rfm
Revista *de la* Facultad *de* Medicina

REFLECTION PAPER
DOI: http://dx.doi.org/10.15446/revfacmed.v68n1.73456
Received: 11/07/2018 **Accepted:** 19/10/2018

Big data, pharmacoepidemiology and pharmacovigilance

Big data, farmacoepidemiología y farmacovigilancia

Jorge Andrés Sánchez-Duque[1,2], Andrés Gaviria-Mendoza[1,3], Paula Andrea Moreno-Gutiérrez[1,3], Jorge Enrique Machado-Alba[1]

[1] Universidad Tecnológica de Pereira - Faculty of Health Sciences - Audifarma S.A., Pharmacoepidemiology and Pharmacovigilance Research Group - Pereira - Colombia.
[2] Universidad Tecnológica de Pereira - Faculty of Health Sciences - Research Group on Epidemiology, Health and Violence - Pereira - Colombia.
[3] Fundación Universitaria Autónoma de las Américas - Pereira Campus - Faculty of Medicine - Biomedical Research Group - Pereira - Colombia.
Corresponding author: Jorge Enrique Machado-Alba. Audifarma S.A. Calle 105 No. 14-140, Zona Industrial de Occidente. Telephone number: +57 6 3137800, ext.: 6119; mobile: +57 3108326970. Pereira. Colombia. Email: machado@utp.edu.co.

Abstract

Big data is a term that comprises a group of technological tools capable of processing extremely large heterogeneous data sets, which are continuously collected and are available to be used at any time, and, therefore, constitutes a source of scientific evidence production.

In the pharmacoepidemiology field, analyses made using these data sets may result in the development of pharmacological therapies that are more efficient, less expensive, and have a lower occurrence rate of adverse reactions. Likewise, the use of tools such as Text Mining or Machine Learning has led to major advances in pharmacoepidemiology and pharmacovigilance areas, so it is likely that these tools will be increasingly used over time.

Keywords: Artificial Intelligence; Automatic Data Processing; Data Accuracy; Data Mining; Machine Learning; Registries (MeSH).

Sánchez-Duque JA, Gaviria-Mendoza A, Moreno-Gutiérrez PA, Machado-Alba JE. Big data, pharmacoepidemiology and pharmacovigilance. Rev. Fac. Med. 2020;68(1):117-20. English. doi: http://dx.doi.org/10.15446/revfacmed.v68n1.73456.

Resumen

Big data es un término que comprende un grupo de herramientas tecnológicas capaces de procesar conjuntos de datos heterogéneos extremadamente grandes, los cuales se recolectan de manera continua, están disponibles para ser usados y constituyen una fuente de evidencia científica.

En el área de la farmacoepidemiología, los análisis generados a partir de estos conjuntos de datos pueden resultar en la obtención de terapias médicas más eficientes, con menor número de reacciones adversas y menos costosas. Asimismo, el uso de herramientas como el *Text Mining* o el *Machine Learning* también ha llevado a grandes avances en las áreas de farmacoepidemiología y farmacovigilancia, por lo que es probable que su empleo sea cada vez mayor.

Palabras clave: Procesamiento automatizado de datos; Minería de datos; Aprendizaje automático; Exactitud de los datos; Inteligencia artificial; Sistema de registros (DeCS).

Sánchez-Duque JA, Gaviria-Mendoza A, Moreno-Gutiérrez PA, Machado-Alba JE. [Big data, farmacoepidemiología y farmacovigilancia]. Rev. Fac. Med. 2020;68(1):117-20. English. doi: http://dx.doi.org/10.15446/revfacmed.v68n1.73456.

