Journal of the Society for Clinical Data Management

Starren J, Rasmussen LV, Schneider DH, Nannapaneni P, Michelson K. A Privacy Nihilist Perspective on Clinical Data Sharing: Open Clinical Data Sharing is Dead, Long Live the Walled Garden. *Journal of the Society for Clinical Data Management*. 2023; 3(3): 4, pp. 1–7. DOI: https://doi.org/10.47912/jscdm.234

OPINION PAPER

A Privacy Nihilist Perspective on Clinical Data Sharing: Open Clinical Data Sharing is Dead, Long Live the Walled Garden

Justin Starren[*], Luke V. Rasmussen[*], Daniel H. Schneider[*], Prasanth Nannapaneni[*,†] and Kelly Michelson[*,‡]

Clinical data sharing, combined with deep learning, and soon quantum computing, has the potential to radically accelerate research, improve health care, and lower costs. Unfortunately, those tools also make it much easier to use the data in ways that can harm patients. This article will argue that the vast amounts of data collected by data brokers, combined with advances in computing, have made reidentification a serious risk for any clinical data that is shared openly. The new National Institute of Health data sharing policy acknowledges this new reality by directing researchers to consider controlled access for any individual-level data. The clinical data sharing community would be well-advised to follow the lead of the physics and astronomy communities and create a "walled garden" approach to data sharing. While the investment will be significant, this approach provides a more optimal combination of both access and privacy. Some design considerations for walled gardens are discussed. The article concludes with a list of recommended actions that can be taken by individuals and institutions today.[1]

Keywords: Manage Clinical Research Data; Define/document data handling process; Secure Data

Introduction

Clinical data sharing combined with deep learning, and soon quantum computing, has the potential to radically accelerate research, improve health care, and lower costs. Unfortunately, those tools also make it much easier to use the data in ways that can harm patients. These competing forces are creating a perilous time for clinical data sharing. Future generations will judge informatics and data science by how we balance those forces.[1]

The term "clinical data sharing" has many potential interpretations. This article focuses on the sharing of clinical data for research, rather than to support clinical care or for public health. However, we will draw from incidents that include clinical care to illuminate the challenges and potential pitfalls involved in data sharing in this sector. This article also focuses on sharing data during "normal times" rather than during a public health emergency.[2] Further, we exclude mandatory data reporting to accrediting or governmental programs,[3] and considerations about knowledge sharing (such as pay-to-publish and open access journals).

The Human Rights Watch has declared data privacy a human right,[4] as has the United Nations High Commissioner for Human Rights.[5] The ethical mandate for health information privacy traces back at least to Hippocrates.[6] Researchers are ethically bound to respect and protect the privacy rights of research participants.[7] Unfortunately, as we will discuss later, data, including clinical data, are frequently used for reasons beyond their original purpose, potentially violating privacy rights. We propose that open clinical data sharing — data that is freely available to any and all users without authentication or repercussions[8] — cannot adequately protect the privacy rights of research participants. We further propose that clinical data sharing for research should limit sharing to known and trusted entities, with severe penalties for data misuse. Doing less risks losing the trust of patients and research subjects.

There is no privacy

In 1999, the chief executive of Sun Microsystems, Scott McNealy, famously declared, "You have zero privacy anyway. Get over it."[9] While possibly hyperbolic two decades ago, his statement becomes truer with each passing year. Unlike many other countries, including Canada, Japan, and those in the European Union, the United States (US) has no overarching law that protects personal information.[10] Instead, there are specific laws that cover specific data types, including driver's licenses,[11] educational records,[12]

[*] Northwestern University Feinberg School of Medicine, US

[†] Northwestern Memorial HealthCare, US

[‡] Ann & Robert H. Lurie Children's Hospital of Chicago, US

Corresponding author: Justin Starren, MD, PhD, FACMI (Justin.starren@northwestern.edu)

credit reports,[13] video rental records,[14] and data produced by covered health care entities.[15] The US regulates privacy based on the entity creating the data rather than the content of the data. For comparison, web search data on the term "diabetes" would be protected in Canada as health information, but not in the US because it was not generated by a covered entity.[15]

