

Foundations of Biomedical Data Sharing & De-Identification: Privacy-Enhancing Technologies

Lisa Pilgram, MD

Postdoctoral Fellow at the Electronic Health Information Laboratory (Khaled El Emam)



Agenda

Sharing Health Data

1

What are barriers to health data sharing?

Potential Risks With Data Sharing

2

What are concerns and real-world risks?

Ways to Mitigate Risk When Sharing Data

3

What is de-identification and synthetic data generation?

Conclusion and Future Perspectives

4

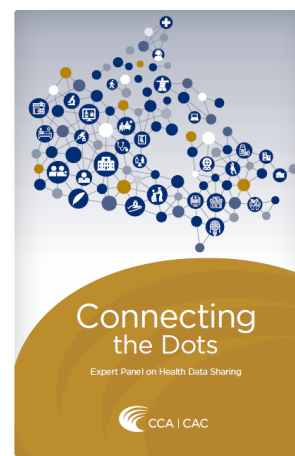
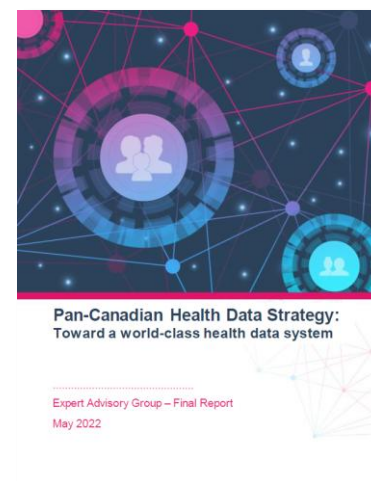
What are implications for practice?



