HHS Public Access

Author manuscript

Lancet Digit Health. Author manuscript; available in PMC 2023 May 03.

Published in final edited form as:

Lancet Digit Health. 2023 May ; 5(5): e288–e294. doi:10.1016/S2589-7500(23)00025-0.

The impact of commercial health datasets on medical research and health-care algorithms

Isabelle Rose I Alberto, BS,

Nicole Rose I Alberto, BS,

Arnab K Ghosh, MD MSc,

Bhav Jain, BS,

Shruti Jayakumar, MBBS,

Nicole Martinez-Martin, JD PhD,

Ned McCague, MPH,

Dana Moukheiber, MS,

Lama Moukheiber, MS,

Mira Moukheiber, MS,

Sulaiman Moukheiber, BS,

Antonio Yaghy, MD,

Andrew Zhang, AB,

Leo Anthony Celi, MD MS

College of Medicine, University of the Philippines, Manila, Philippines (I R I Alberto BS, N R I Alberto BS); Department of Medicine, Weill Cornell Medical College, Cornell University, New York, NY, USA (A K Ghosh MD MSc); Institute for Medical Engineering and Science (B Jain BS, N McCague MPH, D Moukheiber MS, L Moukheiber MS, A Yaghy MD, A Zhang AB, L A Celi MD MS) and The Picower Institute for Learning and Memory (M Moukheiber MS), Massachusetts Institute of Technology, Cambridge, MA, USA; Deloitte Consulting, London, UK (S Jayakumar MBBS); Stanford Center for Biomedical Ethics, Stanford Medicine, Stanford, CA, USA (N Martinez-Martin JD PhD); Markforged, Watertown, MA, USA (N McCague); Department of Computer Science, Worcester Polytechnic Institute, Worcester, MA, USA (S Moukheiber BS); New England Eye Center, Tufts University Medical Center, Boston, MA, USA (A Yaghy); Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA, USA (A Zhang AB); Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA (L A Celi); Department of Biostatistics, Harvard T H Chan School of Public Health, Boston, MA, USA (L A Celi)

Abstract

Correspondence to: Prof Leo Anthony Celi, Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA, lceli@bidmc.harvard.edu.
All authors contributed equally to this manuscript.

Declaration of interests
We declare no competing interests.

As the health-care industry emerges into a new era of digital health driven by cloud data storage, distributed computing, and machine learning, health-care data have become a premium commodity with value for private and public entities. Current frameworks of health data collection and distribution, whether from industry, academia, or government institutions, are imperfect and do not allow researchers to leverage the full potential of downstream analytical efforts. In this Health Policy paper, we review the current landscape of commercial health data vendors, with special emphasis on the sources of their data, challenges associated with data reproducibility and generalisability, and ethical considerations for data vending. We argue for sustainable approaches to curating open-source health data to enable global populations to be included in the biomedical research community. However, to fully implement these approaches, key stakeholders should come together to make health-care datasets increasingly accessible, inclusive, and representative, while balancing the privacy and rights of individuals whose data are being collected.

Introduction

With the increasing digitisation of medical records, the amount of health data produced by medical institutions has grown exponentially.[1,2] Coupled with the advent of cloud data storage, distributed computing, and machine learning, we have entered a new, digital-first era in medicine that has the potential to use huge volumes of health-care data to accelerate scientific discovery, improve health-care quality, enable personalised medicine, and inform evidence-based policy making.[1,3–5] We have also seen an increase in researchers working to combine multiple data sources to identify vulnerable groups, quantify inequities in care, and explore the effect of the social determinants of health.[6,7]

This confluence of factors has turned health-care data—including patient demographics, clinical examination results, laboratory findings, and genomic data—into a premium commodity possessing value for private and public entities. The need for large amounts of health-care data in both academia and private industry has brought about an aggressive and lucrative push towards the commercialisation, spawning a multibillion US$ industry that centres on collecting, analysing, and selling these data.[2,6,8–10]

Although essential to the world of drug development and medical devices, these commercially available datasets might be improperly used by academic researchers who are probably accustomed to publicly available datasets (eg, claims data from the US Centers for Medicare and Medicaid Services[11]) and local datasets from their own institutions and studies. The implications of this knowledge and experience gap are non-trivial—academic research directly influences the development of health-care tools, decisions, and policies that have consequences for patients. Commercial datasets might warrant special caution regarding their implications for research reproducibility and generalisability because these datasets are often expensive to access, and thus are access-restricted. Clinicians, patients, researchers, and policy makers need to understand the broader market for biomedical data, including how the datasets are processed and curated, to better inform how they might be used. In this Health Policy paper, we highlight the need for caution in the use of these data in the growing field of health algorithm development, especially when artificial intelligence (AI) or machine learning is involved.