Few patients or research subjects understand this nuanced distinction between source and content. Many incorrectly assume that sensitive data about their health is automatically protected. An analogy to how the US regulates data would be if chemicals were regulated based on the company that produced them. Under this approach, a chemical produced under the auspices of a pharmaceutical company would be regulated as a drug, but the same chemical produced under the auspices of a food company would not (eg, the cocaine produced as a byproduct of Coca-Cola production[16] would be unregulated). This is not how we regulate chemicals. It should not be the way we regulate data.

Also underappreciated is the vast amount of unregulated data, including medical data, that is collected on every individual. Data brokers — companies that collect, aggregate, and resell vast amounts of personal and sensitive data — are virtually unregulated. Justin Sherman, co-founder of Ethical Tech,[17] in testimony to the US Senate noted:

> Data brokers gather your race, ethnicity, religion, gender, sexual orientation, and income level; major life events like pregnancy and divorce; medical information like drug prescriptions and mental illness; your real-time smartphone location; details on your family members and friends; where you like to travel, what you search online, what doctor's office you visit, and which political figures and organizations you support.[18]

In *Our Bodies, Our Selves*,[19] Adam Tanner profiles the multibillion dollar business of selling medical records. He observes, "medical data miners cross-reference anonymized patient dossiers with named consumer profiles from data brokers," noting that one can easily purchase a fully identified list (ie, name, address, phone number, etc.) of people with a given disease, such as, "clinical depression, irritable bowel syndrome, erectile disfunction, even HIV."[19]

What data brokers do is legal. Even when behavior is clearly illegal, penalties are slight. Cambridge Analytica, a British consulting company, collected personal data on millions of Facebook users without their consent and used the data for political advertising to support the 2016 Trump presidential campaign.[20] While the company was prosecuted and went bankrupt, the punishments for individuals involved were minimal. The CEO was banned from serving as a corporate director for seven years; no one was incarcerated.[21] Although Facebook (Meta) was given a $5 billion fine by the Federal Trade Comission,[20] this was less than 6% of its revenue and Meta's stock price did not fall, which suggests that the stock market viewed this as "business as usual."

Many assume incorrectly that the Health Insurance Portability and Accountability Act (HIPAA) protects all clinical data. Data brokers have found ways around HIPAA.[19] Moreover, most clinical data sharing for research does not involve HIPAA-regulated data. Data transferred to researchers are no longer regulated under HIPAA. Even so, most research-related sharing involves removal of the 18 HIPAA-designated identifiers.[22] This is thought by some to render the data "safe" and freely sharable without restriction. Some institutions do not even consider HIPAA de-identified data to be human subject data.

The re-identification problem

In our experience, not only do many researchers not fully understand the "HIPAA 18" identifiers and fail to correctly remove them, but also "HIPAA 18" censoring does not make clinical data unidentifiable. Under HIPAA, "The covered entity also must have no actual knowledge that the remaining information could be used alone or in combination with other information to identify the individual."[22] Sherman noted that the sheer volume of data now available makes re-identification quite easy.[18] For example, the spacing of days between individual data entries (no actual dates) in a clinical data warehouse of over six million patients contained patterns of spacings that were unique for many patients (including author JS). Since care for JS occurred at specific locations, a cross-reference to cellphone location data could easily re-identify JS's data: the same way January 6 rioters have been identified.[23]

Since Latanya Sweeney re-identified the medical data of Governor Weld of Massachusetts,[24] a string of papers has illustrated the ability to re-identify individuals from supposedly anonymized data sets.[25,26] There is a veritable arms race between developers of anonymization algorithms and developers of re-identification algorithms. It is reasonable to assume that any individual-level data can be re-identified, if not today, then soon.