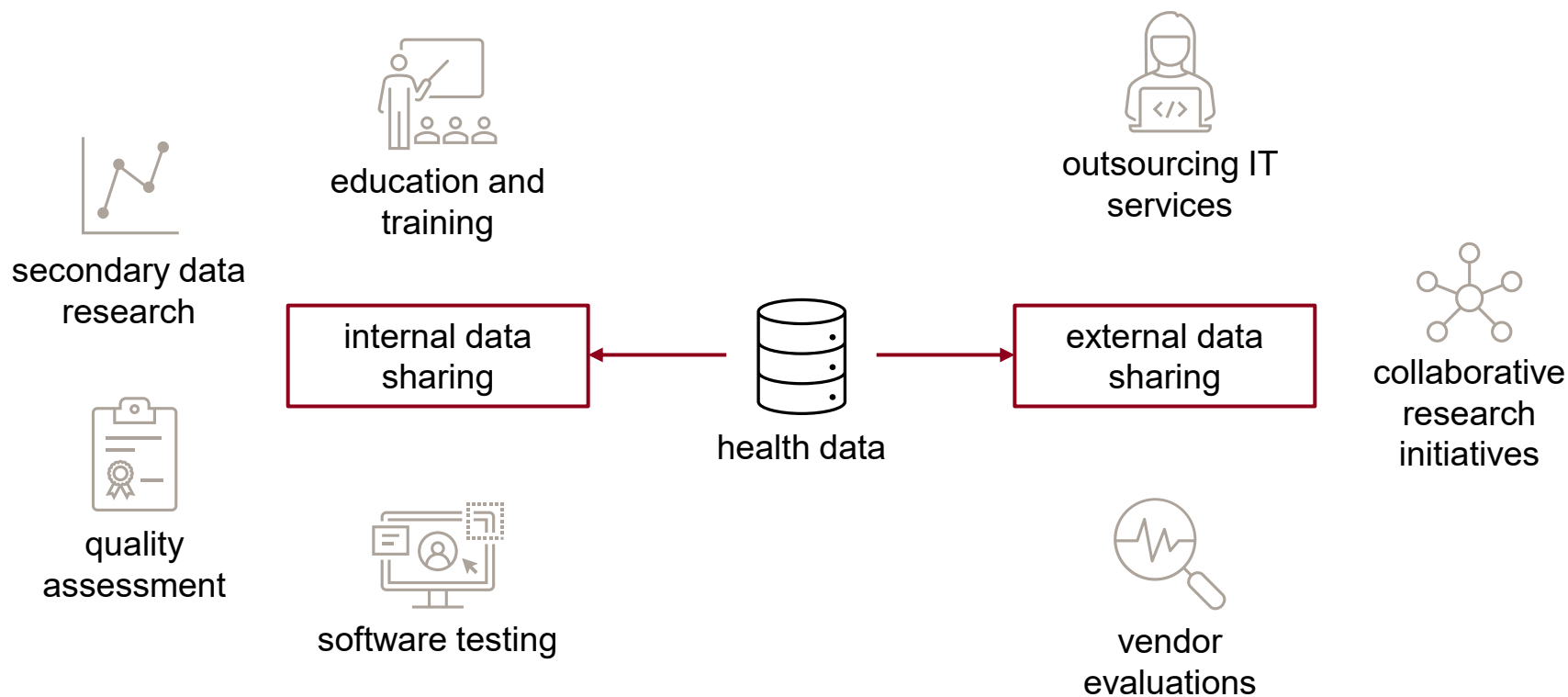
SHARING HEALTH DATA



Reports and Strategies to Access Health Data



Health Data is Essential for Multiple Internal and External Uses



Health Data is Essential for Multiple Internal and External Uses



Advances in data analysis and modeling to inform patient care make high-quality data sharing more important than ever before.



quality
assessment



software testing



vendor
evaluations

Privacy Concerns as a Barrier To Sharing Health Data

Privacy is considered the most prominent issue in big data research.

A. Ferretti et al. "The Challenges of Big Data for Research Ethics Committees: A Qualitative Swiss Study," *J Empir Res Hum Res Ethics*, vol. 17, no. 1–2, pp. 129–143, Feb. 2022, doi: 10.1177/15562646211053538

B.A. Malin, K. El Emam, C.M. O'Keefe. Biomedical data privacy: problems, perspectives, and recent advances. *J Am Med Inform Assoc*. 2013;20(1):2-6. doi:10.1136/amiajnl-2012-001509

Willingness to share health data for secondary purposes is generally high but privacy concerns can act as a barrier to sharing of health data.

K. B. Read et al. "Data-sharing practices in publications funded by the Canadian Institutes of Health Research: a descriptive analysis," *Canadian Medical Association Open Access Journal*, vol. 9, no. 4, pp. E980–E987, Oct. 2021, doi: 10.9778/cmajo.20200303

R. Trestian et al., "Privacy in a Time of COVID-19: How Concerned Are You?," *IEEE Secur. Privacy*, vol. 19, no. 5, pp. 26–35, Sep. 2021, doi: 10.1109/MSEC.2021.3092607

Q. Olsen et al., "Worldwide willingness to share health data high but privacy, consent and transparency paramount, a meta-analysis". *npj Digit. Med*. 8, 540 (2025). doi: 10.1038/s41746-025-01868-9

Privacy concerns can act as a barrier to seeking health care.

Pool J, Akhlaghpour S, Fatehi F, Gray LC. Data privacy concerns and use of telehealth in the aged care context: An integrative review and research agenda. *Int J Med Inform*. 2022;160:104707. doi:10.1016/j.ijmedinf.2022.104707

POTENTIAL RISKS WITH SHARING DATA



Who Wants Your Health Data ? And Why ?

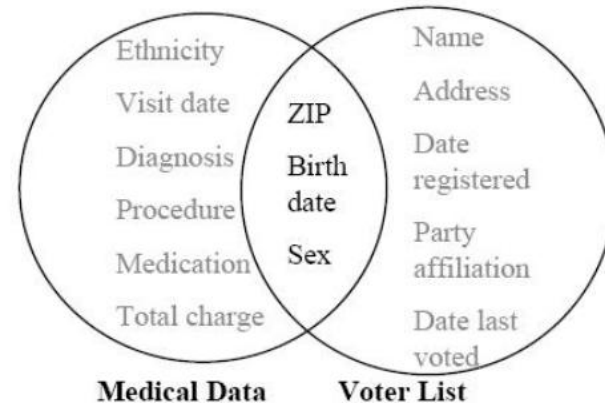
- Personal motives: There may be custody or divorce cases that trigger the search for additional information.
- Public recognition: The media can get a good story by showing that a particular dataset is not well protected.
- Selection or discrimination: An insurance company can select low-risk individuals for coverage.
- Profit: A marketer may link various datasets to get more granular consumer profiles to be sold to advertisers.
- Enforcement or control: A government could leverage data from period tracking apps to prosecute abortion-related cases.
- Self-interest: Individuals may engage in re-identification to discover biological relatives.

Most of this is speculative as we do not know for sure. What we know is that most documented re-identification attacks have been carried out by journalists or researchers.