Sources of commercial health data

There are many sources from which data and specimens used for secondary research are procured.[8] These include patient registries, health-care databases including electronic health records (EHRs), pharmacy and health insurance databases, social media, and patient-powered research networks.[12] The data from these sources are then deidentified in accordance with Health Insurance Portability and Accountability Act (HIPAA) standards, and can be sold thereafter without public transparency or patient consent.[9]

The commercialisation of deidentified data is not restricted to covered entities, like physicians or medical institutions, as their business associates can also partake in the deidentification and sale of health-care data, as long as their contractual agreement expressly allows them to do so.[9] These practices have resulted in many commercial health datasets becoming available to researchers and companies (panel). However, most biomedical research these commercial datasets enable is beyond the original routine clinical care and primary research applications the data and biospecimens were collected for.[8] Patient-centred and business-centred applications of health data are closely linked,[13] and therefore, several commercial health datasets are marketed to private industry for industry-related purposes.[14,15] Due to the opaque nature of health dataset curation and data harmonisation, researchers might not be aware of the sources of bias derived from commercial datasets, nor how best to address these sources of bias and associated limitations with their analyses. These datasets differ from open data sources (eg, Medicare and Medicaid claims), for which decisions regarding curation are made publicly available and have been extensively examined for bias.[16] Such practices mean that commercial health datasets might not be ideally suited for all research applications; researchers using these datasets for academic purposes should be aware of these limitations. The growing use of such datasets in peer-reviewed research might lead to biased results if not evaluated systematically.[8] Therefore, researchers should become familiar with the strengths and limitations of both commercial and non-commercial datasets, to reduce the risk of obtaining incorrect or biased results.

Effect of health data vending on biomedical research: challenges with reproducibility and generalisability

The introduction of commercial health datasets into biomedical research has compounded the problem of reproducibility—a key tenet of rigorous scientific research. Reproducibility is often impeded by the unavailability of the original dataset and code by which the data are processed. Also, commercial databases generally restrict access to data that are essential to reproducing published results.[17] As a result, these financial barriers cause additional hardships that make the pursuit of scientific inquiry structurally inequitable for equally qualified researchers without the necessary financial resources to obtain these datasets. Restricted-access datasets also pose challenges for routine reproducibility audits when journals evaluate studies before publication.[18] For example, in a study designed to validate a prediction model of inflammatory bowel disease, the authors noted that the original model was developed with US Veterans Health Administration data, which are inaccessible to those outside the Veterans Affairs system.[19,20] When choosing an alternative dataset to validate

the model, the authors of the validation study also used a commercial dataset (Optum EHR), which they note is unavailable to the public due to data licensing agreements.[21] As a result, just as with the original model, researchers wishing to examine the validation study must license the data themselves.

There are also inherent limitations in non-public health datasets that hinder the generalisability of health research to a broader population. Proprietary models have been shown to decline in performance over time, but are difficult to assess owing to the scarcity of publicly available data.[22] As an example involving a proprietary health algorithm developed on non-public data, a commonly used sepsis prediction model based on data from Epic EHRs was found to generalise poorly to an external validation set.[23] The Epic Sepsis Model was implemented at hundreds of US hospitals and was found to have poor discrimination and calibration in predicting sepsis onset.[23] The widespread use of this model was not based on thorough independent validation, but rather can be attributed to its ease of integration into existing hospital EHR systems produced by Epic.[23] The data used to produce the model consisted of 405 000 patient encounters from 2013 to 2015, across three hospital systems, making the model susceptible to data drift and other issues in the absence of regular external validation.

