

Rethinking Trust in Synthetic Health Data: Lessons from Seven European Research Initiatives

Jens Declerck, Dipak Kalra, Antti Airola, Ahmed Youssef Ali Amer, Christos Chatzichristos, Maria del Mar Mañu, Bruno M. de Brito Robalo, Francesco Ghini, Alberto Gutierrez-Torre, Sem Hoogteijling, Susanne Hultsch, Jan Ramon, Sara Reidel, Francesco Regazzoni, Luís Silva, Inês Silveira, Tsekeridou Sofia, Christophe Maes

Submitted to: Journal of Medical Internet Research
on: September 01, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

| | |
|----------------------------------|-----------|
| Original Manuscript | 5 |
| Supplementary Files | 18 |
| Multimedia Appendixes | 19 |
| Multimedia Appendix 1..... | 19 |



Rethinking Trust in Synthetic Health Data: Lessons from Seven European Research Initiatives

Jens Declerck^{1,2} MSc; Dipak Kalra¹ Prof Dr; Antti Airola³; Ahmed Youssef Ali Amer⁴; Christos Chatzichristos⁵; Maria del Mar Mañu⁶; Bruno M. de Brito Robalo⁷; Francesco Ghini⁸; Alberto Gutierrez-Torre⁹; Sem Hoogteijling¹⁰; Susanne Hultsch¹¹; Jan Ramon¹²; Sara Reidel^{13,14}; Francesco Regazzoni¹⁵; Luís Silva¹⁶; Inês Silveira¹⁶; Tsekeridou Sofia¹⁷; Christophe Maes^{1,2}

¹ The European Institute for Innovation through Health Data Oosterzele BE

² Unit of Medical Informatics and Statistics Department of Public Health and Primary Care The University of Ghent Ghent BE

³ Department of Computing University of Turku Turku FI

⁴ Innovative Medicine Johnson & Johnson Beerse BE

⁵ Department of Electrical Engineering KU Leuven Leuven BE

⁶ University Hospital Vall d'Hebron Barcelona ES

⁷ Department of Obstetrics and Gynaecology Erasmus MC University Medical Center Rotterdam NL

⁸ Laboratory of Translational Research Reggio Emilia IT

⁹ Barcelona Supercomputing Center Barcelona ES

¹⁰ Department of Neurology and Neurosurgery University Medical Center Utrecht Brain Center Utrecht NL

¹¹ VEIL.AI Helsinki FI

¹² INRIA Lille FR

¹³ Vall d'Hebron Research Institute University Hospital Vall d'Hebron Barcelona ES

¹⁴ BarcelonaTech (UPC) Universitat Politècnica de Catalunya Barcelona ES

¹⁵ University of Amsterdam Amsterdam NL

¹⁶ LIBPhys-UNL NOVA School of Science and Technology Caparica PT

¹⁷ Netcompany SEE & EUI Copenhagen DK

Corresponding Author:

Jens Declerck MSc

The European Institute for Innovation through Health Data
Merebaaistraat 10
Oosterzele
BE

Abstract

Synthetic data generation (SDG) structured health data is increasingly promoted as a solution to longstanding barriers in health data access. It is offering the promise of privacy-preserving data reuse for research, innovation, and policy. Despite rapid technical advances, the adoption of synthetic health data in real-world settings remains limited. Shaped by challenges around data quality, representativeness, infrastructure readiness, trust, and legal uncertainty. This viewpoint draws on experiences from seven European research initiatives within the HealthData4EU cluster to reflect on how SDG is being operationalized in practice. It synthesizes cross-project insights to highlight recurring methodological and governance tensions and to examine their implications for trust and responsible use. The analysis argues that trustworthy SDG cannot be achieved through technical optimization alone, but requires alignment between evaluation practices, upstream data stewardship, regulatory clarity, and sustained stakeholder engagement. Addressing these conditions is essential for moving synthetic data from experimental pilots toward a credible and sustainable component of European health research ecosystems.