Introduction

Big data is a term currently used by computer science to describe a range of technological tools capable of processing extensive data sets. Most such data are observational —also known as "real-world data"— and, when analyzed, can reveal patterns, trends, and associations related to human behavior and its interactions. These large-scale databases may consist of genetic, medical, environmental, economic, geographical, or social network data; for this reason, they are often so extensive and poorly organized that it is not possible to analyze them using traditional computing techniques.[1-4]

Despite its great popularity and multiple uses, there is no clear definition of the concept of big data. Therefore, its definition is based on the four "Vs": volume, velocity, variety, and veracity.[5,6] Volume refers to the availability of massive amounts of data (which requires flexible and easily expandable management, recovery, and storage systems). Velocity is the feature of the big data infrastructure that enables efficient data management. Variety means that the data comes in many formats. Finally, veracity is about reducing errors and unreliable information that affects data analysis and results. In other words, big data involves a large amount of heterogeneous data that is quickly updated and available for use, but it requires checking.[5-7]

Based on the above, this reflection article aims to describe general aspects of the current relevance of big data and its possible application in pharmacoepidemiology and pharmacovigilance. To this end, scientific literature published between 1 January 2000 and 30 November 2018 was searched. The databases consulted were MEDLINE via PubMed, ScienceDirect, and Scopus, and the search strategy included the MeSH terms ["Big data AND Pharmacoepidemiology"; "Big data AND Pharmacovigilance"].

Big data in the health area

Usually, multiple types of data are collected by different health professionals during administrative processes and clinical practice. They include, on the one hand, physicians who record the clinical history of their patients, the prescription of therapies, the results of laboratory tests and the reporting of adverse events, and, on the other hand, pharmacy personnel who record when medications are dispensed. All of this happens routinely.[7,8] Since this information is not collected for scientific research purposes, the data is not always "clean" or available for analysis by researchers; therefore, data accumulates over a long period of time, and its value is not fully recognized or exploited.[5,6] However, the usefulness of this information in health care is increasingly evident, so it is necessary to manage all this data full of scientific evidence.[7]

The use of databases in the health sector began to increase in the 1990s, particularly in Europe, North America and, more recently, Asia, where they have been widely used to assess post-marketing prescription patterns, comparative efficacy, and safety of marketed drugs.[9,10]

The ability to link databases in the health area allows integrating various sources of information to provide an overall picture of the patient's medical history and to carry out collaborative studies through international databases.[5,6,11,12] These techniques are convenient, as it would be extremely costly and time-consuming to collect such information otherwise.[13]

Large healthcare databases often contain information coded according to international classifications such as the International Classification of Diseases (ICD) and the Anatomical, Therapeutic, Chemical (ATC) classification system for drug information. They can also be found in the form of free, unstructured texts that require the use of artificial intelligence technology such as text mining.[7,14] There are two main types of machine learning that have been used in pharmacovigilance for automatic signal generation: supervised learning and unsupervised learning.

Unsupervised machine learning is a computer system that can learn associations between selected data elements on its own, i.e., without being "trained"; this approach has been used to identify complex drug safety signals and discover use patterns. In contrast, supervised machine learning requires "teaching" a computer system how to build an algorithm based on the desired result in advance.[6,15]

Another potential application for big data includes the so-called mobile health (mHealth) area. For some time, applications for smart electronic devices have been developed to help manage a large number of chronic diseases and conditions —such as diabetes and tobacco cessation— and even to improve nutritional habits.[3,16] The information collected from these devices allows for predictive modeling that can result in more efficient and cheaper medical therapies with fewer adverse reactions.[17]

Medical device manufacturers produce tools for use in routine services that monitor clinical marker levels and automatically submit information to complete electronic health records. This information, altogether, allows healthcare providers and government agencies to adjust the treatment plan by phone or applications, e-mails, or directly using the measurement device, thus promoting healthcare compliance.[2,3,5,17]

Big data for drugs in the post-marketing phase

In order to market a novel drug, researchers and manufacturers invest a great deal of time, money, and logistics. Moreover, different phases, which go from pre-clinical research to the first clinical application, must be successfully completed before they are finally approved by the regulatory bodies. Once the drugs are available to patients on the market, pharmacoepidemiology comes into play; it studies their use and effects (beneficial or adverse) in large populations in the post-marketing phase.[1,9,18]

Epidemiological surveillance has been fundamental in public health for decades, as it reports on the health status of patients based on data directly collected from healthcare institutions. These data include sociodemographic variables, clinical conditions, morbidities, laboratory reports, diagnostic and therapeutic strategies, adverse reactions, outcomes, survival, and mortality. This active surveillance is supported by intelligent electronic devices with internet access, in which patients report symptoms and other data that are updated in real time.[1,3] This can be used in the area of pharmacovigilance for reporting adverse drug events.[7,8,19]