Also see: Meurers T, Baum L, Haber AC, et al. Health Data Re-Identification: Assessing Adversaries and Potential Harms. *Stud Health Technol Inform.* 2024;316:1199–203. doi: 10.3233/SHTI240626

The Adversary – A Researcher

- The Massachusetts Group Insurance Commission (GIC) released health records of state employees under the assumption that it was de-identified.
- A researcher, L. Sweeney, used that data and conducted a linking attack with the Massachusetts voter registry.
- She linked the two datasets using ZIP, birth date and sex,
- And successfully re-identified Governor William Weld’s medical record, including diagnoses and medications.
- In fact, 87%* of Americans could be singled out using only ZIP code, birth date, and gender.



From: L. Sweeney, “k-Anonymity: A Model for Protecting Privacy,” International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, vol. 10, no. 5, pp. 557–570, 2002.

*According to P. Golle (Revisiting the uniqueness of simple demographics in the US population. Proceedings of the 5th ACM Workshop on Privacy in Electronic Society. 2006), it is rather at 63% of the US population

Identity Disclosure: When an Individual is Assigned to a Record



Jane
*01.12.1989



Sex	Year of Birth	NDC
Male	1975	009-0031
Male	1988	0023-3670
Male	1972	0074-5182
Female	1993	0078-0379
Female	1989	65862-403
Male	1991	55714-4446
Male	1992	55714-4402
Female	1987	55566-2110
Male	1971	55289-324
Female	1996	54868-6348
Male	1980	53808-0540

Which record belongs to Jane?

Ways to Mitigate Risk When Sharing Data



Privacy Risk Comes from Privacy Vulnerability and Re-Identification Attempt Probability

Vulnerability
of the Data



Probability
of Attempt

↓
“Traditional” De-Identification
Synthetic Data Generation

↓
Controlled access
Remote execution
Remote queries
Secure computation

Identity Disclosure is Driven by Equivalence Class Sizes Based on Quasi-Identifiers

Birth year	Gender	BMI	Pulse (bpm)	Obstructive nephropathy
1958	Female	22.5	87	Yes
1960	Male	28.5	65	Yes
1965	Female	32.5	100	Yes
1964	Male	25.3	96	No



Quasi-Identifiers (QI)

Identity Disclosure is Driven by Equivalence Class Sizes Based on Quasi-Identifiers

Birth year	Gender	BMI	Pulse (bpm)	Obstructive nephropathy
1958	Female	22.5	8	Vulnerability 1/1
1960	Male	28.5	6	Vulnerability 1/1
1965	Female	32.5	1	Vulnerability 1/1
1964	Male	25.3	9	Vulnerability 1/1



Quasi-Identifiers (QI)

* These calculations are for illustrative purposes only and assume that the data used represents the entire population. If you want to learn more about de-identification: subscribe to our mailing list and join our de-identification courses

<https://www.ehealthinformation.ca/Events>

When We Generalize the Class Size Gets Bigger, so the Vulnerability Decreases

Birth year	Gender	BMI	Pulse (bpm)	Obstructive nephropathy
1950-1959	Female	18.5-24.9	8	Vulnerability 1/1
1960-1969	Male	25.0-29.9	6	Vulnerability 1/2
1960-1969	Female	30.0-34.9	1	Vulnerability 1/1
1960-1969	Male	25.0-29.9	9	Vulnerability 1/2



Quasi-Identifiers (QI)

“Traditional” de-identification methods typically leverage generalization and suppression.

* These calculations are for illustrative purposes only and assume that the data used represents the entire population. If you want to learn more about de-identification: subscribe to our mailing list and join our de-identification courses

<https://www.ehealthinformation.ca/Events>

ARX – A Tool to Automate De-Identification

The screenshot displays the ARX Anonymization Tool interface. The main window is titled "ARX Anonymization Tool - anonym-webinar". The top menu includes "File", "Edit", "View", and "Help". The status bar indicates "Attribute: CMS Certification Nu...", "Transformations: 336", "Selected: [4, 3, 0, 0]", and "Applied: [4, 3, 0, 0]".

The "Input data" pane shows a table with columns: Facility Name, CMS Certification Number (CCN), Alternate CCN, Address, City, and State. The table contains 33 rows of data, including entries like "ATLANTIS GUAYAMA", "Atlantis Renal Ce...", "Centro Renal Uni...", "FKC CIUDAD CRI...", "FKC NARANJITO", "FMC Aguadilla D...", "FMC AIBONITO", "FMC Arecibo Dia...", "FMC Arecibo No...", "FMC Bayamon", "FMC Caguas Dial...", "FMC Canovanas", "FMC Carolina Di...", "FMC Guayama D...", "FMC Humacao D...", "FMC Las Piedras", "FMC Los Paseos...", and "FMC Mayaguez".