(JMIR Preprints 01/09/2025:83369)

DOI: <https://doi.org/10.2196/preprints.83369>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ Please make my preprint PDF available to anyone at any time (recommended).

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in [JMIR Publications](#).

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in [JMIR Publications](#).

Preprint
JMIR Publications

Original Manuscript



Rethinking Trust in Synthetic Health Data: Lessons from Seven European Research Initiatives

Abstract

Synthetic data generation (SDG) structured health data is increasingly promoted as a solution to longstanding barriers in health data access. It is offering the promise of privacy-preserving data reuse for research, innovation, and policy. Despite rapid technical advances, the adoption of synthetic health data in real-world settings remains limited. Shaped by challenges around data quality, representativeness, infrastructure readiness, trust, and legal uncertainty. This viewpoint draws on experiences from seven European research initiatives within the HealthData4EU cluster to reflect on how SDG is being operationalized in practice. It synthesizes cross-project insights to highlight recurring methodological and governance tensions and to examine their implications for trust and responsible use. The analysis argues that trustworthy SDG cannot be achieved through technical optimization alone, but requires alignment between evaluation practices, upstream data stewardship, regulatory clarity, and sustained stakeholder engagement. Addressing these conditions is essential for moving synthetic data from experimental pilots toward a credible and sustainable component of European health research ecosystems.

Introduction

As healthcare systems are largely digitized and become increasingly data-driven, access to large-scale, high-quality data is essential for clinical research [1-3], digital health innovation [4], and evidence-based policymaking [5]. However, accessing patient-level health data remains a persistent challenge [6]. Privacy regulations [7], data quality issues [8, 9], and data fragmentation [10] continue to limit data access and sharing, especially in cross-organizational studies [9, 11]. These barriers not only slow innovation but also risk underrepresenting certain diseases and populations, such as rare conditions, in secondary data use. As a result, many promising analytical and clinical applications struggle to move beyond pilots.

Synthetic data generation (SDG) has emerged as a promising response to the barriers limiting access and reuse of patient-level health data [12]. By producing synthetic datasets that resemble the statistical and structural properties of real health data, without exposing personal identities, SDG can drive innovation while maintaining privacy [7]. It enables safer collaboration across different types of health data (e.g. electronic health records, genomics, and medical imaging) and offers value in areas such as AI development [7, 13], rare disease research [14], and model testing [13, 15]. The main aim of SDG is to retain the essential characteristics of original datasets, while retaining privacy (in most use cases). These can be grouped into quantitative [16] (e.g., statistical metrics, analytical patterns, signal-based features), qualitative [16] (e.g., expert-driven assessments), and domain-specific attributes [17] that depend on the data type and specific use case. Conceptual work has also emphasized the importance of incorporating domain knowledge into SDG processes to improve realism and relevance for clinical applications [18].

Despite this growing interest, the adoption of synthetic health data in real-world research and clinical environments remains limited [12, 19]. Concerns persist regarding data quality, representativeness, interpretability, and legal status [20, 21]. As well as the uncertainty about how synthetic data exceed what current methods and data infrastructures can deliver [13]. This gap between promise and practice has contributed to skepticism among multiple stakeholders, for

whom trust in data sources is essential [22]. This lack of trust is partly explained by the focus of SDG literature on technical methods and evaluation metrics [22-24]. While giving limited attention to the organizational, legal, and socio-technical conditions that shape real-world adoption [22, 24].

Several EU-funded initiatives are currently testing SDG in real-world health research settings, all facing a shared question: how can synthetic health data be made trustworthy, useful, and safe? Despite this activity, there has been limited cross-initiative reflection on how these challenges are being addressed in practice.