Data availability is a crucial issue in machine learning for health research. Due to the complexity and opaque nature of machine learning for health algorithms such as deep learning, it is particularly important for datasets used in machine learning for health research to be tested for bias. Datasets that are homogeneous are likely to yield poorly generalisable models.[24] Moreover, machine learning algorithms can exploit patterns in the training data that can be imperceptible to a human observer but have the potential to negatively influence care delivery.[25] Therefore, third parties must evaluate not only machine learning for health models but also the data they are trained on to establish the suitability of these data and model for a particular application. Despite these concerns, only about 55% of machine learning for health studies use publicly available data—a substantially lower percentage than in machine learning studies related to non-health care.[26] For example, a 2020 study published by Google showing AI application for breast cancer screening[27] was met with criticism for not releasing model architecture parameters and for using proprietary datasets. The authors of the letter accuse Google of the “promotion of a closed technology”.[28] Although in this case access to the codebase was also pointed out as an issue, the point remains that restricted-access datasets have made reproducibility and generalisability in machine learning for health particularly challenging.

The repurposing of data from commercial insurance claims for building health-care models can also be problematic. These claims data, which serve as data sources for many commercial databases, are restricted in scope and are subject to upcoding.[29] In some circumstances, prediction models directly created from medical records are more accurate than models created from claims data possibly because clinical characteristics can be better defined in medical records.[30] Models for predicting the severity of rheumatoid arthritis solely built with claims data, have been shown to exhibit low accuracy among patients with high disease activity.[31] Without the purchase of supplemental databases, many key demographic and outcome details such as smoking status, race, and mortality are

inaccessible, thereby limiting the scope and extent of research that can be done with these core claims datasets.[32] Moreover, many claims databases represent working-age patients covered by private insurance supplied by employers rather than a random sample from the US population. For example, employee health data from large employers are used more frequently in databases than data from small or medium firms. Therefore, although these databases cover a large portion of the US population, they might not capture some groups and thereby yield problematic algorithms.

There is also poor consistency in the structure of the underlying data and definition of data across claims of data providers.[33] Interdatabase discrepancies involving data availability, patient populations, and other characteristics might give rise to inconsistent research results. For example, in a study that assessed the risk of sudden cardiac arrest and ventricular arrhythmia among users of second-generation sulfonylureas, inconsistent results were found with claims from five states' Medicaid claims (1999–2012) and Optum Clinformatics commercial claims (2000–16).[34] These discrepancies were partly attributable to differential data capture across insurance plans, data features being present in one dataset but not the other, and demographic differences between the two patient populations (ie, Optum capturing data only from commercially insured individuals; and Medicaid capturing data from publicly insured adults with low-income, older people [≥65 years old], children [<19 years old], pregnant people, and people with disabilities).[34] The presence of these discrepancies raises issues regarding the breadth of research done with a single commercial dataset, meaning that researchers might be pressured to purchase multiple datasets. Therefore, the suitability of claims data for biomedical research should be carefully evaluated on a case-by-case basis per standardised database selection frameworks. For example, international societies such as the International Society of Pharmaceutical Engineering and International Society for Pharmacoeconomics and Outcomes Research have developed guidelines regarding the analysis of secondary data sources for treatment effectiveness research.[35,36]

Ethical considerations in data vending

Discussions of the ethics of big data projects in medicine often focus on privacy and data protection, particularly in terms of the risks to individuals from their personal health data being revealed or used for non-medical purposes. HIPAA protects personal health information that is identifiable, but does not place restrictions on the use of deidentified health data to support potential gains in scientific knowledge enabled by sharing data. Furthermore, HIPAA is limited to data that are generated directly by health-care providers or business associates (when an agreement is in place). HIPAA might also be bypassed if patients sign waivers when obtaining medical care or other benefits, or as a condition of employment. Together with the fact that advances in technology and accessibility of large public databases have allowed the re-identification of data with increasing ease, concern has grown over the limitations of current regulations for protecting privacy in health data at the individual and group levels.[2,37]

However, privacy is a concept that is context-dependent and often should be considered in trade-off with other societal rights and values. Privacy provisions for health data are

intended to balance privacy with the social good that might come from sharing data.[38] For big data projects in medicine, efforts to address privacy are also often viewed in terms of trade-offs with the need to support innovation and scientific knowledge.[39] For example, some intrusions into privacy are considered permissible if they serve a public health or safety purpose.[40] Individuals report greater willingness to share genomic and health data if these data are used for social benefit.[41] For this reason, assessments of the ethics of big data health projects need to include careful consideration of the public benefit and scientific knowledge gained from such projects in the larger discussion of privacy.