The beginning of the technological revolution in the 1970s impacted surveillance systems by improving accessibility and increasing the speed with which data was transmitted between institutions. Similarly, there

was an increase in the number of data sources that can be used in pharmacoepidemiology and pharmacovigilance, covering spontaneous reporting systems, digitized healthcare databases, adverse reaction reports, among others.[3,6,8]

The creation of data systems that collect information on adverse event reports has been a breakthrough in the area of drug safety. Currently, there are international databases that collect such information, continually review it through signal analysis, and issue constant alerts about possible associations between an adverse event and a drug.[8,20,21] This methodology allows the continuous incorporation of data from various sources and its analysis in real time, which in turn allows the detection of possible alerts of unknown adverse reactions or whose magnitude could be greater than expected.[9,13]

Advances in pharmacoepidemiology and pharmacovigilance

Pharmacovigilance appeared more than 50 years ago in response to the harmful side effects caused by the drug thalidomide. In the early years, this science was based on anecdotal evidence and case series through systematic spontaneous reporting, so it did not provide a reliable estimate of incidence or risk. The second-generation shaped important observational studies that sought to understand the contributions of knowledge about potential adverse effects of new and old drugs. Finally, third-generation pharmacovigilance began with meta-analyses on clinical trials and made important contributions.[8]

Furthermore, in recent years, the potential for research based on healthcare databases has generated interest in the results of studies that show the risk association between the consumption of a drug and an adverse effect that could not have been identified during the follow-up time of a conventional clinical trial, such is the case of proton-pump inhibitors usage and the risk of myocardial infarction, [21] or certain drug interactions in the actual clinical context of patients treated with anticoagulants.[22]

The study of big data as a pharmacoepidemiology and pharmacovigilance strategy began in 1990, and, to date, it has proven to be cost-effective, fast, and reliable. Therefore, the Food and Drug Administration (FDA) has not only stated that this strategy has many advantages but has expanded its use to analyze the growing number of reports it receives.[7]

According to the relevant literature, there are several databases with enough information that allow conducting health studies and have a potential application in drug consumption analysis and pharmacovigilance studies. They include the Danish National Health Service Prescription Database,[23,24] the UK's Clinical Practice Research Datalink (CPRD),[25] the US FDA Adverse Event Reporting System (FAERS),[26,27] and the Scottish Prescribing Information System.[28]

In this context, there is evidence that different companies are increasingly using big data and artificial intelligence techniques to support pharmacovigilance activities. However, there is still a long way to go,[29] especially in Latin America, where this type of technology is underdeveloped in the areas of natural sciences and health.[30-32]

Even with the benefits they offer, these techniques have limitations, including the lack of quality standards and validation methods for some of their records, as they may be incomplete, inconsistent, and subject to a great deal of potential bias and confusion. On the other hand, the use of massive amounts of data may cause an existing relationship to go undetected due to the masking or dilution of a phenomenon.[7,33]

Conclusions

The availability of large amounts of healthcare data increases the power of analysis of this information and creates an opportunity to study drug use and safety. Given the high flow of information, big data techniques that allow performing various analysis procedures and obtaining results applicable to routine medical practice are required for the organization and codification of unstructured, and highly complex data. Managing and exploiting these expanding sources of information is the next challenge for the application of research methods in modern pharmacology.[1,6,17,34]

Another relevant advantage of the use of big data in pharmacoepidemiology and pharmacovigilance is the diversity of the data since medical records can be analyzed with information on hospitalization, outpatient consultations, drug prescriptions, and laboratory tests, besides opening up the possibility of continuous monitoring using intelligent electronic devices.[1,2,6]

Due to the limitations of secondary data sources, their interpretation is associated with some important challenges, such as accumulation of estimation errors and spurious correlation.[3] These massive data flows must adjust to changing conditions all the time, so the algorithmic intelligence of digital epidemiology must be harnessed. In this regard, new technologies must be regulated by public health institutions so that data is properly distributed, and high standards of accuracy are maintained.[1,6]

Conflict of interest

The authors are members of the Grupo de Investigación de Farmacoepidemiología y Farmacovigilancia (Pharmacoepidemiology and Pharmacovigilance Research Group) of the Universidad Tecnológica de Pereira in agreement with Audifarma S.A. AGM and JEMA have a contractual relationship with Audifarma S.A.