The "Data transformation" pane shows "Attribute metadata" with "Type: Quasi-identifying" and "Transformation: Generalization". The "Minimum" and "Maximum" are both set to "All". Below this is a table showing the transformation results for various attributes across different levels (Level-0 to Level-6).

Level-0	Level-1	Level-2	Level-3	Level-4	Level-5	Level-6
012306	01230*	0123**	012***	01****	0*****	*****
012500	01250*	0125**	012***	01****	0*****	*****
012501	01250*	0125**	012***	01****	0*****	*****
012502	01250*	0125**	012***	01****	0*****	*****
012505	01250*	0125**	012***	01****	0*****	*****
012506	01250*	0125**	012***	01****	0*****	*****
012507	01250*	0125**	012***	01****	0*****	*****
012508	01250*	0125**	012***	01****	0*****	*****
012509	01250*	0125**	012***	01****	0*****	*****
012512	01251*	0125**	012***	01****	0*****	*****
012513	01251*	0125**	012***	01****	0*****	*****
012515	01251*	0125**	012***	01****	0*****	*****
012516	01251*	0125**	012***	01****	0*****	*****
012517	01251*	0125**	012***	01****	0*****	*****
012519	01251*	0125**	012***	01****	0*****	*****
012520	01252*	0125**	012***	01****	0*****	*****
012521	01252*	0125**	012***	01****	0*****	*****
012522	01252*	0125**	012***	01****	0*****	*****

The "Privacy models" pane shows "Type: Model" and "Attribute: 11-Anonymity". The "General settings" pane includes "Utility measure", "Coding model", and "Attribute weights". The "Suppression limit" is set to 10%, and "Approximate" is checked with "Assume practical monotonicity". "Precomputation" is unchecked with "Enable. Threshold" set to 0%.

The "Sample extraction" pane shows "Size: 7625 / 7625 = 100%" and "Selection mode: None".

Elect Tool: Reference: Prasser F, Kohlmayer F, Lautenschläger R, Kuhn KA. ARX--A Comprehensive Tool for Anonymizing Biomedical Data. AMIA Annu Symp Proc. 2014;2014:984-993. Published 2014 Nov 14. <https://arx.deidentifier.org/> wa

Privacy Risk Comes from Privacy Vulnerability and Re-Identification Attempt Probability

