Acta Biomed 2023; Vol. 94, Supplement 3: e2023121 DOI: 10.23750/abm.v94iS3.14407

Commentary

The potential of digital health records for public health research, policy, and practice: the case of the Lombardy region data warehouse

Lorenzo Blandi[1,2,3], *Alessandro Amorosi*[2], *Olivia Leoni*[2], *Timo Clemens*[3], *Helmut Brand*[3], *Anna Odone*[1]

[1]Department of Public Health, Experimental and Forensic Medicine, University of Pavia, Pavia, Italy; [2]Welfare General Directorate, Regione Lombardia, Milan, Italy; [3]Department of International Health, CAPHRI School for Public Health and Primary Care, Maastricht University, Maastricht, The Netherlands

Abstract. Digital health records can provide advantages to healthcare practice, policy, and research. Several countries have established population-based digitalised data collection, integrated through data linkage techniques. In Lombardy (Italy), a regional population-based registry was established in the 2000s. It collects data from the social and health sector, anonymised immediately after their acquisition and restructured in a single repository. Data can be used for public health interest, planning, monitoring, services evaluation, and research. Indeed, data can also be provided to universities and other scientific institutes. The availability of such data enables to explore the epidemiology of infectious, chronic, and rare diseases. Thus, epidemiological research can support policymakers to tackle public health threats. However, analysis of electronic health records comes along with several challenges, including data inaccuracy, incompleteness, and biases. Researchers should take into consideration limits and barriers related to quality of data. Moreover, health data use must adhere to the national and European privacy legislation, at times limiting the potential of data integration. Therefore, even if big data drives innovation and scientific knowledge, ethical issues regarding privacy should be considered in public debate. (www.actabiomedica.it)

Key words: Big data, data linkage, privacy, data quality, electronic health records

Health systems' digitalisation has great potential to improve healthcare and positively impact on patients' outcomes (1). Analysis of digitalised health data can provide substantial benefits to healthcare practice, monitoring and evaluation, policy, and research (2). With reference to the latter, linked data can be used to supplement follow-up in conventional cohort studies or trials, or to generate real world evidence by creating population-level electronic cohorts that are entirely derived from administrative data (3). Thus, record linkage, a set of techniques and a powerful tool for collecting data from different database, facilitates the conduction of epidemiological studies without specific expensive investments (4). Several countries have established population-based digitalised data collection and linkage as new resources for health services research (5–10).

The regional data warehouse in Lombardy

In Lombardy, the most populated Italian region with 10 million inhabitants, a population-based registry, known as *Regional Data Warehouse* (DWH), was established in the 2000s. The DWH is a supportive tool for the Regional Health Service and for

its institutional functions of protecting the health of citizens; indeed, it allows to carry out monitoring and evaluation activities about effectiveness of the health treatments provided, assessment of the appropriateness and quality of assistance, assessment of satisfaction of the user, assessment of risk factors for health. The DWH collects data from 188 different information flows of the social and health sector. Some of these flows have been active since 1981. At the back-end side of the data warehouse, once data is received from the healthcare organizations in Lombardy, the Region assigns each individual a unique code – through a pseudonymisation technique – for the purpose of verifying the non-duplication of information and any interconnection with other health database. However, that does not allow direct identification of individuals during the data processing. In particular, health records are anonymised immediately after their acquisition by the Region. Then, data are extracted, pre-processed, integrated and restructured in a single repository. At the front-end -the user interface -, only anonymised data can be requested to protect the privacy of individuals.

As front-end terminal, the DWH offers high-performance access to the database resulting from the integration of those non-homogenous sources (11). Data can be used for public health interest, planning, monitoring, services evaluation, and operational research. For this purpose, after an additional process of anonymisation, Lombardy provides data for universities and other scientific institutes which elaborate studies and projects approved by the regional administration. The interested institution, which wants to partner with the Lombardy Region free of charge, must present a formal request of accreditation to the Welfare General Director of the Lombardy Region. After acceptance, a project proposal will be examined by a Commission. Once the accreditation and the project proposal are accepted, the institution and Lombardy Region sign an agreement that states the aims of partnership, the duties of each sides involved, and the duration of the partnership based on the needs of the project. Then, the Lombardy Region sets up the virtual environment to access the DWH. Lombardy provides an explanation of the step-by-step process on its official website, including all the templates to fill out (12).