Funding

This manuscript was financially supported by Audifarma S.A.

Acknowledgements

None stated by the authors.

References

1. Saint-Gerons MD, de la Fuente-Honrubia C, de Andrés-Trelles F, Catalá-López F. Perspectiva futura de la farmacoepidemiología en la era del "Big data" y la expansión de las fuentes de información. Rev Esp Salud Pública. 2016;90(1):1-7.

2. Stokes LB, Rogers JW, Hertig JB, Weber RJ. Big data: Implications for Health system pharmacy. Hosp Pharm. 2016;51(7):599-603. http://doi.org/c8d7.
3. Hernandez I, Zhang Y. Using predictive analytics and big data to optimize pharmaceutical outcomes. Am J Health Syst Pharm. 2017;74(18):1494-500. http://doi.org/gbx3fx.
4. Issa NT, Byers SW, Dakshanamurthy S. Big data: the next frontier for innovation in therapeutics and healthcare. Expert Rev Clin Pharmacol. 2014;7(3):293-298. http://doi.org/f55ppj.
5. Baldwin JN, Bootman JL, Carter RA, Crabtree BL, Piascik P, Ekoma JO, *et al.* Pharmacy practice, education, and research in the era of big data: 2014-15 Argus Commission Report. Am J Pharm Educ. 2015;79(10):S26. http://doi.org/c8ff.
6. Trifirò G, Sultana J, Bate A. From Big Data to Smart Data for Pharmacovigilance: The Role of Healthcare Databases and Other Emerging Sources. Drug Saf. 2018;41(2):143-9. http://doi.org/gc2j4d.
7. Ventola CL. Big Data and pharmacovigilance: data mining for adverse drug events and interactions. P T. 2018;43(6):340-51.
8. Laporte JR. Fifty years of pharmacovigilance-medicines safety and public health. Pharmacoepidemiol Drug Saf. 2016;25(6):725-32. http://doi.org/c8fg.
9. Chen B, Butte AJ. Leveraging big data to transform target selection and drug discovery. Clin Pharmacol Ther. 2016;99(3):285-97. http://doi.org/f8bkzd.
10. More S, Joshi P. Novel approach for Data Mining of Social Media to Improve Health Care using Network-Based Modeling. IJETT. 2017;4(Special).
11. Yang CT, Liu JC, Chen ST, Lu HW. Implementation of a Big Data Accessing and Processing Platform for Medical Records in Cloud. J Med Syst. 2017;41(10):149. http://doi.org/gb456b.
12. Alonso SG, de la Torre Díez I, Rodrigues JJPC, Hamrioui S, López-Coronado M. A Systematic Review of Techniques and Sources of Big Data in the Healthcare Sector. J Med Syst. 2017;41(11):183. http://doi.org/gch262.
13. Wilson AM, Thabane L, Holbrook A. Application of data mining techniques in pharmacovigilance. Br J Clin Pharmacol. 2004;57(2):127-34. http://doi.org/dnvp2h.
14. Ben-Abacha A, Chowdhury MFM, Karanasiou A, Mrabet Y, Lavelli A, Zweigenbaum P. Text mining for pharmacovigilance: Using machine learning for drug name recognition and drug-drug interaction extraction and classification. J Biomed Inform. 2015;58:122-32. http://doi.org/f74w4j.
15. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. JAMA. 2018;319(13):1317-8. http://doi.org/gc7qpm.
16. Fernández-Silano M. La Salud 2.0 y la atención de la salud en la era digital. Revista Médica de Risaralda. 2014;20(1):41-6.
17. Flockhart D, Bies RR, Gastonguay MR, Schwartz SL. Big data: challenges and opportunities for clinical pharmacology. Br J Clin Pharmacol. 2016;81(5):804-6. http://doi.org/c8fk.
18. Sánchez-Duque JA, García-Zuluaga AF, Betancourt-Quevedo R, Alzate-González MF. ¿Es hora de regular los productos y suplementos herbales? CIMEL. 2018;23(2). http://doi.org/c8fm.