Vulnerability
of the Data

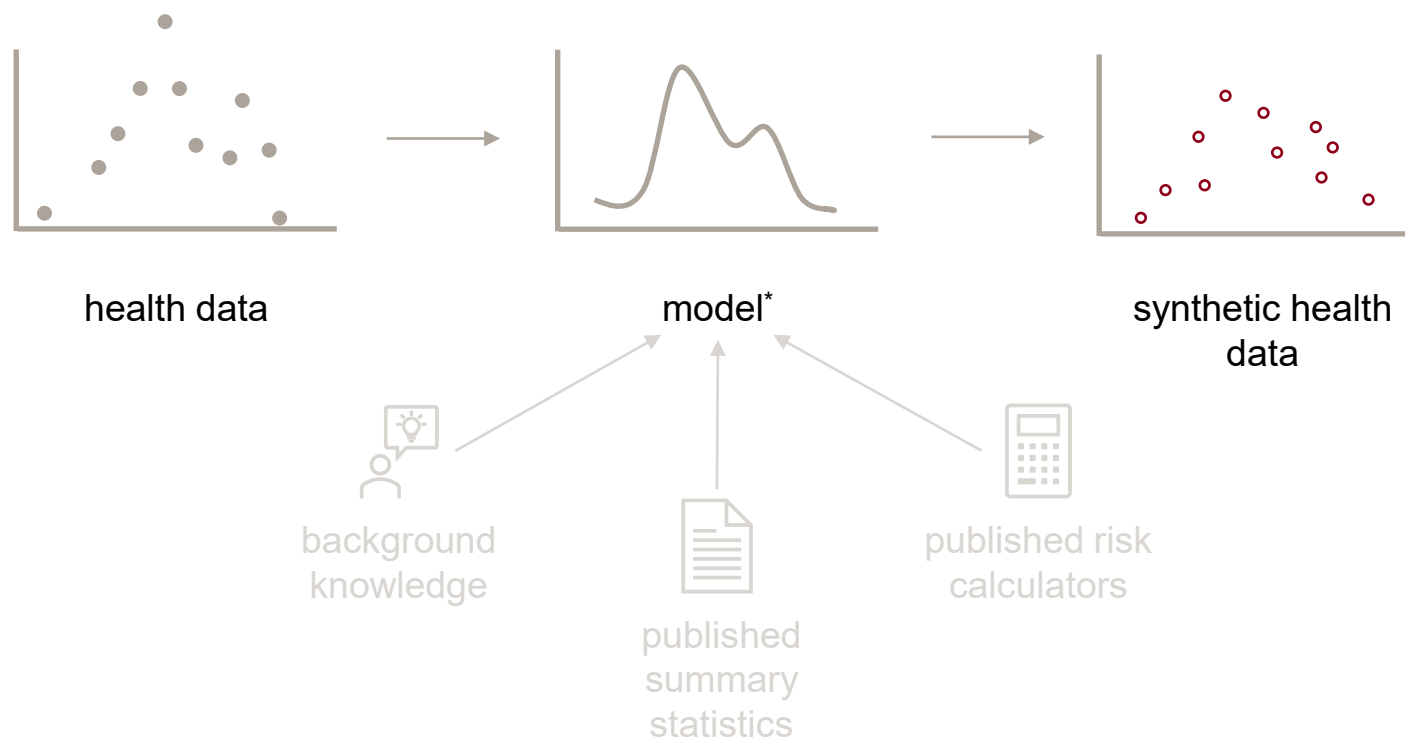


Probability
of Attempt

↓
“Traditional” De-Identification
Synthetic Data Generation

↓
Controlled access
Remote execution
Remote queries
Secure computation

Generating Synthetic Data from Health Data



* The EHIL maintains the Python library **pysdg** with a unified interface to multiple SDG models:
<https://github.com/CHEO-EHIL/pysdg-releases>

Synthetic Data Generation Seems to Have a Better Privacy-Utility Trade-Off

Traditional de-identification can sometimes be hard and can carry relevant utility costs.

Pilgram L, Meurers T, Malin B, Schaeffner E, Eckardt K-U, Prasser F, GCKD Investigators. The Costs of Anonymization: Case Study Using Clinical Data. *J Med Internet Res*. 2024;26:e49445. Published 2024 Apr 24. doi:10.2196/49445

Myers CT, Kumar RD, Pilgram L, Bonomi L, Thomas M, Griffith OL, Fullerton SM, Gibbs RA. Genomic data and privacy. *Clinical chemistry*. 2025 Jan;71(1):10-7.

Synthetic data generation is an evolving technology with promising results in terms of utility.

El Emam K, Mosquera L, Fang X, El-Hussuna A. An evaluation of the replicability of analyses using synthetic health data. *Scientific Reports*. 2024 Mar 24;14(1):6978.

van Breugel B, Liu T, Oglic D, van der Schaar M. Synthetic data in biomedicine via generative artificial intelligence. *Nature Reviews Bioengineering*. 2024 Dec;2(12):991-1004.

Synthetic data should protect against identity disclosure by design, though residual vulnerabilities may still exist and require assessment.

Pilgram L, Dankar FK, Drechsler J, et al. A consensus privacy metrics framework for synthetic data. *Patterns*. 2025 July 29. <https://doi.org/10.1016/j.patter.2025.101320>

Pilgram L, Fineberg A, Jonker E, El Emam K. An Assessment of Synthetic Data Generation, Use and Disclosure Under Canadian Privacy Regulations. *AI Ethics* (2025). <https://doi.org/10.1007/s43681-025-00819-0>



Further Transformation Tools

Computational and Structural Biotechnology Journal 23 (2024) 2892–2910

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj




Synthetic data generation methods in healthcare: A review on open-source tools and methods

Vasileios C. Pezoulas^{a,b}, Dimitrios I. Zaridis^{a,b,c}, Eugenia Mylona^{a,b}, Christos Androutsos^a, Kosmas Apostolidis^{a,b}, Nikolaos S. Tachos^{a,b}, Dimitrios I. Fotiadis^{a,b,*}

^a Unit of Medical Technology and Intelligent Information Systems, Dept. of Materials Science and Engineering, Biomedical Research Institute - FORTH, University of Ioannina, Ioannina GR45110, Greece
^b Biomedical Engineering Laboratory, School of Electrical and Computer Engineering, National Technical University of Athens, Athens GR10700, Greece

