RESS | REVISTA DO SUS

EDITORIAL

doi 10.1590/S2237-96222022000300001

Population data science

The rapid evolution of information technology provides unprecedented growth in data production, storage, exchange, processing and analysis capacity.[1] In addition to structured data, various unstructured data formats are currently made available. Structured data are used in administrative databases, which are produced and used in typical Public Health activities. Regarding this type of database, each row corresponds to an entity, usually an individual, and each column corresponds to an attribute of that entity (for example, date of birth). In turn, unstructured data present different formats, which include, for example, texts in documents or social networks, images, and sensor outputs. These data, especially when linked, have been increasingly used in health, research, surveillance and evaluation activities, as well as in decision-making.[2]

This new complex data ecosystem encouraged researchers from the International Population Data Linkage Network (IPDLN; available at http://www.ipdln.org/) to publish an article in 2018 proposing the creation of a new disciplinary field, which they called population data science.[3] Population Data Science is "a multi-disciplinary field aimed at obtaining population-level insights with public value by organizing, linking or otherwise integrating and analyzing data that pertain to individuals and their social, economic, biological and environmental characteristics and contexts." Or, in short, "the science of data about people".[3] The intensive use of complex databases – resulting from the linkage or integration of individual data of a diverse nature, with population coverage, for the generation of evidence of value to society – is the main characteristic of this new disciplinary field, and that distinguishes it from other related disciplines, such as data science and informatics.[3]

With regard to this disciplinary field, two important challenges must be addressed. Initially, it is necessary to implement technical infrastructure and data access governance policies that respect ethical and privacy norms, as well as meet society's expectations.[4] The data center model – which operates as a reliable third party in the relationship between database custodians and researchers and other actors interested in the use of linked databases – was adopted by several countries,[5] ensuring the balance between guaranteeing privacy preservation and enabling efficient and secure access to linked data, through projects that aim to produce knowledge that is relevant to society.[4]

The second challenge is related to the lack of familiarity of researchers and managers with the use of large and complex secondary databases for surveillance, evaluation and research purposes, which may lead to misinterpretation of the evidence generated.[2,6] Unlike primary data collected to answer a specific research question, researchers and managers generally have no control over the generation and processing of secondary datasets. When using a linked secondary database, it is necessary to know several aspects, including population coverage, field completeness, presence of duplicate records, proportion of missing data, reliability and validity of data, the attribute coding systems, the algorithms used for data transformation, and the data linkage process, including linking errors.[2] Additionally, taking into consideration that these data are collected in different locations and over long periods of time, it is important to evaluate whether these aspects are stable over time, and whether there are regional inequalities.

In Brazil, we have a consolidated experience in the availability and use of unidentified individual microdata for research, evaluation and surveillance purposes. However, in order for this successful

experience to advance, several actions would be necessary, including the implementation of technical infrastructure and governance policies for access to linked unidentified microdata, the dissemination and updating of the metadata of the original databases and derived linked datasets, and the training of researchers and managers in the domains that comprise the field of population data science.

CONFLICTS OF INTEREST

Cláudia Medina Coeli is a member of the Editorial Committee of Epidemiology and Health Services: Journal of the Brazilian National Health System (*Epidemiologia e Serviços de Saúde: revista do SUS* - RESS)

Correspondence: Cláudia Medina Coeli | coelicm@gmail.com

Cláudia Medina Coeli[1]

[1]Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, Brazil

REFERENCES

1. Mabry PL. Making Sense of the Data Explosion: The Promise of Systems Science. Am J Prev Med. 2011;40(5 Supl 2):159-61. doi:10.1016/j.amepre.2011.02.001
2. Christen P, Schnell R. Big Data is not the New Oil: Common Misconceptions about Population Data. arXiv. 2022. arXiv:2112.10912v3. doi:10.48550/arXiv.2112.10912
3. McGrail K, Jones K, Akbari A, Bennett T, Boyd A, Carinci F, et al. A Position Statement on Population Data Science. Int J Popul Data Sci. 2018;3(1):415. doi:10.23889/ijpds.v3i1.415
4. Ark TK, Kesselring S, Hills B, McGrail K. Population Data BC: Supporting population data science in British Columbia. Int J Popul Data Sci. 2020;4(2):1133. doi: 10.23889/ijpds.v5i1.1133
5. Coeli CM, Pinheiro RS, Camargo Junior KR. Conquistas e desafios para o emprego das técnicas de record linkage na pesquisa e avaliação em saúde no Brasil. Epidemiol Serv Saude. 2015;24(4):795-802. doi:10.5123/S1679-49742015000400023
6. Leonelli S. A pesquisa científica na era do Big Data: cinco maneiras que mostram como o big data prejudica a ciência, e como podemos salvá-la. Rio de Janeiro: Fiocruz; 2022. 149 p.