The aim of this viewpoint is to examine why trust in synthetic health data remains fragile despite growing technical maturity. And to reflect on the non-technical conditions required for its responsible adoption. This article is intended for health data researchers, clinicians, data stewards, infrastructure developers, regulators, and policy makers engaged in secondary use of health data across Europe. This article provides a viewpoint informed by experiences from seven European research initiatives within the HealthData4Eu cluster that work with structured synthetic health data. It synthesizes cross-project insights to highlight recurring challenges related to data quality, interoperability, regulatory alignment, and stakeholder acceptance. And reflects on their implications for trust, governance, and responsible use of synthetic data in European health research.

Basis of this viewpoint

Scope and context of the included projects

This viewpoint is informed by seven EU-funded projects that constitute the HealthData4EU cluster, a Horizon Europe initiative focused on advancing synthetic data generation and secure data use in health research across Europe. Collectively, these projects address complementary aspects of the synthetic data landscape, including data generation methods, federated and privacy-preserving architectures, secure data sharing, and AI-enabled clinical applications. All operate within a shared European policy and funding framework while engaging with diverse clinical and institutional contexts relevant to the European Health Data Space (EHDS).

The included initiatives represent the full set of synthetic data projects participating in the HealthData4EU cluster at the time of writing. While they operate within a shared European policy and funding framework, they span a wide range of clinical domains, data modalities (e.g., tabular, imaging, longitudinal, and multimodal data), and technical approaches. This diversity within a coordinated structure provides a useful basis for reflecting on how SDG is being operationalized across healthcare settings and where common challenges and tensions emerge. An overview of the participating initiatives with their clinical domains is provided in Table 1.

| Project | Clinical domain(s) | Example use cases |
|-------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------|
| AISym4MED | Neurological and chronic diseases (i.e., diffuse large B-cell lymphoma, lung cancer, multiple myeloma, Alzheimer's disease, type II diabetes, breast cancer). | Synthetic data for multimodal applications. Support for AI model testing. |
| FLUTE | Oncology (prostate cancer) | Risk stratification. Federated AI validation for prostate cancer diagnosis. |
| PHASE IV AI | Oncology and neurology (i.e., lung cancer, prostate cancer, ischaemic stroke). | Early lung cancer detection. Synthetic MRI for stroke and model testing. |
| PHEMS | Paediatric diseases (i.e., congenital cardiac conditions, sepsis, haemophilia). | Sepsis prediction in paediatric ICUs. Haemophilia care optimization. |
| SECURED | Multiple domains (i.e., mammography, histopathology, chest X-ray, cardiotocography). | Privacy-preserving SDG. Federated AI tool development. |
| SYNTHEMA | Rare haematological diseases (i.e., sickle-cell disease, acute myeloid leukaemia). | Synthetic datasets for rare disease research. Federated validation. |
| SYNTHIA | Multimodal and personalized medicine. | Benchmarking and validation of SDG methods across diverse clinical scenarios. |

Table 1 Overview of projects, clinical domains, and example use cases involving synthetic data

The insights synthesized in this paper draw on cross-project engagement conducted during the active lifetime of the initiatives. These insights were informed by cluster-level workshops, project presentations, and recurring exchanges among project teams. And by review of the project documentation during active implementation phase. Importantly, these interactions took place while projects were still ongoing, allowing challenges and emerging practices to be discussed as they arose rather than retrospectively. The exchanges involved multidisciplinary contributors, including technical developers, clinicians, data stewards, infrastructure providers, and legal and ethical experts.

While these insights are not derived from a formal empirical study. They reflect sustained, longitudinal engagement with projects that are actively developing and validating the real-world use of synthetic health data within regulated clinical and research environments. To improve transparency and address requests for additional project-level context, we have incorporated a

supplementary table providing a descriptive overview of the included projects.

Lessons from seven SDG projects

A shared landscape of methodological tensions

Across projects, SDG is not pursued as a single technical task but as a complex socio-technical intervention embedded in heterogeneous health systems [25]. In practice, SDG sits at the intersection of machine learning, clinical data infrastructures, regulatory governance, and institutional trust.