ARTICLE INFO

Keywords:
 Synthetic data generation
 Data privacy
 Healthcare
 Artificial intelligence
 Tabular data
 Imaging data
 Radiomics data
 Time-series data
 Omics data
 Multimodal data



ABSTRACT

Synthetic data generation has emerged as a solution to data scarcity and privacy concerns, offering a means to generate synthetic data with sufficient variability and unbiased data with sufficient volume. This study systematically searched the PubMed database for articles on synthetic data generation methods used for the synthesis of time-series, and omics data. Studies were included in the synthesis of statistical, probabilistic, machine learning, and natural language processing languages used for the implementation of synthetic data generators to generate synthetic data under various conditions, (ii) enhance the predictive performance of machine learning models for fair treatment recommendations across different patient groups, (iii) improve the quality, representative multimodal data, and (iv) underline the wide use of deep learning models in synthetic data generation. Finally provided to accelerate research in the field of synthetic data generation.

Abu Attieh et al. *BMC Medical Informatics and Decision Making* (2025) 25:128
<https://doi.org/10.1186/s12911-025-02958-0>

BMC Medical Informatics and Decision Making

SYSTEMATIC REVIEW **Open Access**





Pseudonymization tools for medical research: a systematic review

Hammam Abu Attieh^{1*}, Armin Müller¹, Felix Nikolaus Wirth¹ and Fabian Prasser¹

Briefings in Bioinformatics, 2022, 23(6), 1–10
<https://doi.org/10.1093/bib/bbac440>
 Advance access publication date 10 October 2022

Open tools for quantitative anonymization of tabular phenotype data: literature review

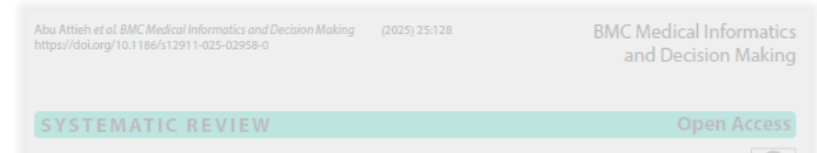
Anna C. Haber, Ulrich Sax and Fabian Prasser  on behalf of the NFDI4Health Consortium
 Corresponding author: Fabian Prasser, Health Data Science Center, Berlin Institute of Health at Charité—Universitätsmedizin Berlin. Tel.: +49 152 04 31 80 73; E-mail: fabian.prasser@charite.de

Abstract
 Precision medicine relies on molecular and systems biology methods as well as bidirectional association studies of phenotypes and (high-throughput) genomic data. However, the integrated use of such data often faces obstacles, especially in regards to data protection. An important prerequisite for research data processing is usually informed consent. But collecting consent is not always feasible, in particular when data are to be analyzed retrospectively. For phenotype data, anonymization, i.e. the altering of data in such a way that individuals cannot be identified, can provide an alternative. Several re-identification attacks have shown that this is a complex task and that simply removing directly identifying attributes such as names is usually not enough. More formal approaches are needed that use mathematical models to quantify risks and guide their reduction. Due to the complexity of these techniques, it is challenging and not advisable to implement them from scratch. Open software libraries and tools can provide a robust alternative. However, also the range of available anonymization tools is heterogeneous and obtaining an overview of their strengths and weaknesses is difficult due to the complexity of the problem space. We therefore performed a systematic review of open anonymization tools for structured phenotype data described in the literature between 1990 and 2021. Through a two-step eligibility assessment process, we selected 13 tools for an in-depth analysis. By comparing the supported anonymization techniques and further aspects, such as maturity, we derive recommendations for tools to use for anonymizing phenotype datasets with different properties.

Keywords: privacy, phenotype, data anonymization, software, review



Further Transformation Tools



Synthetic data generation methods in healthcare: A review on open-source

Pseudonymization tools for medical research: a systematic review

There are many approaches to reduce the identifiability of health data. Ultimately, it does not matter so much which one you use, but that you can demonstrate the risks are acceptably low.