The consent process is an essential part of informing the people from whom data are being collected of privacy and data protection issues. In primary research, the informed consent process is used to notify patients about privacy issues, such as the type and scope of data collection and potential downstream risks from the data. Traditionally, informed consent takes place once for a single study. Single-instance consent, however, might not convey the range of potential privacy implications and downstream uses, which might include repurposing or aggregation of data; considering the complexities of future inferential analyses that are not yet developed; and the downstream effects of their data, such as use by third parties to determine insurance or mortgage rates.[42] Furthermore, when health data are collected through consumer devices, informed consent might not apply, and instead terms and conditions, which are often long, dense, and difficult to read, are the main means for informing consumers of data collection.[43] Nonetheless, much of the justification for proceeding with big data projects rests on the potential social benefits and medical knowledge developed from projects that aggregate and reuse personal data.

Large-scale data collection projects in health-care institutions, such as learning health systems, have prompted efforts to formulate appropriate means to balance privacy and innovation through responsible data governance or stewardship.[44] Such approaches include minimising risks to individuals and groups from uses of their data, and mechanisms for generating input from stakeholders on data collection and use. For these frameworks, use of personal data to produce scientific gains and social benefit is prioritised. As commercial datasets play an increasing role in health research, it becomes imperative to examine the extent to which they contribute to scientific knowledge and the public good. Leveraging key data governance principles, including regularly consulting marginalised populations and indigenous communities regarding the collection and processing of personal health data, applying best practices in data deidentification to protect patient data privacy and reduce breach risks, and periodically revisiting governance mechanisms as new technologies are introduced through engagement with the aforementioned stakeholders, all represent potential means through which societal benefits can be maximised from health data.[45]

Discussion

Clive Humby, a British mathematician, commented that “data is the new oil” in 2006.[46] However, most people neglect the important addendum in the second part of his quip: “[oil is] valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc., to create a valuable entity that drives profitable activity; so must data be broken down, analyzed for it to have value.”[46] This phrase proved to be prescient.

Multibillion dollar tech companies such as Google, Microsoft, Amazon, and Facebook have collected petabytes of data to power and improve their products.[47] In health care, Google has partnered with Meditech, an EHR company, and Oracle has purchased Cerner, an EHR company secondary only to Epic.[48,49] Despite massive troves of data existing in the health-care realm, for them to be useful, they must first be aggregated, processed, and curated. Consequently, health-care databases are a necessity for researchers to develop tools that will improve the quality of care.

With the many growing needs for health data, AI-based and machine learning-driven solutions will form the backbone of precision medicine. Currently, such solutions are largely informed by commercially available datasets created from high-income countries comprised of higher-income populations living near hospitals and with lower barriers to care. By contrast, lower-income countries have fewer capabilities to generate and refine data, a process that often requires advanced infrastructure and expertise. Therefore, a large swath of the global population is excluded from most health-care research and innovations. Ibrahim and colleagues[50] introduced the term data poverty, which they defined as “the inability for individuals, groups, or populations to benefit from a discovery or innovation due to insufficient data that are adequately representative”. Data poverty might have major downstream effects on the quantity and quality of AI-based health-care tools, meaning lower-income populations are less likely to benefit from the rapid technological advances in clinical care.

The task of curating health-care datasets is costly, and a sustainable solution is required to produce them. Commercialisation is one such strategy, but alternative approaches are also being pursued in a range of initiatives from academic and government institutions. With research funding, cooperation of health-care institutions, and support of patients, several groups have created open-access health-care databases. For example, the Massachusetts Institute of Technology Laboratory for Computational Physiology (MIT LCP; Cambridge, MA, USA) has constructed a critical care database called the Medical Information Mart for Intensive Care (MIMIC).[51] In contrast to commercially available datasets, access to MIMIC only requires users to confirm their identity, complete human research training, and sign a data use agreement. MIMIC’s code repository provides a forum for public discussion and code sharing, promoting transparency, reproducibility, and collaboration between research groups. Now in its fourth iteration, MIMIC has enabled more than 4000 studies that leverage electronic health record data from a single, large US hospital.[52] Although MIMIC has proved valuable for research studies, the fact that the data are restricted to a single US hospital limits their value for research. How can the success of MIMIC be reproduced on a grand scale?

