

Neither better nor worse, simply different

Nem melhor nem pior, apenas diferente

Ni mejor ni peor, sólo diferente

*Cláudia Medina Coeli*¹
*Rejane Sobrino Pinheiro*¹
*Marília Sá Carvalho*²

¹ Instituto de Estudos em Saúde Coletiva, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil.
² Programa de Computação Científica, Fundação Oswaldo Cruz, Rio de Janeiro, Brasil.

Correspondence

C. M. Coeli
 Instituto de Estudos em Saúde Coletiva, Universidade Federal do Rio de Janeiro.
 Pça. Jorge Moreira Machado
 1000, Cidade Universitária,
 Rio de Janeiro, RJ
 21941-598, Brasil.
 coelicm@gmail.com

Have you ever suffered discrimination because you used secondary data in your research? Since the principal area of research for this article's three authors involves the development and application of techniques for using secondary data, our answer is definitely no. However, we frequently hear complaints by colleagues who have encountered barriers to developing their theses or obtaining research funding because they opted to use secondary data.

A recent article by Rothman¹ discusses six erroneous perceptions regarding aspects of epidemiological research that are often reinforced in classrooms and textbooks. Although the author did not discuss data sources, we believe that the list should add a seventh misconception: the notion that primary data are the only valid source for epidemiological studies.

Population, vital, epidemiological, administrative, and clinical data have undergone important changes in their production and dissemination. They are now available in online databases that include millions of individual micro-data. In addition to the above-mentioned traditional sources, other modalities have emerged. The digital trails produced in accessing different web-based communication platforms and mobile phones have been used in studies about how patterns of behavior and mobility influence the determination and spread of diseases².

Secondary data have the potential to back studies on highly relevant public health issues, particularly due to their wide availability, scope, and coverage. They are actually the best data to answer questions on the determinants of incidence rates in populations, as suggested by Rose³. Even so, it is important to discuss how the two worlds are brought together. For example, gene-environment interaction requires the use of increasingly larger study populations. The context of "big epidemiology"⁴ stimulates the practice of "data sharing", whereby the data collected for specific studies are used by researchers not originally involved in their planning and execution.

The age of "big data" has brought about the recommendation of using this wealth of data in research⁵, including population health research⁶. However, several authors have emphasized the need for responsible use of such databases⁷. The main criticisms aimed at secondary data sources are the absence of mechanisms for data quality assurance and control and the lack of necessary variables for adequately testing causal hypotheses at the individual level.

Quality is a crucial issue. One should evaluate the different dimensions of quality⁸ before using a secondary data source. Meanwhile, database custodians should employ techniques to prevent, detect, and repair errors⁹ and make extensive documentation available on their data

collections. Financing infrastructure for data management and access is an essential element in policies to encourage the use of secondary data^{5,6}. In relation to the available variables for analysis, the integration of databases through record linkage techniques¹⁰ can contribute to better specification of exposure and outcome variables, in addition to expanding the number of variables available for adjustment for confounding. In addition, some methodological solutions have been proposed to mitigate the problem of unmeasured confounding factors¹¹. Finally, interest has grown in answering non-etiological questions, which do not require adjustment for confounding. One example are questions regarding the evaluation of public health interventions, which can be answered using different types of data, together with the application of new analytical techniques, for example data mining and computational modeling of complex systems^{2,6,10}.

Beyond the methodological issues, responsible use should also contemplate respect for privacy. This requires the development of an ethical framework that considers the specificities of research based on secondary data, especially informed consent¹². Brazil recently passed *Law n. 12,527*, regulating access to public information¹³. Care should be taken to prevent overly conservative interpretations of the law from resulting in unnecessary restrictions on the disclosure of anonymous database contents or on access to identified databases (while maintaining the necessary safeguards). According to a study by the U.S. National Research Council, the American legislation governing health information trans-

fer (*HIPAA Privacy Rule*) had negative impacts on relevant research for public health¹⁴. In Brazil, the legislation should seek a balance between individual rights and collective interests to avoid jeopardizing studies that aim to improve health, healthcare, and living conditions for users of the Unified National Health System.