Health data integration: a powerful tool

The DWH shares a single framework for heterogenous sources and includes datasets from the following areas: hospital care, infectious diseases surveillance, immunization registry, primary care services, mental health services, drug abuse services, pharma prescriptions, emergency system, residential and semi-residential services. The DWH uses a *star schema*, containing a set of large central repositories. These repositories contain the main core of data, without any redundancy, and a set of complementary repositories, one for each information dimension (11). Thus, health data integration enables a collaborative use of information across different systems and actors (13). As a result, DWH can be used to elaborate epidemiological measures and trends of chronic and infectious diseases, as well as to report about preventive interventions, diagnosis, and therapies. Data integration and linkage enables also to develop algorithm and to find a tracking variable which can identify cohorts of patients with similar conditions or diagnosis. This approach allows to develop longitudinal studies on real world data which help us to analyse large sample sizes or whole populations and consider risk factors and outcomes (3). The availability of such data enables not only to explore the epidemiology of infectious diseases and acute conditions, but also of multifactorial and slow onset chronic (14) and rare diseases. Epidemiological research conducted with DWH can inform health services planning, implementation and monitoring and might support policymakers tackling key public health challenges (15).

Data quality issues

Many relevant studies have been conducted, using data from the Lombardy DHW (16–20). Indeed, secondary reuse of electronic health data for research is increasing in importance and popularity (21). Electronic health data become more available and analytic methods become more powerful. However, analysis of electronic health records (EHR) still comes along with several social and technical challenges, including data inaccuracy, incompleteness, and biases implicit in the healthcare recording process (22,23), and data

quality analysis on DWH data has not been undertaken. In fact, data might be inaccurate, resulting in a loss of predictive power; the extent and bias of the noise could require different methods for data analysis (21). Researchers should take into considerations limits and barriers of these data and control potential errors and biases, also using validated algorithm from the national and international literature or from good practices (24).

Privacy issues

According to the national and regional laws, personal data processing must take place in compliance with the rights and fundamental liberties of the individuals. Thus, the reuse of data is accomplished for public interest purposes after the process of data anonymisation (25). However, health data integration must adhere to further national and European laws and must pursue the following statements: 1) the *principle of knowability*, whereby everyone has the right to know the existence of automated decision-making processes of own concern and to receive significant information on the logic used; 2) the *principle of non-exclusivity of the algorithmic decision*: the person who is the recipient of the legal effects of an automated decision has the right that the same is not based solely on an automated process; 3) The *principle of algorithmic non-discrimination*, according to which the data controller should use appropriate mathematical or statistical procedures for profiling, implementing appropriate technical and organizational measures to ensure that risk of errors is minimized and to guarantee the security of personal data and which prevents discriminatory effects (26,27). Hence, it emerges a privacy issue when personal data are used to stratify populations and to profile individuals through algorithms (26). Thus, privacy legislation has arisen barriers in the DHW use, even for public interest purposes. On one hand it is imperative to avoid any generalisation about those limitations and correctly analyse every specific situation of personal data processing. On the other hand, any project aimed at developing predictive approach seems to be precluded by the current laws, limiting the scientific efforts towards *proactive medicine*. These activities constitute a delicate ethical issue that requires a specific regulatory framework adopted by the national legislator. Before the processing of personal data, Lombardy Region carries out a privacy impact assessment. Indeed, it may present a high risk to the rights and freedoms of individuals. In this regard, the national authority pointed out that this requirement was not waived by the emergency regulations adopted with reference to the pandemic context (28).

What's next for the future?

Amount of data from different sources is likely to increase over the coming years, together with demand for access to high-quality linked data. All governmental organizations, universities, research institutes and other stakeholders must collaborate to use the powerful potentialities of data integration. The digital environment of the future needs to meet the expectations of data owners and data users, providing an easy-to-use and safe tool. By increasing the number of data users, new technical advances will permit to share techniques and validated algorithm. Thus, findings from these data will be of higher quality and more comparable between different research. Also, the legal framework should be updated. The Italian Privacy Guarantor suggested the national legislator to pay attention to the following concerns and considerations: i) ensuring specific information for interested individuals and the right to obtain human intervention, ii) to express one's opinion, iii) to obtain an explanation of the decision reached after assessment and to contest the decision (26). Meanwhile, big data is leading to innovation and improving the scientific knowledge for public interest purpose. However, these data could play an ever-growing importance for future regional strategies and plans, especially for diseases and conditions with uncertain epidemiology (14). These data also need to be analysed by ever-powerful tools. For instance, the Artificial Intelligence techniques can make decisions and predictions much earlier and more accurately than humans would, especially with huge amount of unstructured data (29), but the current legal framework does not permit its implementation. Privacy and ethical issues should be more relevant in the public debate, to

understand how to use them and to enhance our data resources with respect for the individual rights.

