CSP CADERNOS DE SAÚDE PÚBLICA
REPORTS IN PUBLIC HEALTH

De dados secundários à Ciência de Dados Populacionais: recordando 40 anos da produção científica nas páginas de CSP

Cláudia Medina Coeli [1]

doi: 10.1590/0102-311XPT087624

[1] *Instituto de Estudos em Saúde Coletiva, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil.*

Foi com grande alegria que aceitei o convite para escrever este editorial. Uma oportunidade especial para celebrar junto a Marilia Sá Carvalho, Luciana Dias de Lima, Luciana Correia Alves e toda a comunidade de CSP os 40 anos deste importante projeto editorial, no qual tive a honra de atuar por nove anos como Coeditora-Chefe. Tendo como meu foco principal em pesquisa o desenvolvimento de técnicas e o uso de bases de dados secundários, rever a produção científica desse tema em CSP me permitiu recordar artigos que foram referências fundamentais para minha formação e desenvolvimento de meus projetos de pesquisa.

O primeiro número de CSP nasceu em 1985. No contexto internacional, a venda de computadores pessoais (PC) [1] ganhava momento, sendo seguida, no início dos anos 1990, pela abertura ao público do acesso à World Wide Web (WWW) [2]. Esses avanços foram significativos para a popularização das tecnologias de informação.

Bases administrativas passaram a ser empregadas como fontes de dados secundários na pesquisa em Saúde Coletiva [3]. Nos anos 1990 e na primeira década dos anos 2000, Centros de Dados foram implantados na Austrália, Canadá e Reino Unido. Nessas organizações, bases administrativas são vinculadas continuamente, e os conjuntos de dados anonimizados resultantes podem ser acessados por pesquisadores para o desenvolvimento de seus projetos [4].

No Brasil, foi criado, em 1991, o Departamento de Informática do Sistema Único de Saúde (DATASUS) [5], que contribuiu significativamente para a acessibilidade às bases administrativas brasileiras. O modelo adotado para a disseminação de dados foi, contudo, diferente do Centro de Dados mencionado anteriormente. Foram disponibilizadas duas modalidades de acesso. A primeira é por meio de um tabulador *online* que permite criar tabelas dos principais sistemas de informações em saúde nacionais. A segunda consiste na disseminação de microdados não-identificados. Inicialmente, as bases eram distribuídas em *compact discs* (CDs) mensais, sendo, posteriormente, disponibilizadas para transferências *online*. Informações sobre nascimentos, óbitos, doenças e agravos de notificação, atenção básica, cuidados ambulatoriais e hospitalares, estabelecimentos de saúde e orçamento público passaram a ser disponibilizados não apenas para pesquisadores, mas também para toda a população. Esse modelo de dados abertos é singular por sua inovação, variedade

de dados, abrangência temporal e territorial das bases e acesso inclusivo. Informações em formato digital de interesse para a saúde também passaram a ser disponibilizados por diferentes instituições como o Instituto Brasileiro de Geografia e Estatística (IBGE), a Agência Nacional de Saúde Suplementar (ANS), a Agência Nacional de Vigilância Sanitária (Anvisa), além das secretarias de saúde de estados e municípios.

Mesmo antes da disseminação digital, dados administrativos, especialmente sobre mortalidade, eram utilizados no Brasil para a pesquisa em saúde coletiva. Contudo, a facilidade de acesso proporcionada pela adesão de instituições brasileiras ao modelo de dados abertos incentivou esse tipo de uso. Por meio da consulta ao PubMed, identifiquei 461 artigos publicados em CSP que usaram dados administrativos, dos quais 86 abordaram temas relacionados à qualidade. Destacam-se, entre esses, o artigo que é fruto da tese de Claudia Risso de Araujo Lima [6]. Claudia, que foi membro da equipe do DATASUS, foi uma das responsáveis pela implementação da política de disseminação de informações sobre saúde no Brasil. Publicado em 2009, seu artigo continua sendo referenciado até hoje (96 citações na base Scopus). Sua grande contribuição é realizar uma revisão sobre dimensões de qualidade na avaliação de sistemas de informação em saúde do Brasil.

A publicação de artigos que avaliam a qualidade, sejam dos sistemas de informação, sejam de processos para a vinculação de bases, atende a uma crescente demanda para a adoção de boas práticas na condução e relato de estudos que usam dados secundários [7,8]. Um editorial [9] e um artigo de perspectivas [10] reforçam a política editorial de CSP de valorização do uso responsável de bases administrativas na pesquisa.

CSP também publicou quatro artigos metodológicos que apresentam rotinas computacionais para o processamento de bases de dados. Três soluções foram voltadas para o relacionamento de dados (*record linkage*) [11,12,13], enquanto a última, o pacote Microdatasus [14], otimiza o *download* e o pré-processamento de microdados disponibilizados pelo DATASUS. O software Reclink foi publicado em 2000, como software livre de código fechado [11]. A nova versão OpenReclink foi publicada em 2015, já com código aberto [12]. Também são de código aberto o EPPD [13] e o Microdatasus [14], atendendo à política editorial de CSP de adesão à ciência aberta [15].