Despite differences in disease focus, data modalities, and maturity, projects consistently encounter several recurring challenges that shape how synthetic data generation is operationalized in practice. Across the initiatives examined in this paper, three themes emerged repeatedly: the absence of shared consensus on how synthetic data quality should be defined and evaluated, the limitations of existing of existing data infrastructures to support reliable SDG workflows, and uncertainty surrounding the regulatory and governance status of synthetic data. These challenges reflect structural characteristics of health data ecosystems, where technical innovation must coexist with legal safeguards, clinical accountability, and fragmented infrastructures. Understanding SDG through these challenges helps explain why synthetic data adoption remains difficult to scale beyond pilot environments.

Data quality without consensus

A central tension concerns how “data quality” in synthetic health data is defined and demonstrated. In practice, data quality is assessed through overlapping but fragmented lenses: statistical fidelity, analytical or clinical utility, and privacy protection. These dimensions are widely acknowledged across initiatives, yet there is no shared understanding of how they should be balanced, prioritised, or interpreted in different contexts [26].

Existing tools and metric often assume that real-world source data are themselves unbiased, complete, and representative [3, 6, 8, 9]. This assumption is rarely valid in health systems shaped by structural inequities, fragmented data capture, and variable clinical practices across institutions and regions. When real data reflect structural inequities or incomplete capture, synthetic data derived from them may inherit the same distortions [20].

As a result, SDG risks reinforcing existing biases rather than mitigating them. While synthetic data is frequently promoted as a mechanism for improving representativeness or enabling rare disease research, representational properties are frequently assessed primarily at the level of the resulting dataset [6]. In practice, however, representativeness is shaped not only by the quality of the synthetic generation process but also by the characteristics of the underlying source data and the transformations applied during the data lifecycle (e.g., data extraction, harmonization, mapping to common data models, and other ETL steps) [27]. These can introduce or amplify biases by selectively filtering variables, standardizing heterogeneous data sources, or excluding smaller subpopulations. These upstream influences may remain difficult to detect when evaluation focuses primarily on statistical similarity. As a result, population coverage and the potential impact of data preparation processes on representation may receive less attention in current evaluation practices.

In this sense, data quality cannot be reduced to statistical resemblance alone. We argue that there is a need for a more explicit data quality assessment at source and across the data lifecycle. In that sense results need to be interpreted in relation to the populations, decisions, and governance contexts in which synthetic data are intended to be used.

Infrastructure as a bottleneck to trust

Federated and decentralized SDG architectures are frequently presented as solutions to privacy and governance constraints [13]. In practice, infrastructure maturity at data sources emerges as a critical limiting factor. Heterogeneous hospital IT systems, inconsistent data models, and uneven data governance practices complicate federated training and validation. These challenges affect not only technical integration but also the reliability of data inputs and the feasibility of data quality assessments.

Interoperability challenges persist even when projects adopt common data models such as the OMOP CDM. Differences in local coding practices, laboratory workflows, and semantic interpretation (e.g., variation in LOINC mappings for similar laboratory measurements) require substantial upfront coordination. These nuances complicate federated validation and cross-site comparability, reinforcing that standardization alone does not eliminate heterogeneity in real-world health data. Within the HealthData4EU cluster, several projects reported that aligning local coding practices and laboratory mappings across institutions required substantial preprocessing effort before federated synthetic data generation or validation could be performed.

These infrastructural constraints shape what forms of SDG are feasible and who can participate, reinforcing a gap between technical potential and institutional readiness.

Trust, transparency, and regulatory uncertainty

Trust is not guaranteed by compliance

Most initiatives emphasize compliance with data protection and ethical frameworks. While compliance is necessary, it is not sufficient to generate trust. Trust emerges through transparency, interpretability, and meaningful stakeholder engagement across the SDG lifecycle [22].