19. Salathé M. Digital Pharmacovigilance and Disease Surveillance: Combining Traditional and Big-Data Systems for Better Public Health. JID. 2016;214(Suppl 4):S399-S403. http://doi.org/f9pvm7.
20. Harpaz R, DuMochel W, Shah NH. Big data and adverse drug reaction detection. Clin Pharmacol Ther. 2016;99(3):268-70. http://doi.org/c8fn.
21. Shah NH, LePendu P, Bauer-Mehren A, Ghebremariam YT, Iyer SV, Marcus J, *et al.* Proton pump inhibitor usage and the risk of myocardial infarction in the general population. PLoS One. 2015;10(6):e0124653. http://doi.org/f743hs.
22. Chang SH, Chou IJ, Yeh YH, Chiou MJ, Wen MS, Kuo CT, *et al.* Association between use of non-vitamin k oral anticoagulants with and without concurrent medications and risk of major bleeding in nonvalvular atrial fibrillation. JAMA. 2017;318(13):1250-9. http://doi.org/gbzw2f.
23. Pedersen LH, Petersen OB, Nørgaard M, Ekelund C, Pedersen L, Tabor A, *et al.* Linkage between the Danish National Health Service Prescription Database, the Danish Fetal Medicine Database, and other Danish registries as a tool for the study of drug safety in pregnancy. Clin Epidemiol. 2016;8:91-5. http://doi.org/c8fp.
24. Pottegård A, Schmidt SAJ, Wallach-Kildemoes H, Sørensen HT, Hallas J, Schmidt M. Data Resource Profile: The Danish National rescription Registry. Int J Epidemiol. 2017;46(3):798-798f. http://doi.org/c8fq.
25. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, *et al.* Data Resource Profile: Clinical Practice Research Datalink (CPRD). Int J Epidemiol. 2015;44(3):827-36. http://doi.org/f7ndhg.
26. Blau JE, Tella SH, Taylor SI, Rother KI. Ketoacidosis associated with SGLT2 inhibitor treatment: Analysis of FAERS data. Diabetes Metab Res Rev. 2017;33(8):e2924. http://doi.org/c8fr.
27. Wang K, Wan M, Wang RS, Weng Z. Opportunities for Web-based Drug Repositioning: Searching for Potential Antihypertensive Agents with Hypotension Adverse Events. J Med Internet Res. 2016;18(4):e76. http://doi.org/f8w652.
28. Álvarez-Madrazo S, McTaggart S, Nangle C, Nicholson E, Bennie M. Data Resource Profile: The Scottish National Prescribing Information System (PIS). Int J Epidemiol. 2016;45(3):714-715f. http://doi.org/c8fs.
29. Donzanti BA. Pharmacovigilance is Everyone's Concern: Let's Work It Out Together. Clin Ther. 2018;40(12):1967-72. http://doi.org/gfs8tk.
30. Fernández A, Gómez A, Lecumberry F, Pardo A, Ramírez I. Pattern Recognition in Latin America in the "Big Data" Era. Pattern Recognit. 2015;48(4):1185-96. http://doi.org/c8ft.
31. Noreña-P A, González-Muñoz A, Mosquera-Rendón J, Botero K, Cristancho MA. Colombia, an unknown genetic diversity in the era of Big Data. BMC Genomics. 2018;19(Suppl 8):859. http://doi.org/c8fv.
32. Lombi F, Varela CF, Martinez R, Greloni G, Campolo-Girard V, Rosa-Diez G. Acute kidney injury in Latin America in "big data" era. Nefrologia. 2017;37(5):461-4. http://doi.org/c8fw.
33. Purcell PM. Data Mining in Pharmacovigilance. Int J Pharm Med. 2003;17(2):63-4. http://doi.org/dkbgbf.
34. Xie L, Draizen EJ, Bourne PE. Harnessing big data for systems pharmacology. Annu Rev Pharmacol Toxicol. 2017;57:245-62. http://doi.org/c8fx.