Nesses 40 anos, houve uma expansão acelerada das tecnologias de informação. Avanços na capacidade de captação, processamento, armazenamento, transmissão e análise de dados foram ocorrendo sucessivamente, com avanços incrementais em cada área, estimulando o avanço nas demais. Atualmente, é possível processar grandes quantidades de informações em tempo real. Dados não estruturados que apresentam diferentes formatos, como textos em documentos ou redes sociais, imagens e saídas de sensores, são novas fontes para uso secundário em pesquisas. Adicionalmente, ocorre a introdução na pesquisa em saúde de técnicas desenvolvidas pela Ciência da Informação como a mineração de dados, o aprendizado de máquinas e os modelos de linguagem ampla (*large language models* – LLM). Essas inovações levaram à criação de um novo campo disciplinar, denominado Ciência de Dados Populacionais [16,17], que, por meio de organização, integração, vinculação e análise de dados individuais e contextuais, objetiva gerar evidências ao nível populacional com valor para a sociedade. Os artigos sobre o desenvolvimento ou aplicação de técnicas de *record linkage* são publicados em CSP desde os anos 2000. Recentemente, com a maior disseminação na Saúde Coletiva das técnicas desenvolvidas pela Ciência da Informação, foram publicados artigos usando mineração de dados, texto e aprendizado de máquina.

Além das questões técnicas, a Ciência de Dados Populacionais busca modelos de gestão de acesso à informação que equilibrem o direito à proteção de informações pessoais com os benefícios potenciais da utilização das bases administrativas na pesquisa, tema tratado em mais de um artigo publicado em CSP [18,19,20].

Ao longo de 40 anos, CSP publicou artigos abordando os principais tópicos da Ciência de Dados Populacionais, valorizando as boas práticas no emprego de dados secundários em pesquisas de interesse para a sociedade. Coerentemente com sua missão, mostrou-se um veículo crucial para a circulação de ideias e métodos desse campo.

Informação adicional

ORCID: Cláudia Medina Coeli (0000-0003-1757-3940).

1. McCracken H. TIME's Machine of the Year, 30 years later. https://techland.time.com/2013/01/04/times-machine-of-the-year-30-years-later (accessed on 09/May/2024).
2. Redator Rock Content. Conheça a história da Internet, sua finalidade e qual o cenário atual. https://rockcontent.com/br/blog/historia-da-internet/ (accessed on 09/May/2024).
3. Boslaugh S. Secondary data sources for public health: a practical guide. Cambridge: Cambridge University Press; 2007.
4. Coeli CM, Pinheiro RS, Camargo Jr. KR. Conquistas e desafios para o emprego das técnicas de record linkage na pesquisa e avaliação em saúde no Brasil. Epidemiol Serv Saúde 2015; 24:795-802.
5. Ministério da Saúde. Departamento de Informática do SUS. Trajetória 1991-2002. Brasília: Ministério da Saúde; 2002.
6. 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.
7. 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: Editora Fiocruz; 2022.
8. Christen P, Schnell R. Thirty-three myths and misconceptions about population data: from data capture and processing to linkage. Int J Popul Data Sci 2023; 8:2115.
9. Coeli CM. A qualidade do linkage de dados precisa de mais atenção. Cad Saúde Pública 2015; 31:1349-50.
10. Coeli CM, Pinheiro RS, Carvalho MS. Neither better nor worse, simply different. Cad Saúde Pública 2014; 30:1363-5.
11. Camargo Jr. KR, Coeli CM. Reclink: aplicativo para o relacionamento de bases de dados, implementando o método probabilistic record linkage. Cad Saúde Pública 2000; 16:439-47.
12. Camargo Jr. KR, Coeli CM. Going open source: some lessons learned from the development of OpenRecLink. Cad Saúde Pública 2015; 31:257-63.
13. Brustulin R, Marson PG. Inclusão de etapa de pós-processamento determinístico para o aumento de performance do relacionamento (linkage) probabilístico. Cad Saúde Pública 2018; 34:e00088117.
14. Saldanha RF, Bastos RR, Barcellos C. Microdatasus: pacote para download e pré-processamento de microdados do Departamento de Informática do SUS (DATASUS). Cad Saúde Pública 2019; 35:e00032419.
15. Carvalho MS. Aberto, por quê? Cad Saúde Pública 2015; 31:221-2.
16. McGrail K, Jones K, Akbari A, Bennett T, Boyd A, Carinci F, et al. A position statement on population data science: the science of data about people. Int J Popul Data Sci 2018; 3:415.
17. Coeli CM. Ciência de dados populacionais. Epidemiol Serv Saúde 2022; 31:e2022119.
18. Ventura M. Lei de acesso à informação, privacidade e a pesquisa em saúde. Cad Saúde Pública 2013; 29:636-8.
19. Ventura M, Coeli CM. Para além da privacidade: direito à informação na saúde, proteção de dados pessoais e governança. Cad Saúde Pública 2018; 34:e00106818.
20. Keinert TMM, Cortizo CT. Dimensões da privacidade das informações em saúde. Cad Saúde Pública 2018; 34:e00039417.

Recebido em 10/Mai/2024
Aprovado em 13/Mai/2024