Other open-access databases have emerged worldwide, including the AmsterdamUMCdb,[53] the eICU-CRD,[54] and the High Time-Resolution Intensive Care Unit Dataset.[55] As an example of a large, publicly available, multi-institutional EHR dataset, eICU-CRD contains data collected from 208 hospitals across the USA and made possible through a collaboration between Philips Healthcare and MIT LCP.[54] As the value of publicly available research data is increasingly recognised, the US National Institutes of Health (NIH) in October, 2020, announced a new Data Management and Sharing policy, which requires NIH-funded

researchers to share their data publicly.[56] These examples show that an open-access approach to health-care data is a feasible route alongside commercial endeavours. However, like commercial datasets, current open-access datasets are not without issues. They do not necessarily mitigate the problem of generalisability, because many are derived from just one or a few institutions. Truly comprehensive coverage of the global population would require datasets from diverse geographical regions, necessitating a standardised format for data harmonisation. Industry–academia collaborations, convened through partnerships of commercial entities and clinician scientists, are a promising avenue for the creation of large datasets similar to eICU-CRD. The datasets resulting from these collaborations represent a potential solution to the generalisability concerns that relate to data accessibility as outlined. To facilitate these multi-institutional data pipelines, initiatives such as the Observational Health Data Sciences and Informatics programme[57] and the US Food and Drug Administration's Sentinel system[58] have developed guidelines for data quality and standardisation, and software that enables federated data analytics.

There are opportunities to parallel these successes with commercial datasets. To ensure that inherent biases of commercial datasets are well understood, so that researchers can understand the limitations of them, we have two recommendations. First, data providers should provide detailed documentation of the processes and algorithms used to curate and harmonise the data. This provision would allow researchers to develop a full understanding of the data they are working with and anticipate any potential issues that might arise. Second, commercial datasets should be made open to independent third-party assessors to analyse the limitations, validity, and potential issues with their data curation and harmonisation processes. Ideally, this analysis would be done by the researchers themselves, with access to the code and data, who can then interrogate these processes to quantify and describe any sources of bias.

As patients entrust institutions with their health data, they should benefit from high-quality research that advances medical care for all. The curation of health data requires careful consideration of their downstream research uses and their many ethical implications, particularly when these data are used to develop health-care algorithms. Current frameworks of health data collection and distribution, whether from industry, academia, or government institutions, are imperfect and do not allow researchers to use the full potential of downstream analytical efforts. To take advantage of the wealth of data from clinical care, key stakeholders should come together to make health-care datasets increasingly accessible, inclusive, and representative, at the same time balancing the privacy and rights of the patients involved.

Acknowledgments

LAC was funded by the National Institute of Health through the NIBIB R01 grant EB017205.

References

1. Yannoukakou A, Kitsos P, Milossi M, Nikita M. Big and open data privacy risks in health sector: developing a trend or establishing the future? 5th International Conference on E-Democracy, Security, Privacy and Trust in a Digital World; Dec 5–6, 2013.

2. Glenn T, Monteith S. Privacy in the digital world: medical and health data outside of HIPAA protections. Curr Psychiatry Rep 2014; 16: 494. [PubMed: 25218603]

3. Hansen MM, Miron-Shatz T, Lau AYS, Paton C. Big data in science and healthcare: a review of recent literature and perspectives. Yearb Med Inform 9: 21–26. [PubMed: 25123717]

4. Adam NR, Wieder R, Ghosh D. Data science, learning, and applications to biomedical and health sciences. Ann N Y Acad Sci 2017; 1387: 5–11. [PubMed: 28122121]

5. Pastorino R, De Vito C, Migliara G, et al. Benefits and challenges of big data in healthcare: an overview of the European initiatives. Eur J Public Health 2019; 29 (suppl 3): 23–27.

