RESS | REVISTA DO SUS

EDITORIAL

doi 10.1590/S2237-96222022000300001

Ciência de dados populacionais

Population data science

Ciencia de datos poblacionales

A rápida evolução das tecnologias de informação proporciona um crescimento sem precedentes da capacidade de produção, armazenamento, troca, processamento e análise de dados.[1] Além dos dados estruturados, vários formatos de dados não estruturados são hoje disponibilizados. Dados estruturados são usados em bases administrativas produzidas e utilizadas nas atividades típicas da Saúde Coletiva. Nesse tipo de base, cada linha corresponde a uma entidade, geralmente um indivíduo, e cada coluna, a um atributo dessa entidade (por exemplo, a data de nascimento). Por seu turno, os dados não estruturados apresentam diferentes formatos, nos quais se incluem, por exemplo, textos em documentos ou em redes sociais, imagens, e saídas de sensores. Esses dados, especialmente quando vinculados, vêm, cada vez mais, sendo empregados na saúde, em atividades de pesquisa, vigilância e avaliação, bem como na tomada de decisão.[2]

Esse novo complexo ecossistema de dados estimulou pesquisadores da International Population Data Linkage Network (IPDLN; disponível em http://www.ipdln.org/) a publicarem, em 2018, um artigo propondo a criação de um novo campo disciplinar, ao qual deram a denominação de ciência de dados populacionais (*population data science*).[3] A ciência de dados populacionais é definida como "um campo multidisciplinar destinado a obter informação, ao nível populacional, com valor para a sociedade, por meio da organização, integração e análise de dados de indivíduos, assim como dos seus contextos social, econômico, biológico e ambiental", ou, de forma mais resumida, "a ciência de dados sobre pessoas".[3] O uso intensivo de bases de dados complexas – resultantes da vinculação ou integração de dados individuais de natureza diversa, com abrangência populacional, para a geração de evidências com valor para a sociedade – é a principal característica desse novo campo disciplinar, que o distingue de outras disciplinas correlacionadas, como a ciência de dados e a informática.[3]

Nesse campo disciplinar, dois desafios importantes devem ser enfrentados. Em primeiro lugar, faz-se necessária a implantação de infraestrutura técnica e de políticas de governança de acesso a dados que respeitem normas éticas e de privacidade, bem como atendam às expectativas da sociedade.[4] O modelo de centro de dados – que opera como terceira parte confiável na relação entre gestores responsáveis pela custódia de bases de dados e pesquisadores e outros atores interessados no uso de bases vinculadas – foi adotado por vários países,[5] assegurando o equilíbrio entre a garantia da preservação da privacidade e a viabilização do acesso eficiente e seguro a dados vinculados, por projetos que tenham como objetivo a produção de conhecimento relevante para a sociedade.[4]

O segundo desafio diz respeito à falta de familiaridade de pesquisadores e gestores com a utilização de bases de dados secundárias volumosas e complexas para fins de vigilância, avaliação e pesquisa, o que pode levar a interpretações incorretas das evidências geradas.[2,6] Ao contrário do que ocorre com dados primários coletados para resposta a uma pergunta específica de pesquisa, pesquisadores e gestores, em geral, não têm controle sobre os processos de geração e processamento dos conjuntos de dados secundários. Ao se utilizar um banco de dados secundários

vinculado, é fundamental conhecer vários aspectos, incluindo-se a cobertura populacional, a completude de campos, a presença de registros duplicados, a proporção de dados faltantes, a confiabilidade e validade dos dados, os sistemas de codificação de atributos, os algoritmos empregados para a transformação de dados, e o processo de vinculação de dados, inclusive os erros de vinculação.[2] Adicionalmente, considerando-se que esses dados são coletados em diferentes locais e em períodos longos de tempo, é importante avaliar se esses aspectos são estáveis no tempo, e se existem desigualdades regionais.

No Brasil, temos uma experiência já consolidada na disponibilidade e uso para fins de pesquisa, avaliação e vigilância de microdados individuais não identificados. Entretanto, para que essa experiência bem-sucedida possa avançar, são necessárias várias ações, incluindo-se a implantação de infraestrutura técnica e de políticas de governança de acesso a microdados vinculados não identificados, a divulgação e atualização dos metadados das bases de dados originais e dos conjuntos de dados vinculados derivados, e a formação de pesquisadores e gestores nos domínios que compreendem o campo da ciência de dados populacionais.

CONFLITOS DE INTERESSE

Cláudia Medina Coeli é membro do Comitê Editorial da Revista *Epidemiologia e Serviços de Saúde: revista do SUS* (RESS).

Correspondência: Cláudia Medina Coeli | coelicm@gmail.com

Cláudia Medina Coeli[1]

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

REFERÊNCIAS

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.