Threats to health care data privacy are increasing

On top of increased cyberattacks that have targeted health care data,[27] recent political events have accelerated the need to protect patient and research subject privacy.[28] Open data are not only available to researchers, they are also available to corporations and to the government. Police use public genetic databases to search for suspected criminals,[29] and can subpoena newborn genetic screening results.[30] Concerns about governmental access to clinical records are increasingly acute, following the Supreme Court's Dobbs v. Jackson Women's Health decision.[31] Some state Attorneys General are already attempting to subpoena privileged medical records, as has occurred in Indiana.[32] This behavior is not new. In 2004, after Congress passed the Partial-Birth Abortion Ban Act, Attorney General John Ashcroft subpoenaed medical records from multiple hospitals, including New York Presbyterian (NYP), and Northwestern Memorial Hospital (NMH). In the case of NYP, the hospital refused; the judge ruled against the hospital and later found the hospital in contempt. In the case of NMH, the hospital successfully blocked the subpoena, but the government appealed. The government later dropped the subpoenas when it became clear that NYP and the other hospitals were ready and willing to fight this all the way to the Supreme Court.[33]

These examples are relevant to clinical data sharing for research for several reasons. First, the government is legally allowed to obtain data from third parties that it cannot obtain directly. For example, the US government can bulk purchase data that it is legally forbidden to collect directly without a warrant, such as cellphone location data.[34] A government attorney, like Ashcroft, looking for evidence of a crime, could analyze openly shared clinical data and would only need to re-identify a fraction of the individuals to argue that a crime had been committed. Second, the Indiana case reinforces that this risk is not merely hypothetical. The International Classification of Diseases 10th Revision (ICD10-CM) contains many reproductive health codes (**Table 1**) that indicate activities considered "crimes" in certain states. Similarly, the Systematized Nomenclature of Human Medicine (SNOMED)[35] contains roughly 100 codes related to elective or attempted abortions. Normally, researchers do not purge reproductive health codes from shared data sets.

The new NIH Data Sharing Policy is a step in the right direction

The new NIH Data Sharing Policy[36] incorporates many of the concerns described above. It instructs, "Researchers should consider whether access to scientific data derived from humans, even if de-identified and lacking explicit limitations on subsequent use, should be controlled." The U.S. Department of Health and Human Services Secretary's Advisory Committee on Human Research Protections (SACHRP) states these concerns even more forcefully, declaring, "Increasingly, the protections afforded by removing the eighteen identifying data elements cited in HIPAA have become out of date, as technological advances and the combining of data sets increase the risk of re-identification."[37] The SACHRP further states, "Genomic data are also particularly susceptible to re-identification." The SACHRP makes several recommendations including controlled access for data from human participants, and stronger measures to deter misuse.

Table 1: ICD10-CM Codes Related to Reproductive Health Activities that are Illegal in some Jurisdictions.

Code	Definition
004.X	Complications following (induced) termination of pregnancy
007.X	Failed attempted termination of pregnancy
004.82	Renal failure following (induced) termination of pregnancy
099.32	Drug use complicating pregnancy, childbirth, and the puerperium
Z33.2	Encounter for elective termination of pregnancy
10A0	Abortion, Products of Conception

The walled garden approach to data sharing

The approaches to protecting privacy in shared research data can be divided into three broad categories: distributed computing, data-centric, and process-centric. With distributed computing, the data are not actually shared. Instead, each site typically creates identical data structures. Queries and algorithms are then run locally against those structures, and the results are aggregated. The Observational Health Data Sciences and Informatics (OHDSI) network is one of the largest examples of distributed computing.[38] While very useful for simple queries, this approach can be limiting for machine learning because of the necessity for every site to support the correct computer environment for the algorithm.

The data-centric approach is to create data that "can defend itself".[39] The goal is to modify or "harden" the data to the point that they can be shared without restrictions because there is a sufficiently low risk of reidentification. Many approaches have been used, which include: removing certain data types, such as the 18 HIPAA identifiers; censoring cells when the number of subjects is below a certain threshold; and censoring extreme values. As we now know, these approaches reduce, but do not eliminate, reidentification risk.