6. Dickens A. From information to valuable asset: the commercialization of health data as a human rights issue. Health Hum Rights 2020; 22: 67–69. [PubMed: 33390695]

7. Kind AJH, Buckingham WR. Making neighborhood-disadvantage metrics accessible—the neighborhood atlas. N Engl J Med 2018; 378: 2456–58. [PubMed: 29949490]

8. Spector-Bagdady K. Governing secondary research use of health data and specimens: the inequitable distribution of regulatory burden between federally funded and industry research. J Law Biosci 2021; 8: lsab008. [PubMed: 34055367]

9. McGraw D, Petersen C. From commercialization to accountability: responsible health data collection, use, and disclosure for the 21st century. Appl Clin Inform 2020; 11: 366–73. [PubMed: 32434225]

10. Tang C, Plasek JM, Bates DW. Rethinking data sharing at the dawn of a health data economy: a viewpoint. J Med Internet Res 2018;20: e11519. [PubMed: 30467103]

11. Mues KE, Liede A, Liu J, et al. Use of the Medicare database in epidemiologic and health services research: a valuable source of real-world evidence on the older and disabled populations in the US. Clin Epidemiol 2017; 9: 267–77. [PubMed: 28533698]

12. Dagenais S, Russo L, Madsen A, Webster J, Becnel L. Use of real-world evidence to drive drug development strategy and inform clinical trial design. Clin Pharmacol Ther 2022; 111: 77–89. [PubMed: 34839524]

13. Trinidad MG, Platt J, Kardia SLR. The public's comfort with sharing health data with third-party commercial companies. Humanit Soc Sci Commun 2020; 7: 149. [PubMed: 34337435]

14. Clinformatics? Data Mart. 2017. https://www.optum.com/content/dam/optum/resources/productSheets/Clinformatics_for_Data_Mart.pdf (accessed July 19, 2022).

15. IBM. IBM MarketScan research databases for life sciences researchers. 2021. https://www.ibm.com/downloads/cas/OWZWJ0QO (accessed July 19, 2022).

16. Government Accountability Office. Medicaid: data completeness and accuracy have improved, though not all standards have been met. 2021. http://resource.nlm.nih.gov/9918250407506676 (accessed Jul 19, 2022).

17. Peng RD, Hicks SC. Reproducible research: a retrospective. Annu Rev Public Health 2021; 42: 79–93. [PubMed: 33467923]

18. National Academy of Sciences. Improving reproducibility in the face of data access restrictions and other data complexities. 2019. www.nationalacademies.org/reproducibilityinscience (accessed July 19, 2022).

19. Waljee AK, Lipson R, Wiitala WL, et al. Predicting hospitalization and outpatient corticosteroid use in inflammatory bowel disease patients using machine learning. Inflamm Bowel Dis 2017; 24: 45–53. [PubMed: 29272474]

20. University of San Francisco California. National Veterans Affairs (VA) Data. 2023. https://peppercenter.ucsf.edu/department-veterans-affairs-va-data (accessed March 6, 2023).

21. Gan RW, Sun D, Tatro AR, et al. Replicating prediction algorithms for hospitalization and corticosteroid use in patients with inflammatory bowel disease. PLoS One 2021; 1: 16.

22. Finlayson SG, Subbaswamy A, Singh K, et al. The clinician and dataset shift in artificial intelligence. N Engl J Med 2021; 385: 283–86. [PubMed: 34260843]

23. Wong A, Otles E, Donnelly JP, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern Med 2021; 181: 1065–70. [PubMed: 34152373]

