CSP CADERNOS DE SAÚDE PÚBLICA
REPORTS IN PUBLIC HEALTH

De los datos secundarios a la Ciencia de Datos Poblacional: recordando los 40 años de producción científica en CSP

Cláudia Medina Coeli [1]

doi: 10.1590/0102-311XES087624

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

Me complació mucho la invitación a escribir este editorial. Esta es una gran oportunidad para celebrar junto a Marilia Sá Carvalho, a Luciana Dias de Lima, a Luciana Correia Alves y a toda la comunidad CSP el 40º aniversario de este importante proyecto editorial, en el que tuve el honor de participar durante nueve años como Coeditora en Jefe. Teniendo como foco principal estudios sobre el desarrollo de técnicas y el uso de bases de datos secundarias, revisar la producción científica en esta temática en CSP me permitió recordar artículos fundamentales en mi formación y desarrollo de mis proyectos de investigación.

La primera edición de CSP se publicó en 1985. En el contexto internacional, cobraba impulso la venta de computadoras personales (PC) [1], seguida de la apertura al público del acceso a la World Wide Web (WWW) a principios de los 1990 [2]. Estos avances resultaron en la popularización de las tecnologías de la información.

Las bases administrativas comenzaron a utilizarse como fuentes de datos secundarias en la investigación en Salud Pública [3]. En los años 1990, y en la primera década de los 2000, se implementaron Centros de Datos en Australia, en Canadá y en Reino Unido. En estas organizaciones, las bases administrativas están continuamente vinculadas, y los conjuntos de datos anonimizados resultantes pueden ser accedidos por los investigadores para desarrollar sus proyectos [4].

En 1991 Brasil creó el Departamento de Informática del Sistema Único de Salud (DATASUS) [5], lo que implicó un gran aporte a la accesibilidad en las bases administrativas brasileñas. Pero adoptó un modelo para la difusión de datos distinto del Centro de Datos mencionado anteriormente. En el modelo estaban habilitadas dos modalidades de acceso. La primera se daba por un tabulador en línea que permitía crear tablas de los principales Sistemas de Información Sanitaria nacionales. La segunda consistía en la difusión de microdatos no identificados. Al principio, las bases se distribuían en discos compactos (CD) mensuales y, posteriormente, se pusieron a disposición para descarga en línea. Los datos de nacimientos, defunciones, enfermedades y enfermedades de declaración obligatoria, atención primaria, atención ambulatoria y hospitalaria, centros de salud y presupuesto público comenzaron a ponerse a disposición no solo a los investigadores, sino también a toda la población. Ese modelo de Datos Abiertos es innovador por la variedad de datos, la cobertura temporal y territorial de bases de datos y el acceso inclusivo. La información en formato

digital en torno a la salud también fue puesta a disposición por varias instituciones como el Instituto Brasileño de Geografía y Estadística (IBGE), la Agencia Nacional de Salud Suplementaria (ANS), la Agencia Nacional de Vigilancia Sanitaria (Anvisa), además de los departamentos de salud de estados y municipios.

Incluso antes de la difusión digital, los datos administrativos, especialmente sobre mortalidad, se utilizaban en Brasil en los estudios de Salud Pública. Sin embargo, el acceso proporcionado por la adhesión de las instituciones brasileñas al modelo de datos abiertos fomentó este tipo de uso. En una búsqueda en PubMed, encontré 461 artículos publicados en CSP que utilizaban datos administrativos; de los cuales 86 abordaban temas relacionados con la calidad. Entre estos últimos, se destaca el artículo que es resultado de la tesis de Claudia Risso de Araujo Lima [6]. Claudia, ex miembro del equipo de DATASUS, fue una de las responsables de implementar la política de difusión de información de salud en Brasil. Su artículo publicado en 2009 sigue siendo un referente (96 citas en la base de datos Scopus). Su gran aporte fue realizar una revisión de las dimensiones de calidad en la evaluación de los sistemas de información de salud en Brasil.

La publicación de artículos que evalúan la calidad de sistemas de información o de procesos de vinculación de bases de datos responde a una creciente demanda de adopción de buenas prácticas en la realización y reporte de estudios que utilizan datos secundarios [7,8]. Un editorial [9] y un artículo de perspectivas [10] refuerzan la política editorial de CSP de valorar el uso responsable de las bases administrativas en la investigación.

CSP también publicó cuatro artículos metodológicos que presentan rutinas computacionales para el procesamiento de bases de datos. Tres soluciones estaban dirigidas a la vinculación de registros [11,12,13], mientras que la última, el paquete Microdatasus [14], optimiza la descarga y el preprocesamiento de los microdatos puestos a disposición por DATASUS. El software Reclink se lanzó en 2000 como software libre, pero de código cerrado [11]. La nueva versión de OpenReclink se publicó en 2015, ya con código abierto [12]. EPPD [13] y Microdatasus [14] también son de código abierto en cumplimiento con la política editorial de CSP de adhesión a la ciencia abierta [15].

En estos 40 años ha habido una creciente expansión de las tecnologías de la información. Los avances en la capacidad de capturar, procesar, almacenar, transmitir y analizar datos se produjeron sucesivamente, con avances en cada área estimulando el progreso en las demás. Actualmente, es posible procesar grandes cantidades de información en tiempo real. Los datos no estructurados que presentan diferentes formatos, como textos en documentos o redes sociales, imágenes y salidas de sensores, son nuevas fuentes de uso secundario en la investigación. Además, existe la introducción en la investigación en salud de técnicas desarrolladas por la Ciencia de la Información como la minería de datos, el aprendizaje automático y los modelos lingüísticos de gran tamaño (*large language models* -LLM). Estas innovaciones posibilitaron la creación de un nuevo campo, la Ciencia de Datos Poblacionales [16,17], que mediante la organización, integración, vinculación y análisis de datos individuales y contextuales pretende generar evidencia a nivel poblacional con valor para la sociedad. Los artículos sobre el desarrollo o la aplicación de técnicas de vinculación de registros se han publicado en CSP desde los años 2000. Con la recién difusión en Salud Pública de las técnicas desarrolladas por la Ciencia de la Información, se han publicado artículos utilizando minería de datos, texto y aprendizaje automático.

Además de las cuestiones técnicas, la Ciencia de Datos Poblacionales busca modelos de gestión de acceso a la información que equilibren el derecho a la protección de la información personal con los potenciales beneficios para la sociedad del uso de bases administrativas en la investigación, un tema abordado en más de un artículo ya publicado en CSP [18,19,20].

A lo largo de estos 40 años CSP ha publicado artículos que abordan los principales temas de la Ciencia de Datos Poblacionales, valorando las buenas prácticas con el uso de datos secundarios en la investigación de interés para la sociedad. En consonancia con su misión, demostró ser una herramienta esencial en la circulación de ideas y métodos en este campo.

Información 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.

Recibido el 10/May/2024
Aprobado el 13/May/2024