Recently, there has been considerable interest in synthetic data as a way to truly anonymize data (that is, to make it never re-identifiable) while preserving its inferential scientific value. It is worth noting that synthetic data comes in two forms: fully synthetic and partially synthetic. Fully synthetic data are generated completely *de novo* — without relying on preexisting data — and are intended to provide data that looks like real data, but that may have little inferential value, and consequently little research value. The Medicare Claims Synthetic Public Use Files (SynPUFs) are examples of fully synthetic data.[40] Partially synthetic data are derived from real data with the intent of preserving the inferential value while reducing the reidentification risk.[41,42] Although some claim that synthetic data is immune from reidentification risk,[43] partially synthetic data is vulnerable to several risks, including membership inference,[44] which is when an adversary can infer that a target individual is in the data set, and can thereby infer other facts about the individual. Stadler, et al., evaluated five different algorithms for generating synthetic data from the *All of Us* data set and found that synthetic data did not provide a better tradeoff between privacy and utility than traditional deidentification techniques.[45] Stadler's work suggests that synthetic data will not be a "magic bullet" that slays the reidentification monster.

Given this, should researchers simply stop sharing data, even though data sharing can accelerate research and save lives? No; which brings us to process-centric approaches. These approaches assume that the data can be reidentified and, instead, focus on process controls and contractual obligations to ensure that the data recipient will not attempt to reidentify individuals or use the data for other than intended uses, and impose penalties as a deterrent. Research institutions are very familiar with process-centric approaches in the form of Data Use Agreements (DUAs) and Material Transfer Agreements (MTAs). Typically, these are bilateral agreements between two organizations. There are 142 medical schools that receive NIH grants, bilateral agreements between each pair of these institutions would require slightly over ten thousand separate agreements and would likely involve massive duplication of research data. In addition, DUAs and MTAs are typically limited to a single project, potentially resulting in hundreds of thousands of separate agreements. To be efficient and to reduce infrastructure duplication, large-scale, multi-institution clinical data sharing for research typically involves consortia of multiple institutions and a single, shared technology infrastructure. We call this the "walled garden" approach.

Physics and astronomy provide examples of the walled garden approach. The Large Hadron Collider (LHC) project, held up as the archetypal big data sharing example, rapidly sends petabytes of data to collaborators globally.[46] However, this is not *open* data sharing. Data sharing occurs only within the consortium. The first release of open data was in 2021, roughly a decade after the data was originally collected.[47] The Laser Interferometer Gravitational-Wave Observatory (LIGO)[48] has a similar approach: there is extensive sharing within the consortium, with clear rules for data use and consequences for misuse, but little data release outside the walls. With a walled garden approach, data is shared only with known and trusted individuals and institutions that are held accountable if trust is misplaced. In the biomedical domain, the *All of Us* research project[49] and the National Covid Cohort Consortium (N3C)[50] are among the best known examples of walled gardens.