phenotype data: literature review

and efficacy of synthetic data methods in healthcare considering the complexity of medical data. To our end, we systematically searched the PubMed database for studies that have used synthetic data generation in the context of time-series, and omics data. Studies involving multi-modal synthetic data generation were also explored. The type of method used for the synthetic data generation process was identified in each study and was categorized into statistical, probabilistic, machine learning, and generative adversarial networks (GANs). The languages used for the implementation of each method. Our evaluation revealed that the majority of the studies utilize synthetic data generators to: (i) reduce the cost and time required for clinical trials for rare diseases and conditions, (ii) enhance the predictive power of AI models in personalized medicine, (iii) ensure the delivery of fair treatment recommendations across diverse patient populations, and (iv) enable researchers to access high-quality, representative multimodal data without exposing sensitive patient information, among others. We underline the wide use of deep learning based synthetic data generators in 72.6 % of the included studies, with 75.3 % of the generators being implemented in Python. Finally provided to accelerate research in precision medicine relies on molecular and systems biology methods as well as bidirectional association studies of phenotypes and high-throughput genomic data. However, the integrated use of such data often faces obstacles, especially in regards to data protection and privacy. This study evaluated the current landscape of pseudonymization tools, of which 10 met our inclusion criteria and were assessed. The results show that there are differences between the tools that make them more or less suited for research projects differing in regards to the dimensions described above, enabling us to provide targeted recommendations.

Conclusions The landscape of existing pseudonymization tools is heterogeneous, and researchers need to carefully select the tool that best fits their needs. Our findings highlight two Software-as-a-Service-based tools for retrospective pseudonymization of smaller, short-term projects, and two tools well-suited for larger, long-term projects.

Keywords: privacy, phenotype, data anonymization, software, review


Risk in De-Identified and Synthetic Data is Quantifiable and Automatable



Trust in Data

De-identified Data Evaluation Report

Data custodian: ClientName
Data recipient: RecipientName
Date: September 02, 2025 at 2:23 AM (EDT)
Generated by: Evidata by Woodway Assurance Ltd (version 0.1, contact: info@woodway-assurance.com)



The Identity Disclosure Risk is HIGH.

Disclaimer: This assessment is based on information provided by ClientName, along with assumptions made in the context of the assessment. ClientName is responsible for verifying that these assumptions remain valid for their use case. A re-assessment is required if conditions or assumptions change in a material way. Results are valid for up to 2 years from the above date or until any condition or assumption is no longer met.

Table 1: Key Assumptions.

Key Assumptions:
<ul style="list-style-type: none"> • The context is non-public reuse or data sharing. • Medium (security and privacy) controls have been implemented by the anticipated

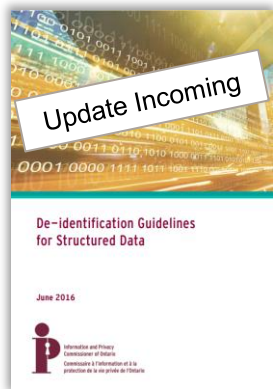
Tool: Woodway Assurance is a spin-off from our lab that provides automated, independent third-party risk assessment of de-identified, anonymized and synthetic data. <https://www.woodway-assurance.com/>

Electronic Health Information Laboratory, Children’s Hospital of Eastern Ontario Research Institute

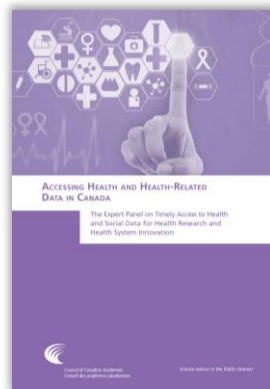
Conclusion and Future Perspectives



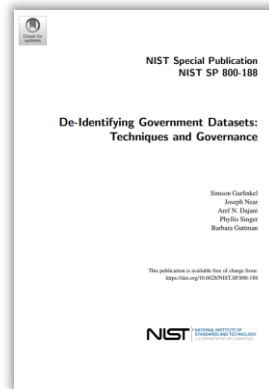
Well-Established Standards and Guidelines for De-Identification