24. Celi LA, Cellini J, Charpignon M-L, et al. Sources of bias in artificial intelligence that perpetuate healthcare disparities—a global review. PLoS Digit Health 2022; 1: e0000022. [PubMed: 36812532]
25. Adam H, Yang MY, Cato K, et al. Write it like you see it: detectable differences in clinical notes by race lead to differential model recommendations. arXiv 2022; published online May 8. https://arxiv.org/abs/2205.03931v1 (preprint).
26. McDermott MBA, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M. Reproducibility in machine learning for health research: still a ways to go. Sci Transl Med 2021; 13: eabb1655. [PubMed: 33762434]
27. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature 2020; 577: 89–94. [PubMed: 31894144]
28. Haibe-Kains B, Adam GA, Hosny A, et al. Transparency and reproducibility in artificial intelligence. Nature 2020; 586: E14–16. [PubMed: 33057217]
29. Coustasse A, Layton W, Nelson L, Walker V. Upcoding Medicare: is healthcare fraud and abuse increasing? Perspect Health Inf Manag 2021; 18: 1f.
30. Ouwerkerk W, Voors AA, Zwinderman AH. Factors influencing the predictive power of models for predicting mortality and/or heart failure hospitalization in patients with heart failure. JACC Heart Fail 2014; 2: 429–36. [PubMed: 25194294]
31. Sauer BC, Teng CC, Accortt NA, et al. Models solely using claims-based administrative data are poor predictors of rheumatoid arthritis disease activity. Arthritis Res Ther 2017; 19: 86. [PubMed: 28482933]
32. Kulaylat AS, Schaefer EW, Messaris E, Hollenbeak CS. Truven Health Analytics MarketScan databases for clinical research in colon and rectal surgery. Clin Colon Rectal Surg 2019; 32: 54–60. [PubMed: 30647546]
33. Voss EA, Ma Q, Ryan PB. The impact of standardizing the definition of visits on the consistency of multi-database observational health research. BMC Med Res Methodol 2015; 15: 13. [PubMed: 25887092]
34. Dhopeshwarkar N, Brensinger CM, Bilker WB, et al. Risk of sudden cardiac arrest and ventricular arrhythmia with sulfonylureas: an experience with conceptual replication in two independent populations. Sci Rep 2020; 10: 10070. [PubMed: 32572080]
35. Hall GC, Sauer B, Bourke A, Brown JS, Reynolds MW, LoCasale R. Guidelines for good database selection and use in pharmacoepidemiology research. Pharmacoepidemiol Drug Saf 2012; 21: 1–10.
36. Berger ML, Sox H, Willke RJ, et al. Good practices for real-world data studies of treatment and/or comparative effectiveness: recommendations from the Joint ISPOR-ISPE Special Task Force on real-world evidence in health care decision making. Value Health 2017; 20: 1003–08. [PubMed: 28964430]
37. Rocher L, Hendrickx JM, de Montjoye YA. Estimating the success of re-identifications in incomplete datasets using generative models. Nat Commun 2019; 10: 3069. [PubMed: 31337762]
38. Allen AL. Protecting one's own privacy in a big data economy. 2016. https://scholarship.law.upenn.edu/faculty_scholarship/1716 (accessed July 19, 2022).
39. Price WN 2nd, Cohen IG. Privacy in the age of medical big data. Nat Med 2019; 25: 37–43. [PubMed: 30617331]
40. Martinez-Martin N, Wieten S, Magnus D, Cho MK. Digital contact tracing, privacy, and public health. Hastings Cent Rep 2020; 50: 43–46. [PubMed: 32596893]
41. Mello MM, Lieou V, Goodman SN. Clinical trial participants' views of the risks and benefits of data sharing. N Engl J Med 2018; 378: 2202–11. [PubMed: 29874542]
42. Mittelstadt B. Ethics of the health-related internet of things: a narrative review. Ethics Inf Technol 2017; 19: 157–75.
43. Spector-Bagdady K. Reconceptualizing consent for direct-to-consumer health services. Am J Law Med 2015; 41: 568–616. [PubMed: 26863850]
44. Institute of Medicine (US). Digital infrastructure for the learning health system: the foundation for continuous improvement in health and health care: workshop series summary. Washington, DC: National Academies Press (US), 2011.