The use of secondary data in research requires investments in human resource training. If, on the one hand, research teams increasingly need to incorporate information technology professionals, on the other, we need public health researchers capable of interacting with them, as interactive experts as defined by Collins et al.¹⁵. The necessary skill set and minimum expected level of expertise remain open questions. Relevant contents include SQL (Structured Query Language), record linkage, unstructured data integration, data mining, and computational modeling of complex systems.

We finished this paper in Rio de Janeiro during Carnival, which features the parade of samba schools as one of the city's most important tourist events. The article's title was inspired by a samba refrain coined in the 1960s by Nelson de Andrade, then-president of the Salgueiro samba school. The original refrain, "*Neither better nor worse, simply a different School!*" meant to highlight a creative revolution in Rio's Carnival led by Fernando Pamplona and Arlindo Rodrigues¹⁶. Secondary data represent a valuable source for research in public health. Taking maximum advantage of the data also requires a revolution: thinking differently, training differently, and doing differently.

Contributors

All the authors participated in all stages of the article's production.

1. Rothman KJ. Six persistent research misconceptions. *J Gen Intern Med* 2014; Epub ahead of print.
2. Salathé M, Bengtsson L, Bodnar TJ, Brewer DD, Brownstein JS, Buckee C, et al. Digital epidemiology. *PLoS Comput Biol* 2012; 8:e1002616.
3. Rose G. Sick individuals and sick populations. *Int J Epidemiol* 1985; 14:32-8.
4. Thompson A. Thinking big: large-scale collaborative research in observational epidemiology. *Eur J Epidemiol* 2009; 24:727-31.
5. Community cleverness required (Editorial). *Nature* 2008; 455:1.
6. Mabry PL. Making sense of the data explosion: the promise of systems science. *Am J Prev Med* 2011; 40(5 Suppl 2):S159-61.
7. Hernán MA. With great data comes great responsibility. *Epidemiology* 2011; 22:290-1.
8. Lima CRA, Schramm JMA, Coeli CM, Silva MEM. Revisão das dimensões de qualidade dos dados e métodos aplicados na avaliação dos sistemas de informação em saúde. *Cad Saúde Pública* 2009; 25:2095-109.
9. Herzog TN, Scheuren FJ, Winkler WE. *Data quality and record linkage techniques*. New York: Springer; 2007.
10. Christen P. *Data matching concepts and techniques for record linkage, entity resolution, and duplicate detection*. Berlin/New York: Springer; 2012.
11. Toh S, García-Rodríguez LA, Hernán MA. Analyzing partially missing confounder information in comparative effectiveness and safety research of therapeutics. *Pharmacoepidemiol Drug Saf* 2012; 21:13-20.
12. da Silva MEM, Coeli CM, Ventura M, Palacios M, Magnanini MM, Camargo TM, et al. Informed consent for record linkage: a systematic review. *J Med Ethics* 2012; 38:639-42.
13. Ventura M. Lei de acesso à informação, privacidade e a pesquisa em saúde. *Cad Saúde Pública* 2013; 29:636-8.
14. National Research Council. *Beyond the HIPAA privacy rule: enhancing privacy, improving health through research*. Washington DC: The National Academies Press; 2009.
15. Collins H, Evans R, Ribeiro R, Hall M. Experiments with interactional expertise. *Stud Hist Philos Sci Part A* 2006; 37:656-74.
16. Faria GJM. Nem melhor, nem pior, apenas uma escola diferente: os Acadêmicos do Salgueiro e as transformações estéticas e ideológicas na cultura brasileira (1959-1971). *Revista Litteris* 2010; (6). <http://www.revistaliteris.com.br>.

Submitted on 14/Apr/2014

Approved on 15/Apr/2014