Walled garden design considerations

For walled gardens to succeed, several considerations must be addressed. First, developing and maintaining the garden will involve significant infrastructure investment. For the LHC, LIGO, *All of Us,* and N3C, the data management infrastructures were massive multi-year, multimillion dollar endeavors that required large development teams and ongoing multimillion dollar operations expenditures. Entities, such as the European Council for Nuclear Research (CERN), the NIH or the National Science Foundation have resources of that scope. Second, someone must build the walls. Access to the garden should be though national centralized identity management. Only a large, national entity, such as the NIH, has identity management systems of sufficient scale. Third, the governance of the garden must be trusted and trustworthy. There should be clearly articulated principles and allowed uses for the data. For example, will for-profit researchers be granted access? What are researchers' obligations for disseminating results derived from the shared data? Whether the governance is appointed or democratic, public or private, the users of the garden must have confidence that data within the garden will only be used for appropriate purposes. Fourth, rules without punishment are not deterrents. The penalties for data misuse must, as SACHRP recommends, be severe enough to be an effective deterrent. Rule violations should be considered scientific misconduct and addressed accordingly. Scientific misconduct is much more common that we often like to admit.[51] In physics or astronomy, getting thrown out of the LHC or LIGO consortia as a result of bad behavior could be career ending. Penalties should also anticipate that not all garden users will be traditional academics, and that purely academic penalties may not be sufficient. Each email that violates the CAN-SPAM Act[52] can result in a penalty of $46,517 with no limit to the total fine. Clinical data privacy violations are generally considered more serious violations than spam email, and the penalties should reflect that. Finally, to maximize the scientific and societal benefit, entry to and use of the garden must be affordable. Some current gardens, such as *All of Us,* charge for cloud compute time above a basic allocation. User fees should never make access to the garden so onerous or expensive that only major corporations and elite universities can pass through the metaphorical gate.

What can data managers do today?

Adequate general purpose walled gardens for potentially identifiable biomedical data are not widely available today. So, what can individuals and institutions do now?

- Convene institutional leadership to establish acceptable thresholds for reidentification risk and criteria for evaluating that risk. For example, some institutions are comfortable with the level of protection provided by current synthetic data approaches; others are not.
- Establish processes for evaluating data sets prior to any sharing outside the institution. For example, who evaluates a data set prior to release? At Northwestern University, central review is now required for any clinical data.
- Reify these criteria and processes in publicly available policies.
- Identify data sharing repositories that meet institutional criteria for various risk levels. For example, PhysioNet is a repository for biomedical data that has the ability to enforce DUAs and that requires users to receive human subjects research training prior to data access.[53] We have used this at Northwestern for sharing data that, though HIPAA safe-harbor deidentified, presented a reidentification risk that was adjudged too high for open sharing.[54]
- Ensure that study consent documents honestly communicate that absolute anonymity cannot be guaranteed. Telling our research participants otherwise would be disingenuous and unethical.
- Increase awareness of reidentification risk among researchers, data managers, and data analysts, making it everyone's responsibility to consider the implications of reidentification risk.
- Encourage the developers of institutional data sharing software to support DUAs and the validation of external users.
- Encourage relevant governmental bodies to support the development and operation of appropriate walled gardens to accelerate research through the sharing of clinical data.
- Follow the reidentification literature and periodically reevaluate institutional criteria and policies.

Conclusion

In conclusion, we would suggest that Scott McNealy's privacy nihilist view was half right.[9] With everything happening today, we have close to zero privacy. But we do not need to "get over it" and give up. We can and should do better. The first step is to stop believing that we can truly anonymize data.

Competing Interests

The authors have no competing interests to declare.