Conflicts of Interest: Each author declares that he or she has no commercial associations (e.g. consultancies, stock ownership, equity interest, patent/licensing arrangement etc.) that might pose a conflict of interest in connection with the submitted article.

Authors Contribution: All authors contributed to the project and the study design. All authors contributed to subsequent revisions and editing. All authors read and approved the final manuscript.

Ethic Committee: not applicable

References

1. Shull JG. Digital Health and the State of Interoperable Electronic Health Records. JMIR Med Inform. 2019 Nov 1;7(4):e12712. doi: 10.2196/12712.
2. Wang SJ, Middleton B, Prosser LA, et al. A cost-benefit analysis of electronic medical records in primary care. Am J Med. 2003 Apr 1;114(5):397–403. doi: 10.1016/s0002-9343(03)00057-3.
3. Harron K. Data linkage in medical research. BMJ Med. 2022 Mar 1;1(1):e000087. doi: 10.1136/bmjmed-2021-000087.
4. Silva MEM da, Coeli CM, Ventura M, et al. Informed consent for record linkage: a systematic review. J Med Ethics. 2012 Oct 1;38(10):639–42. doi: 10.1136/medethics-2011-100208.
5. Mourby M, Doidge J, Jones KH, et al. Health Data Linkage for Public Interest Research in the UK: Key Obstacles and Solutions. Int J Popul Data Sci. 2019 Apr 2;4(1). doi: 10.23889/ijpds.v4i1.1093.
6. Tully MP, Bernsten C, Aitken M, Vass C. Public preferences regarding data linkage for research: A discrete choice experiment comparing Scotland and Sweden. BMC Med Inform Decis Mak. 2020 Jun 16;20(1):1–13. doi: 10.1186/s12911-020-01139-5.
7. Thorvaldsen G, Andersen T, Sommerseth HL. Record Linkage in the Historical Population Register for Norway. In: Bloothooft, G., Christen, P., Mandemakers, K., Schraagen, M. Population Reconstruction, pp. 155–171. Springer, Cham; 2015.
8. Sortsø C, Caspar Thygesen L, Brønnum-hansen H. Database on Danish population-based registers for public health and welfare research. Scand J Public Health. 2011 Jul;39(7):17–9. doi: 10.1177/1403494811399171.
9. Smith M, Flack F. Data Linkage in Australia: The First 50 Years. Int J Environ Res Public Health. 2021 Nov 1;18(21):11339. doi: 10.3390/ijerph182111339.
10. Domhoff D, Seibert K, Stiefler S, Wolf-Ostermann K, Peschke D. Data linkage of German statutory health insurance claims data and care needs assessments preceding a population-based cohort study on nursing home admission. 2022 Jun 30;12(6):e063475. doi: 10.1136/bmjopen-2022-063475.
11. Cesana G, Fornari C, Chiodini V, Madotto F, Merlino L, Mantovani LG. DENALI: il Data Warehouse di Sanità Pubblica della Regione Lombardia. Farmeconomia Health economics and therapeutic pathways. 2011 May 15;12(2S):19–23. doi: 10.7175/fe.v12i2S.991
12. Emergenza Covid-19 - Indicazioni per l'accesso al patrimonio informativo regionale tramite Daas 2.0 [Internet]. [cited 2023 Feb 10]. Available from: https://www.regione.lombardia.it/wps/portal/istituzionale/HP/DettaglioServizio/servizi-e-informazioni/Enti-e-Operatori/sistema-welfare/Accreditamento/accesso-db-covid/accesso-db-covid
13. Peng C, Goswami P, Bai G. A literature review of current technologies on health data integration for patient-centered health management. Health Informatics J. 2020 Sep 1;26(3):1926–51. doi: 10.1177/1460458219892387.
14. Krysinska K, Sachdev PS, Breitner J, Kivipelto M, Kukull W, Brodaty H. Dementia registries around the globe and their applications: A systematic review. Alzheimers Dement. 2017 Sep 1;13(9):1031–47. doi: 10.1016/j.jalz.2017.04.005.
15. DGR XI/6793 - Fondo per l'Alzheimer e le Demenze di cui alla legge 30 dicembre 2020, n. 178: piano triennale delle attività di Regione Lombardia [Internet]. [cited 2023 Feb 10]. Available from: https://www.lombardianotizie.online/demenze-centri-cura/
16. Corrao G, Franchi M, Cereda D, et al. Persistence of protection against SARS-CoV-2 clinical outcomes up to 9 months since vaccine completion: a retrospective observational analysis in Lombardy, Italy. Lancet Infect Dis. 2022 May 1;22(5):649. doi: 10.1016/S1473-3099(21)00813-6.
17. Mancia G, Rea F, Ludergnani M, Apolone G, Corrao G. Renin–Angiotensin–Aldosterone System Blockers and the Risk of Covid-19. N Engl J Med. 2020 Jun 18;382(25):2431–40. doi: 10.1056/NEJMoa2006923.
18. Merlo I, Crea M, Berta P, et al. Detecting early signals of COVID-19 outbreaks in 2020 in small areas by monitoring healthcare utilisation databases: first lessons learned from the Italian Alert_CoV project. Euro Surveill. 2023 Jan 5;28(1):2200366. doi: 10.2807/1560-7917.ES.2023.28.1.2200366.
19. Grasselli G, Zanella A, Carlesso E, et al. Association of COVID-19 Vaccinations With Intensive Care Unit Admissions and Outcome of Critically Ill Patients With COVID-19 Pneumonia in Lombardy, Italy. JAMA Netw Open. 2022 Oct 3;5(10):e2238871. doi: 10.1001/jamanetworkopen.2022.38871.
20. Corrao G, Franchi M, Cereda D, et al. Factors associated with severe or fatal clinical manifestations of SARS-CoV-2

