Big Data means with respect to health research, or at least a consensus of what this term means, was proposed by the Health Directorate of the Directorate-General for Research and Innovation of the European Commission: "*Big Data in health encompasses high volume, high diversity biological, clinical, environmental, and lifestyle information collected from single individuals to large cohorts, in relation to their health and wellness status, at one or several time points*" [2].

Big Data analytics techniques and methods, such as statistical analysis, data mining, machine learning, and deep learning, have made notable progress in the recent years, and are expected to develop even further in the near future [3]. With the ever-increasing quantities of data that are digitally collected and stored within healthcare organizations, there is a growing enthusiasm in the potential applications for Big Data analytics in the fields of diagnostics, precision medicine, computerized decision support for clinicians, pharmacological research aiming to cure diseases and develop new treatments, the early detection of adverse drug reactions, cost reduction in patient care, preventive medicine, and population health research [4].

At the same time, the term Real-World Data (RWD) is commonly used to describe data derived from sources other than traditional randomized-controlled trials (RCTs). These sources may include EHRs, pragmatic clinical trials, prospective or retrospective observational studies, health insurance claims, case reports, data obtained as part of routine public health surveillance, product and disease registries, patient surveys, or other real-world sources [5].

The Association of the British Pharmaceutical Industry has defined the RWD as "*data obtained by any non-interventional methodology that describes what is happening in normal clinical practice*" [6], while according to the U.S. Food and Drug Administration, the term RWD refers to "*data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources*" [7]. In the last few years, a consensus has formed that RWD offer valuable opportunities for generating robust clinical evidence (i.e., real-world evidence) regarding the use and potential benefits or harms of new therapies outside the context of RCTs [8,9].

Researchers typically understand RWD as observational data, distinct from data sourced from RCTs, and in a way similar to Big Data. Nevertheless, RWD and Big Data are not synonymous. In fact, Big Data represent a special kind of RWD, which are characterized by high volume, high velocity, high variety, high veracity, and high value (5 Vs) [10].

The promise of Big Data in healthcare depends on the ability to extract meaningful information from real-world, large-scale resources that may pave the way to scientific

Citation: Piovani, D.; Bonovas, S. Real World—Big Data Analytics in Healthcare. *Int. J. Environ. Res. Public Health* **2022**, *19*, 11677. https://doi.org/10.3390/ijerph191811677

Received: 1 September 2022
Accepted: 15 September 2022
Published: 16 September 2022

discoveries in the pathogenesis, diagnosis, prevention, treatment, and prognosis of diseases, and eventually revolutionize clinical medicine and public health [11–13]. When Big Data are analyzed with the aim of causal inference and not as a hypothesis-generating tool, special attention should be made to the very serious risk of residual confounding and a variety of biases including time-related biases [14]. Recent methodological advancements have made causal inferences from Big Data possible and, in certain cases, are highly effective in replicating the results of RCTs that are made available later [14,15]. This can be achieved successfully if a careful application of a comprehensive methodological and statistical analysis plan called "target trial emulation" is pursued [14]. The failure to correctly apply such a plan may result in Big Data yielding overly optimistic estimates of effects.

Despite these advancements, until now, Big Data analytics have not fulfilled the oversized expectations in the health sector, possibly because of several significant challenges that are summarized below [16–19]:

(a) Big Data are often unstructured, fragmented, heterogeneous, and in incompatible formats, and are thus difficult to aggregate and analyze;
(b) There are important issues regarding data security (privacy and confidentiality);
(c) A lack of data standardization, language barriers, and different terminologies;
(d) There are often problems with the accuracy and precision of data;
(e) Storage and transfers of data are associated with significant costs;
(f) Budget constraints—there is a shortage of focused and sustained funding;
(g) The awareness of Big Data analytics' capabilities among health care professionals is rather limited;
(h) A shortage of researchers with skills in Big Data—due to the constant evolution of science and technology, professionals who collect, process, extract, or analyze data (i.e., data scientists, biostatisticians, epidemiologists, and experts in advanced analytics and artificial intelligence) need to be regularly trained and kept up-to-date;
(i) There are often issues regarding data governance and data ownership;
(j) Healthcare organizations implementing Big Data analytics as a part of their information systems need to comply with high standards and regulatory legislation.

Additionally, Sir David Cox and colleagues have methodically discussed several challenges of a statistical/epidemiological nature that arise when analyzing Big Data in healthcare [20]:

(a) The relevance of the data for the purpose of the investigation (the data's fitness for purpose)—big datasets may not be representative of the target population, and the largeness of a dataset does not imply that the findings of the investigation (e.g., the patterns, trends, and associations) are free of bias;
(b) The need for well-established quality control and assurance procedures (data reliability)—Big Data are not collected for a specific purpose and may be subject to particular quality issues (e.g., measurement errors, missing data, errors in coding information buried in textual reports, etc.);
(c) The potential for overconfidence in the results obtained from statistical analyses of Big Data (i.e., conclusions being seriously overoptimistic) due to superficially highly precise, but potentially biased, estimates.

In conclusion, the application of Big Data in healthcare is a fast-growing field with great advances in data-generation and data-analysis methodologies. Despite the challenges outlined above, Big Data analytics have the potential for positive impacts and global implications; they are becoming increasingly important, as they enable investigations to be conducted and conclusions to be drawn that would otherwise be very difficult or even impossible. However, we should keep in mind that, when analyzing Big Data in health care research, we need to make careful use of statistical and epidemiological concepts together with an in-depth understanding of the data themselves.