References

1. **Horvitz E, Mulligan D.** Data, privacy, and the greater good. *Science*. 2015; 349(6245): 253–255. DOI: https://doi.org/10.1126/science.aac4520
2. **Subbian V, Solomonides A, Clarkson M,** et al. Ethics and informatics in the age of COVID-19: challenges and recommendations for public health organization and public policy. *J Am Med Inform Assoc*. 2021; 28(1): 184–189. DOI: https://doi.org/10.1093/jamia/ocaa188
3. **Centers for Medicare and Medicaid Services.** Hospital Inpatient Quality Reporting Program. CMS.gov. Published December 1, 2021. Accessed April 4, 2023. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/HospitalRHQDAPU
4. **St. Vincent S.** Data Privacy is a Human Right. Europe is moving toward Recognizing that. *Foreign Policy in Focus*. Published April 19, 2018. Accessed December 1, 2022. https://fpif.org/data-privacy-is-a-human-right-europe-is-moving-toward-recognizing-that/.
5. **United Nations.** *The Right to Privacy in the Digital Age. Report of the United Nations High Commissioner for Human Rights*.; 2021. Accessed December 1, 2022. https://documents-dds-ny.un.org/doc/UNDOC/GEN/G21/249/21/PDF/G2124921.pdf?OpenElement
6. **Greek Medicine.** History of Medicine Division, National Library of Medicine. Published February 7, 2012. Accessed December 2, 2022. https://www.nlm.nih.gov/hmd/greek/greek_oath.html
7. **Arellano AM, Dai W, Wang S, Jiang X, Ohno-Machado L.** Privacy Policy and Technology in Biomedical Data Science. *Annu Rev Biomed Data Sci*. 2018; 1(1): 115–129. DOI: https://doi.org/10.1146/annurev-biodatasci-080917-013416
8. What is Open Data? Open Data Handbook. Accessed December 1, 2022. https://opendatahandbook.org/guide/en/what-is-open-data/
9. **Sprenger P.** Sun on Privacy: "Get Over It." *Wired*. Published online January 26, 1999. Accessed December 1, 2022. https://www.wired.com/1999/01/sun-on-privacy-get-over-it/
10. **Privacy Laws Around the World.** pdpEcho. Published December 1, 2022. Accessed December 1, 2022. https://pdpecho.com/privacy-laws-around-the-world/
11. The Drivers Privacy Protection Act (DPPA) and the Privacy of Your State Motor Vehicle Record. Epic.org. Accessed December 1, 2022. https://epic.org/dppa/
12. **U.S. Department of Education.** Family Educational Rights and Privacy Act (FERPA). U.S. Department of Education. Published August 25, 2021. Accessed December 1, 2022. https://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html
13. **Bureau of Justice Assistance.** Fair Credit Reporting Act. Bureau of Justice Assistance. Accessed December 1, 2022. https://bja.ojp.gov/program/it/privacy-civil-liberties/authorities/statutes/2349
14. Video Privacy Protection Act. Wikipedia. Accessed December 1, 2022. https://en.wikipedia.org/wiki/Video_Privacy_Protection_Act
15. **U.S. Department of Health and Human Services.** HIPAA Home. U.S. Department of Health and Human Services. Accessed December 1, 2022. https://www.hhs.gov/hipaa/index.html
16. Coca-Cola Formula. Wikipedia. Accessed December 1, 2022. https://en.wikipedia.org/wiki/Coca-Cola_formula
17. **Ethical Tech Research Policy Education.** Ethical Tech. Accessed December 13, 2022. https://ethicaltech.duke.edu
18. **U.S. Senate Committee on Finance.** Data Brokerage and Threats to U.S. Privacy and Security Written Testimony. Accessed December 1, 2022. https://www.finance.senate.gov/imo/media/doc/Written%20Testimony%20-%20Justin%20Sherman.pdf
19. **Tanner A.** *Our Bodies, Our Data: How Companies Make Billions Selling Our Medical Records*. Beacon Press; 2017.
20. Facebook–Cambridge Analytica data scandal. Wikipedia. Accessed December 1, 2022. https://en.wikipedia.org/wiki/Facebook–Cambridge_Analytica_data_scandal
21. **Davies R.** Former Cambridge Analytica chief receives seven-year directorship ban. *The Guardian*. Published September 24, 2020. Accessed December 1, 2022. https://www.theguardian.com/uk-news/2020/sep/24/cambridge-analytica-directorship-ban-alexander-nix.
22. **National Institutes of Health.** How Can Covered Entities Use and Disclose Protected Health Information for Research and Comply with the Privacy Rule? HIPAA Privacy Rule. Accessed December 1, 2022. https://privacyruleandresearch.nih.gov/pr_08.asp
23. **Hall M.** The DOJ is creating maps from subpoenaed cell phone data to identify rioters involved with the Capitol insurrection. *Business Insider*. Published March 24, 2021. Accessed December 1, 2022. https://www.businessinsider.com/doj-is-mapping-cell-phone-location-data-from-capitol-rioters-2021-3.
24. **Meyer M.** Law, Ethics & Science of Re-identification Demonstrations. Bill of Health. Accessed December 1, 2022. https://blog.petrieflom.law.harvard.edu/symposia/law-ethics-science-of-re-identification-demonstrations/
25. **Narayanan A, Shmatikov V.** Robust De-anonymization of Large Sparse Datasets. In: *2008 IEEE Symposium on Security and Privacy (Sp 2008)*. IEEE; 2008; 111–125. DOI: https://doi.org/10.1109/SP.2008.33