In practice, these dimensions are unevenly operationalised. Some initiatives invest in dataset documentation, model cards, and participatory validation involving clinicians or data stewards. Others limit engagement to formal approval processes or internal review. This variability affects adoption, particularly in clinical and regulatory contexts where accountability and explainability are essential.

In the European context, this includes emerging obligations under the EU Artificial Intelligence Act, specific for high-risk medical AI systems. Where the provenance and validation of training data, including synthetic data, may become subject to increased scrutiny.

The unresolved legal status of synthetic data

Regulatory ambiguity remains one of the most significant barriers to uptake. Synthetic data is often assumed to fall outside data protection regulation, yet institutions frequently adopt conservative interpretations, particularly for rare diseases or small cohorts [28]. The absence of harmonized guidance leaves data controllers and ethics boards navigating uncertainty case by case.

Clarification is particularly needed in relation to GDPR anonymization thresholds, the interaction with the EU Artificial Intelligence Act and national interpretations by data protection authorities and ethics committees.

This uncertainty raises a fundamental but often unspoken question: what is the value of synthetic data if it cannot be confidently reused beyond its original context? If synthetic datasets remain confined to local pilots due to legal or institutional caution, their potential to support cross-border

research, reproducibility, and capacity building is limited. Addressing reusability therefore becomes central to the trust debate, linking legal clarity, documentation, and transparent validation of the broader promise of synthetic data as a shared research asset rather than a one-off technical item.

Emerging responses and their limits

Encouragingly, initiatives are beginning to respond through federated validation pipelines, benchmarking libraries, interoperability standards, and public-facing platforms. However, these responses remain fragmented, and no shared consensus has yet emerged on what constitutes trustworthy SDG across contexts.

Reflections and recommendations

The experiences discussed in this article suggest that synthetic data generation in health research has reached a point where technical feasibility is no longer the primary bottleneck. Instead, progress increasingly depends on how synthetic data is evaluated, governed, communicated, and trusted in practice.

First, there is a need to reframe how success in SDG is defined and assessed. Much of the current emphasis remains on demonstrating technical performance, such as statistical similarity or model accuracy. While these measures are important, they do not on their own indicate whether synthetic data are appropriate for research or operational contexts. In practice, synthetic datasets may serve different purposes (e.g., exploratory analysis, educational use, or hypothesis testing). These uses imply different expectations regarding the level of fidelity, reliability and validation required. Clarifying the intended context of use can therefore help guide how synthetic datasets are evaluated and interpreted. This does not imply that all potential analytical applications must be evaluated in advance, but rather that evaluation strategies should be transparent about the contexts for which synthetic datasets are considered suitable. At the same time, general evaluation metrics remain essential to characterize the overall fidelity and privacy properties of synthetic datasets. In many cases, findings derived from exploratory analysis using synthetic data may subsequently require validation using the original data sources where appropriate access mechanisms exist.

Second, greater attention should be given to data stewardship and governance upstream of SDG. Synthetic data reflects the characteristics and limitations of the infrastructures from which it is

generated. Without robust practices for data quality management, interoperability, and documentation at source, SDG risks perpetuating existing shortcomings rather than addressing them. Aligning SDG initiatives with broader efforts to strengthen health data infrastructures would therefore increase both the reliability of synthetic outputs and institutional confidence in their reuse.

Third, addressing the legal uncertainty surrounding synthetic health data is essential for wider adoption. In the absence of shared guidance, institutions often adopt cautious positions that limit reuse. Particularly in sensitive or cross-border contexts. Developing common interpretations, best practices, or certification pathways that clarify when synthetic data can be considered sufficiently anonymized would provide a more stable foundation for decision-making. Such approaches would not eliminate the need for contextual oversight, but they would reduce fragmentation and support more consistent application of safeguards.