Canada



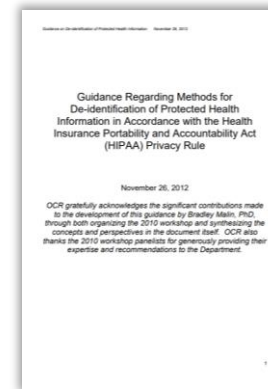
Canada



USA



USA



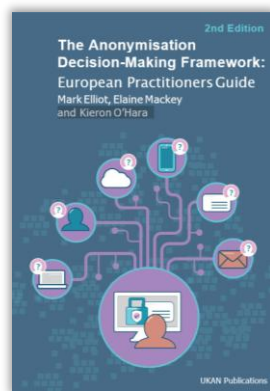
USA



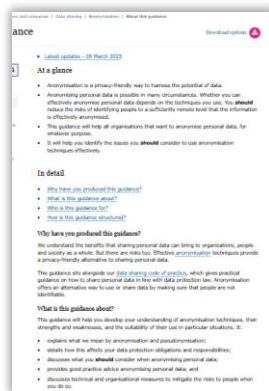
International



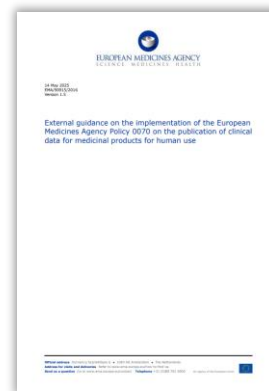
UK



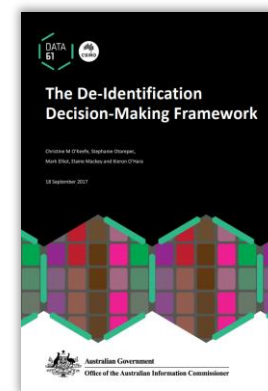
UK



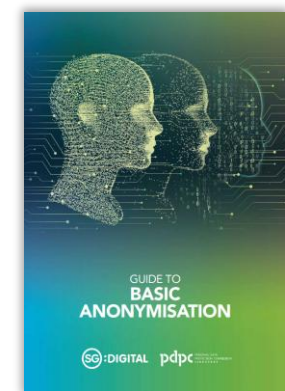
UK



Europe



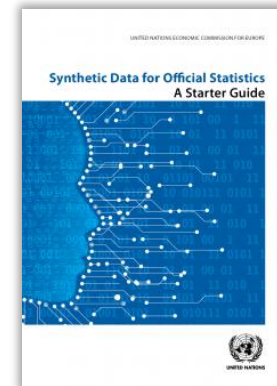
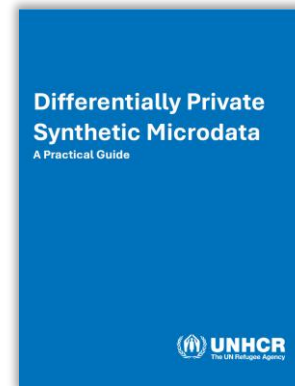
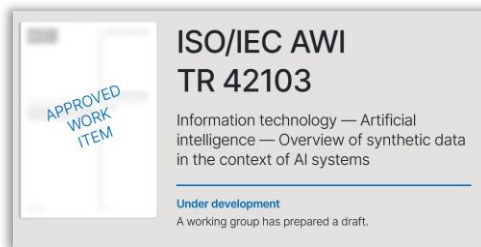
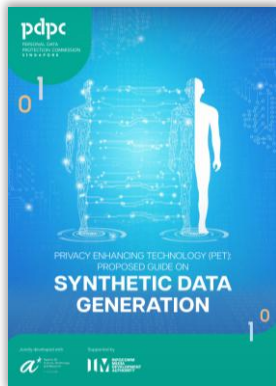
Australia



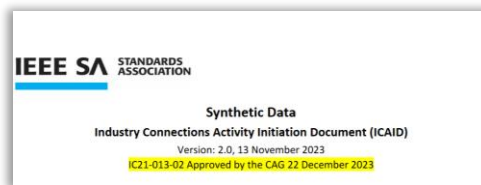
Singapore



Emerging International Frameworks and Standards for Synthetic Data

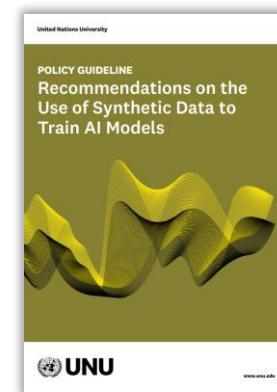


international standards



regulatory guidelines

non-governmental guidelines



Privacy-Preserving Health Data Sharing - Learnings

- Access to health data accelerates **research** and **innovation**, improves **quality of care** and fosters **transparency**.
- Privacy-enhancing technologies such as **de-identification and synthetic data generation** can relevantly reduce risks.
- Residual vulnerabilities can be **quantified** and should be documented.
- **Automated tools** can help with de-identification, synthetic data generation and risk assessment.
- There are **governance frameworks and standards** that ensure best practice.

If you want to learn more about de-identification: subscribe to our mailing list and join our de-identification courses <https://www.ehealthinformation.ca/Events>



THANK YOU

Lisa Pilgram, MD

Postdoctoral Fellow at the Electronic Health Information Laboratory (Khaled El Emam)