Author Contributions: Conceptualization, D.P. and S.B.; writing—original draft preparation, S.B.; writing—review and editing, D.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Heads of Medicines Agencies (HMA) and European Medicines Agency (EMA). HMA-EMA Joint Big Data Taskforce Phase II Report: Evolving Data-Driven Regulation. 2019. Available online: https://www.ema.europa.eu/en/documents/other/hma-ema-joint-big-data-taskforce-phase-ii-report-evolving-data-driven-regulation_en.pdf (accessed on 1 September 2022).
2. Auffray, C.; Balling, R.; Barroso, I.; Bencze, L.; Benson, M.; Bergeron, J.; Bernal-Delgado, E.; Blomberg, N.; Bock, C.; Conesa, A.; et al. Making Sense of Big Data in Health Research: Towards an EU Action Plan. *Genome Med.* **2016**, *8*, 71. [CrossRef]
3. Chan, C.-L.; Chang, C.-C. Big Data, Decision Models, and Public Health. *Int. J. Environ. Res. Public Health* **2020**, *17*, 6723. [CrossRef]
4. Vayena, E.; Dzenowagis, J.; Brownstein, J.S.; Sheikh, A. Policy Implications of Big Data in the Health Sector. *Bull. World Health Organ.* **2018**, *96*, 66–68. [CrossRef] [PubMed]
5. Sherman, R.E.; Anderson, S.A.; Dal Pan, G.J.; Gray, G.W.; Gross, T.; Hunter, N.L.; LaVange, L.; Marinac-Dabic, D.; Marks, P.W.; Robb, M.A.; et al. Real-World Evidence—What Is It and What Can It Tell Us? *N. Engl. J. Med.* **2016**, *375*, 2293–2297. [CrossRef] [PubMed]
6. The Association of the British Pharmaceutical Industry (ABPI). Demonstrating Value with Real World Data: A Practical Guide. 2011. Available online: http://www.abpi.org.uk/publications/real-world-data (accessed on 1 September 2022).
7. U.S. Food and Drug Administration (FDA). Framework for FDA's Real-World Evidence Program. 2018. Available online: https://www.fda.gov/media/120060/download (accessed on 1 September 2022).
8. National Academies of Sciences, Engineering, and Medicine. *Real-World Evidence Generation and Evaluation of Therapeutics: Proceedings of a Workshop*; The National Academies Press: Washington, DC, USA, 2017. Available online: https://www.ncbi.nlm.nih.gov/books/NBK441694 (accessed on 1 September 2022).
9. International Society for Pharmacoepidemiology (ISPE). ISPE's Position on Real-World Evidence. 2020. Available online: https://pharmacoepi.org/pub/?id=136DECF1-C559-BA4F-92C4-CF6E3ED16BB6 (accessed on 1 September 2022).
10. Mehta, N.; Pandit, A. Concurrence of Big Data Analytics and Healthcare: A Systematic Review. *Int. J. Med. Inform.* **2018**, *114*, 57–65. [CrossRef] [PubMed]
11. Obermeyer, Z.; Emanuel, E.J. Predicting the Future—Big Data, Machine Learning, and Clinical Medicine. *N. Engl. J. Med.* **2016**, *375*, 1216–1219. [CrossRef] [PubMed]
12. Benke, K.; Benke, G. Artificial Intelligence and Big Data in Public Health. *Int. J. Environ. Res. Public Health* **2018**, *15*, 2796. [CrossRef] [PubMed]
13. Shilo, S.; Rossman, H.; Segal, E. Axes of a Revolution: Challenges and Promises of Big Data in Healthcare. *Nat. Med.* **2020**, *26*, 29–38. [CrossRef] [PubMed]
14. Hernán, M.A.; Robins, J.M. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *Am. J. Epidemiol.* **2016**, *183*, 758–764. [CrossRef] [PubMed]
15. Lodi, S.; Phillips, A.; Lundgren, J.; Logan, R.; Sharma, S.; Cole, S.R.; Babiker, A.; Law, M.; Chu, H.; Byrne, D.; et al. Effect Estimates in Randomized Trials and Observational Studies: Comparing Apples with Apples. *Am. J. Epidemiol.* **2019**, *188*, 1569–1577. [CrossRef] [PubMed]
16. Baro, E.; Degoul, S.; Beuscart, R.; Chazard, E. Toward a Literature-Driven Definition of Big Data in Healthcare. *BioMed Res. Int.* **2015**, *2015*, 639021. [CrossRef] [PubMed]
17. Kruse, C.S.; Goswamy, R.; Raval, Y.; Marawi, S. Challenges and Opportunities of Big Data in Health Care: A Systematic Review. *JMIR Med. Inform.* **2016**, *4*, e38. [CrossRef] [PubMed]
18. Galetsi, P.; Katsaliaki, K.; Kumar, S. Values, Challenges and Future Directions of Big Data Analytics in Healthcare: A Systematic Review. *Soc. Sci. Med.* **2019**, *241*, 112533. [CrossRef]
19. Borges do Nascimento, I.J.; Marcolino, M.S.; Abdulazeem, H.M.; Weerasekara, I.; Azzopardi-Muscat, N.; Gonçalves, M.A.; Novillo-Ortiz, D. Impact of Big Data Analytics on People's Health: Overview of Systematic Reviews and Recommendations for Future Studies. *J. Med. Int. Res.* **2021**, *23*, e27275. [CrossRef]
20. Cox, D.R.; Kartsonaki, C.; Keogh, R.H. Big Data: Some Statistical Issues. *Stat. Probab. Lett.* **2018**, *136*, 111–115. [CrossRef] [PubMed]