infection after receiving the third dose of vaccine. J Intern Med. 2022 Nov 1;292(5):829–36. doi: 10.1111/joim.13551.
21. Ta CN, Weng C. Detecting Systemic Data Quality Issues in Electronic Health Records. Stud Health Technol Inform. 2019 Aug 8;264:383. doi: 10.3233/SHTI190248.
22. Kharrazi H, Chi W, Chang HY, et al. Comparing Population-based Risk-stratification Model Performance Using Demographic, Diagnosis and Medication Data Extracted From Outpatient Electronic Health Records Versus Administrative Claims. Med Care. 2017 Aug;55(8):789–96. doi: 10.1097/MLR.0000000000000754.
23. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc. 2013 Jan 1;20(1):117–21. doi: 10.1136/amiajnl-2012-001145.
24. Tavolo per il monitoraggio del recepimento e implementazione del Piano Nazionale Demenze. Linee di indirizzo Nazionali sull'uso dei Sistemi Informativi per caratterizzare il fenomeno delle demenze [Internet]. 2017 [cited 2023 Feb 10]. Available from: https://www.salute.gov.it/imgs/C_17_pagineAree_4893_listaFile_itemName_1_file.pdf
25. Banca dati del Consiglio Regionale della Lombardia [Internet]. [cited 2023 Feb 10]. Available from: https://normelombardia.consiglio.regione.lombardia.it/NormeLombardia/Accessibile/main.aspx?iddoc=rr002012122400003&view=showdoc
26. Parere al Consiglio di Stato sulle nuove modalità di ripartizione del... - Garante Privacy [Internet]. [cited 2023 Feb 10]. Available from: https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9304455
27. Regulation (EU) 2016/679 of the European Parliament and of the Council [Internet]. [cited 2023 Feb 10]. Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN
28. Provvedimento correttivo e sanzionatorio nei confronti dell'Azienda... - Garante Privacy [Internet]. [cited 2023 Feb 10]. Available from: https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9845312
29. Troisi E. Automated Decision Making and right to explanation. The right of access as ex post information. EJPLT. 2022 Sep 13;181-202. doi: 10.57230/EJPLT221ET.

Correspondence:
Received: 13 February 2023
Accepted: 13 March 2023
Lorenzo Blandi, MD
Department of Public Health,
Experimental and Forensic Medicine,
University of Pavia, Pavia, Italy
Via Forlanini 2, Pavia, 27100 Italy
E-mail: lorenzo.blandi@unipv.it