26. **Malin B, Sweeney L.** How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. *J Biomed Inform*. 2004; 37(3): 179–192. DOI: https://doi.org/10.1016/j.jbi.2004.04.005
27. **Southwick R.** Cyberattacks in healthcare surged last year, and 2022 could be even worse. *Chief Healthcare Executive*. Published January 24, 2022. Accessed December 11, 2022. https://www.chiefhealthcareexecutive.com/view/cyberattacks-in-healthcare-surged-last-year-and-2022-could-be-even-worse.
28. **Clayton EW, Embí PJ, Malin BA.** Dobbs and the future of health data privacy for patients and healthcare organizations. *J Am Med Inform Assoc*. 2023; 30(1): 155–160. Erratum: *J Am Med Inform Assoc*. 30(1) January 2023, Page 208. DOI: https://doi.org/10.1093/jamia/ocac155
29. **Kaiser J.** A judge said police can search the DNA of 1 million Americans without their consent. What's next? *Science*. Published online November 7, 2019. DOI: https://doi.org/10.1126/science.aba1428
30. **Grant C.** Police Are Using Newborn Genetic Screening to Search for Suspects, Threatening Privacy and Public Health. ACLU News and Comentary. Published July 26, 2020. Accessed December 1, 2022. https://www.aclu.org/news/privacy-technology/police-are-using-newborn-genetic-screening
31. **Supreme Court of the United States.** *Dobbs v. Jackson Women's Health*.(Supreme Court of the United State 2022). Accessed December 1, 2022. https://www.supremecourt.gov/opinions/21pdf/19-1392_6j37.pdf
32. **Sasani A, Stolberg SG.** Indiana Attorney General Asks Medical Board to Discipline Abortion Doctor. *New York Times*. Published November 30, 2022. Accessed December 1, 2022. https://www.nytimes.com/2022/11/30/us/indiana-attorney-general-abortion-doctor.html.
33. **Freiden T.** U.S. drops fight to get abortion records. CNN.com Law Center. Published June 1, 2004. Accessed December 1, 2022. https://www.cnn.com/2004/LAW/04/27/abortion.records/
34. **Cyphers B.** How the Federal Government Buys Our Cell Phone Location Data. Electronic Frontier Foundation. Published June 13, 2022. Accessed December 1, 2022. https://www.eff.org/deeplinks/2022/06/how-federal-government-buys-our-cell-phone-location-data
35. **National Library of Medicine.** SNOMED CT Browsers. National Library of Medicine. Published December 5, 2022. Accessed December 5, 2022. https://www.nlm.nih.gov/research/umls/Snomed/snomed_browsers.html
36. **National Institutes of Health.** Final NIH Policy for Data Management and Sharing. Published October 29, 2023. Accessed December 1, 2022. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html
37. **U.S. Department of Health and Human Services.** Attachment A – NIH Data Sharing Policy. Office for Human Research Protections. Published September 17, 2020. Accessed December 5, 2022. https://www.hhs.gov/ohrp/sachrp-committee/recommendations/august-12-2020-attachment-a-nih-data-sharing-policy/index.html
38. **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: 574–578. PMID: PMID: 26262116; PMCID: PMC4815923.
39. **Medicare Claims Synthetic Public Use Files (SynPUFs).** CMS.gov. Published December 1, 2021. Accessed April 4, 2023. https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs
40. **Medicare Claims Synthetic Public Use Files (SynPUFs).** CMS.gov. Published December 1, 2021. Accessed April 4, 2023. https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs
41. **El Emam K, Mosquera L, Fang X.** Validating a membership disclosure metric for synthetic health data. *JAMIA Open*. 2022; 5(4): ooac083. DOI: https://doi.org/10.1093/jamiaopen/ooac083
42. **Kuo NIH, Polizzotto MN, Finfer S,** et al. The Health Gym: synthetic health-related datasets for the development of reinforcement learning algorithms. *Sci Data*. 2022; 9(1): 693. DOI: https://doi.org/10.1038/s41597-022-01784-7
43. **Platzer M.** AI-based Re-Identification Attacks – and how to Protect Against Them. Mostly.ai. Published April 22, 2022. Accessed April 3, 2023. https://mostly.ai/blog/synthetic-data-protects-from-ai-based-re-identification-attacks/
44. **Zhang Z, Yan C, Malin BA.** Membership inference attacks against synthetic health data. *J Biomed Inform*. 2022; 125: 103977. DOI: https://doi.org/10.1016/j.jbi.2021.103977
45. **Stadler T, Oprisanu B, Troncoso C.** Synthetic Data — Anonymisation Groundhog Day. 2022;(arXiv:2011.07018). Accessed April 3, 2023. http://arxiv.org/abs/2011.07018
46. **CERN, the European Organization for Nuclear Research.** The Network Challenge. CERN. Accessed June 27, 2023. https://home.cern/science/computing/network
47. CMS releases heavy-ion data from 2010 and 2011. opendata CERN. Published December 21, 2021. Accessed December 1, 2022. https://opendata.cern.ch/docs/cms-releases-heavy-ion-data
48. **Abramovici A, Althouse WE, Drever RWP,** et al. LIGO: The Laser Interferometer Gravitational-Wave Observatory. *Science*. 1992; 256(5055): 325–333. DOI: https://doi.org/10.1126/science.256.5055.325
49. **The All of Us Research Program Investigators.** The "All of Us" Research Program. *N Engl J Med*. 2019; 381(7): 668–676. DOI: https://doi.org/10.1056/NEJMsr1809937