Finally, trust building must be recognised as an active and ongoing process. Trust does not arise automatically from compliance with legal or ethical requirements. Nor from technical sophistication alone. It depends on transparency, interpretability, and the ability of stakeholders to understand how synthetic data was generated, evaluated and intended to be used. Embedding mechanisms such as clear documentation, accessible validation summaries, and engagement with clinicians, data stewards, and other end users can support accountability. And enable more nuanced judgments about risk and value.

Overall, these reflections point to a broader shift in focus: from demonstrating that synthetic data can be generated, to establishing the conditions under which it can be used responsibly, communicated transparently, and governed consistently. Advancing SDG along this path will require coordination across technical, institutional, and regulatory domains. If achieved, synthetic data can become not only a powerful analytical tool, but also a credible and trusted component of health research ecosystems.

Use of Generative AI

No generative artificial intelligence tools were used in the generation of this paper.

Funding

This paper was funded by the Innovative Health Initiative through the SYNTHIA project.

The study draws on insights from seven research projects that form part of the HealthData4EU cluster.

Projects within the cluster and their funding sources include:

- SYNTHEMA, PHASE IV AI, SECURED, FLUTE, and AI SYM4MED are funded by the Horizon Europe programme.
- PHEMS is funded by Horizon Europe and UK Research and Innovation (UKRI).
- SYNTHIA is funded through the Innovative Health Initiative (IHI).

Acknowledgements

The authors would like to thank all seven HealthData4EU projects for their valuable input, collaboration, and contributions to this paper. Their openness and insights were essential to shaping the cross-case analysis. The authors are also grateful to Iraia Nuñez and Anna Lorenzini for their dedicated coordination and communications support. While not involved in the technical content of this study, their efforts in organizing exchanges and facilitating collaboration across the projects played a vital role in enabling this synthesis.

1. Eden R, Burton-Jones A, Scott I, Staib A, Sullivan C. Effects of eHealth on hospital practice: synthesis of the current literature. *Australian Health Review*. 2018;42(5):568. doi: 10.1071/ah17255.

2. Zheng K, Abraham J, Novak LL, Reynolds TL, Gettinger A. A Survey of the Literature on Unintended Consequences Associated with Health Information Technology: 2014-2015. *Yearb Med Inform.* 2016 Nov 10(1):13-29. PMID: 27830227. doi: 10.15265/iy-2016-036.
3. Declerck J, Lee J, Sen A, Palmeri A, Oostenbrink R, Giannuzzi V, et al. The Potential to Leverage Real-World Data for Pediatric Clinical Trials: A Proof-of-Concept Study. *J Med Internet Res.* 2025;27:e72573. doi: 10.2196/72573.
4. Wang Z, Penning M, Zozus M. Analysis of Anesthesia Screens for Rule-Based Data Quality Assessment Opportunities. *Stud Health Technol Inform.* 2019;257:473-8. PMID: 30741242.
5. Wiebe N, Xu Y, Shaheen AA, Eastwood C, Boussat B, Quan H. Indicators of missing Electronic Medical Record (EMR) discharge summaries: A retrospective study on Canadian data. *Int J Popul Data Sci.* 2020 Dec 11;5(1):1352. PMID: 34007880. doi: 10.23889/ijpds.v5i3.1352.
6. Declerck J KD, Vander Stichele R, Coorevits P. Frameworks, dimensions, definitions of aspects and assessment methods for the appraisal of quality of health data for secondary use: a review of reviews. *JMIR Medical Informatics.* 2024.
7. Shabani M, Marelli L. Re-identifiability of genomic data and the GDPR. *EMBO reports.* 2019;20(6):e48316. doi: <https://doi.org/10.15252/embr.201948316>.
8. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *EGEMS (Wash DC).* 2016;4(1):1244. PMID: 27713905. doi: 10.13063/2327-9214.1244.
9. Declerck J, Vandenberg B, Deschepper M, Colpaert K, Cool L, Goemaere J, et al. Building a Foundation for High-Quality Health Data: Multihospital Case Study in Belgium. *JMIR Med Inform.* 2024 Dec 20;12:e60244. PMID: 39727158. doi: 10.2196/60244.
10. Wei WQ, Leibson CL, Ransom JE, Kho AN, Caraballo PJ, Chai HS, et al. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. *J Am Med Inform Assoc.* 2012 Mar-Apr;19(2):219-24. PMID: 22249968. doi: 10.1136/amiajnl-2011-000597.
11. Oja M, Tamm S, Mooses K, Pajusalu M, Talvik HA, Ott A, et al. Transforming Estonian health data to the Observational Medical Outcomes Partnership (OMOP) Common Data Model: lessons learned. *JAMIA Open.* 2023 Dec;6(4):ooad100. PMID: 38058679. doi: 10.1093/jamiaopen/ooad100.
12. Gonzales A, Guruswamy G, Smith SR. Synthetic data in health care: A narrative review. *PLOS Digital Health.* 2023;2(1):e0000082. doi: 10.1371/journal.pdig.0000082.
13. Giuffrè M, Shung DL. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *npj Digital Medicine.* 2023 2023/10/09;6(1):186. doi: 10.1038/s41746-023-00927-3.
14. Al-Dhamari I, Abu Attieh H, Prasser F. Synthetic datasets for open software development in rare disease research. *Orphanet Journal of Rare Diseases.* 2024 2024/07/15;19(1):265. doi: 10.1186/s13023-024-03254-2.
15. Goncalves A, Ray P, Soper B, Stevens J, Coyle L, Sales AP. Generation and evaluation of synthetic patient data. *BMC Medical Research Methodology.* 2020