45. OECD. Health data governance: privacy, monitoring and research—policy brief. 2015. https://www.oecd.org/health/health-systems/Health-Data-Governance-Policy-Brief.pdf (accessed Dec 24, 2022).
46. Arthur C. Tech giants may be huge, but nothing matches big data. The Guardian. Aug 23, 2013. https://www.theguardian.com/technology/2013/aug/23/tech-giants-data (accessed July 19, 2022).
47. Mitchell G. How much data is on the internet? BBC Science Focus. https://www.sciencefocus.com/future-technology/how-much-data-is-on-the-internet/ (accessed July 19, 2022).
48. Lopez E. How Google and MEDITECH are charting a new course in digital healthcare. July 20, 2021. https://blog.meditech.com/howgoogle-and-meditech-are-charting-a-new-course-in-digital-healthcare. (accessed July 19, 2022).
49. Egbert M. Oracle completes acquisition of Cerner. June 7, 2022. https://www.oracle.com/news/announcement/oracle-completes-acquisition-of-cerner-2022-06-07/ (accessed July 19, 2022).
50. Ibrahim H, Liu X, Zariffa N, Morris AD, Denniston AK. Health data poverty: an assailable barrier to equitable digital health care. Lancet Digit Health 2021; 3: e260–65. [PubMed: 33678589]
51. Johnson AE, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016; 3: 160035. [PubMed: 27219127]
52. Johnson AEW, Stone DJ, Celi LA, Pollard TJ. The MIMIC Code Repository: enabling reproducibility in critical care research. J Am Med Inform Assoc 2018; 25: 32–39. [PubMed: 29036464]
53. Thoral PJ, Peppink JM, Driessen RH, et al. Sharing ICU patient data responsibly under the Society of Critical Care Medicine/European Society of Intensive Care Medicine joint data science collaboration: the Amsterdam University Medical Centers Database (AmsterdamUMCdb) example. Crit Care Med 2021; 49: e563–77. [PubMed: 33625129]
54. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data 2018; 5: 180178. [PubMed: 30204154]
55. Hyland SL, Faltys M, Hüser M, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. Nat Med 2020; 26: 364–73. [PubMed: 32152583]
56. Kozlov M. NIH issues a seismic mandate: share data publicly. Nature 2022; 602: 558–59. [PubMed: 35173323]
57. Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform 2015; 216: 574–78. [PubMed: 26262116]
58. Platt R, Brown JS, Robb M, et al. The FDA sentinel initiative—an evolving national resource. N Engl J Med 2018; 379: 2091–93. [PubMed: 30485777]

Panel: Major commercial health datasets and vendors

Optum (UnitedHealth Group)

Optum, a subsidiary of UnitedHealth, offers a database of administrative health insurance claims as part of its Clinformatics Data Mart. The database comprises deidentified medical, pharmacy, and laboratory data collected since 2000 across the USA.

Merative MarketScan (formerly IBM MarketScan)

Merative MarketScan Research Databases comprise deidentified patient-level health data from privately and publicly insured Americans, including prescriptions, laboratory results, and electronic medical records. The dataset covers more than 273 million unique patients since 1995.

Flatiron Enhanced Datamart

Flatiron, which was acquired by Roche in 2018, offers 22 datasets as part of its Enhanced Datamart, with a focus on clinical, genomics, and outcomes data related to oncology.

IQVIA Real World Data

IQVIA, formed from the merger of IMS Health and Quintiles in 2016, offers an expansive portfolio of health data from electronic medical records, prescriptions, insurance claims, and other sources.

Decision Resources Group Real World Evidence

Clarivate Real World Data offers data on over 300 million US patients, including over 120 million claims and electronic health records. These data cover demographics, laboratory results, diagnoses, and more.

Bluehealth Intelligence (Blue Cross Blue Shield)

BlueHealth Intelligence offers a dataset of insurance claims with high geographic coverage and uniformity of structure for over 234 million unique patients.

Symphony Integrated Dataverse

Symphony Health’s Integrated Dataverse connects longitudinal data from physicians, pharmacies, patients, and hospitals collected for 17 years from more than 317 million patients.

Premier Healthcare Database

The Premier Healthcare Database includes hospital admissions data from more than 1000 US health-care providers and 231 million patients. Around three-quarters of the data are classified as originating from urban areas, with the remainder from rural regions.

Cerner Real-World Data

Cerner Real-World Data offers a US national, deidentified, longitudinal patient dataset.

PointClickCare Real World Data

PointClickCare offers more than 20 years of deidentified real world data from long-term and post-acute care, including nursing and assisted living facilities.

Ontada Real World Data

Ontada is a data platform launched by McKesson, which offers patient demographics, diagnostics, treatments, insurance claims, and other data related to oncology care.

Search strategy and selection criteria

For this Health Policy, we searched Google Scholar, PubMed, and Google for studies, commercial press releases, and reports published between Jan 1, 2011, and Dec 31, 2022, with the terms: “open health data”, “health data vendors”, “electronic health records”, “commercial health database”, “medical big data”, “real-world evidence”, “administrative claims data”, “biomedical reproducibility”, “biomedical generalizability”, “biomedical data accessibility”, and “biomedical data privacy”. We also included publications cited in the documents when relevant. We restricted the language of the studies, commercial press releases, and reports to those published in English. We included publications from searches if they presented a concept, principle, case study, or approach that was considered applicable either directly or indirectly to the use of commercial or non-commercial health datasets in biomedical research and health-care algorithms development.