50. **Haendel MA, Chute CG, Bennett TD,** et al. The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. *J Am Med Inform Assoc*. 2021; 28(3): 427–443. DOI: https://doi.org/10.1093/jamia/ocaa196
51. **Fanelli D.** How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data. Tregenza T (ed.), *PLoS ONE*. 2009; 4(5): e5738. DOI: https://doi.org/10.1371/journal.pone.0005738
52. **Federal Trade Comission.** CAN-SPAM Act: A Compliance Guide for Business. Federal Trade Commission. Published January 1, 2022. Accessed December 2, 2022. https://www.ftc.gov/business-guidance/resources/can-spam-act-compliance-guide-business
53. **Moody GB, Mark RG, Goldberger AL.** PhysioNet: a research resource for studies of complex physiologic and biomedical signals. *Comput Cardiol*. 2000; 27: 179–182. PMID: 14632011
54. **Markov N, Gao CA, Stoeger T,** et al. SCRIPT CarpeDiem Dataset: demographics, outcomes, and per-day clinical parameters for critically ill patients with suspected pneumonia. PhysioNet. DOI: https://doi.org/10.13026/5PHR-4R89

How to cite this article: Starren J, Rasmussen LV, Schneider DH, Nannapaneni P, Michelson K. A Privacy Nihilist Perspective on Clinical Data Sharing: Open Clinical Data Sharing is Dead, Long Live the Walled Garden. *Journal of the Society for Clinical Data Management*. 2023; 3(3): 4, pp. 1–7. DOI: https://doi.org/10.47912/jscdm.234

Submitted: 15 December 2022 **Accepted:** 25 May 2023 **Published:** XX Month 202X

JSCDM

Journal of the Society for Clinical Data Management is a peer-reviewed open access journal published by Society for Clinical Data Management.

OPEN ACCESS