2020/05/07;20(1):108. doi: 10.1186/s12874-020-00977-1.

16. Park N, Mohammadi M, Gorde K, Jajodia S, Park H, Kim Y. Data synthesis based on generative adversarial networks. arXiv preprint arXiv:180603384. 2018.

17. Esteban C, Hyland SL, Rätsch G. Real-valued (medical) time series generation with recurrent conditional gans. arXiv preprint arXiv:170602633. 2017.

18. Hashemi AS, Soliman A, Lundström J, Etminani K. Domain Knowledge-Driven Generation of Synthetic Healthcare Data. In: Maria H, Madeleine B, Stefano B, Lina N, Inge Cort M, Sylvia P, et al., editors. The 33rd Medical Informatics Europe Conference, MIE2023, Gothenburg, Sweden, 22-25 May, 2023; 2023; Amsterdam: IOS Press; 2023. p. 352-3.

19. Ogwel B, Mzazi VH, Awuor AO, Otieno G, Ogolla S, Nyawanda BO, et al. A quarter-century of synthetic data in healthcare: Unveiling trends with structural topic modeling. *Digit Health*. 2025 Jan-Dec;11:20552076251404530. PMID: 41346942. doi: 10.1177/20552076251404530.

20. Alaa A, Van Breugel B, Saveliev ES, Van Der Schaar M, editors. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. *International conference on machine learning*; 2022: PMLR.

21. Stenger M, Leppich R, Foster I, Kounev S, Bauer A. Evaluation is key: a survey on evaluation measures for synthetic time series. *Journal of Big Data*. 2024 2024/05/07;11(1):66. doi: 10.1186/s40537-024-00924-7.

22. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019 Jan;25(1):44-56. PMID: 30617339. doi: 10.1038/s41591-018-0300-7.

23. Jordon J, Szpruch L, Houssiau F, Bottarelli M, Cherubin G, Maple C, et al. Synthetic Data -- what, why and how?2022.

24. Shabani M, Marelli L. Re-identifiability of genomic data and the GDPR: Assessing the re-identifiability of genomic data in light of the EU General Data Protection Regulation. *EMBO Rep*. 2019 Jun;20(6). PMID: 31126909. doi: 10.15252/embr.201948316.

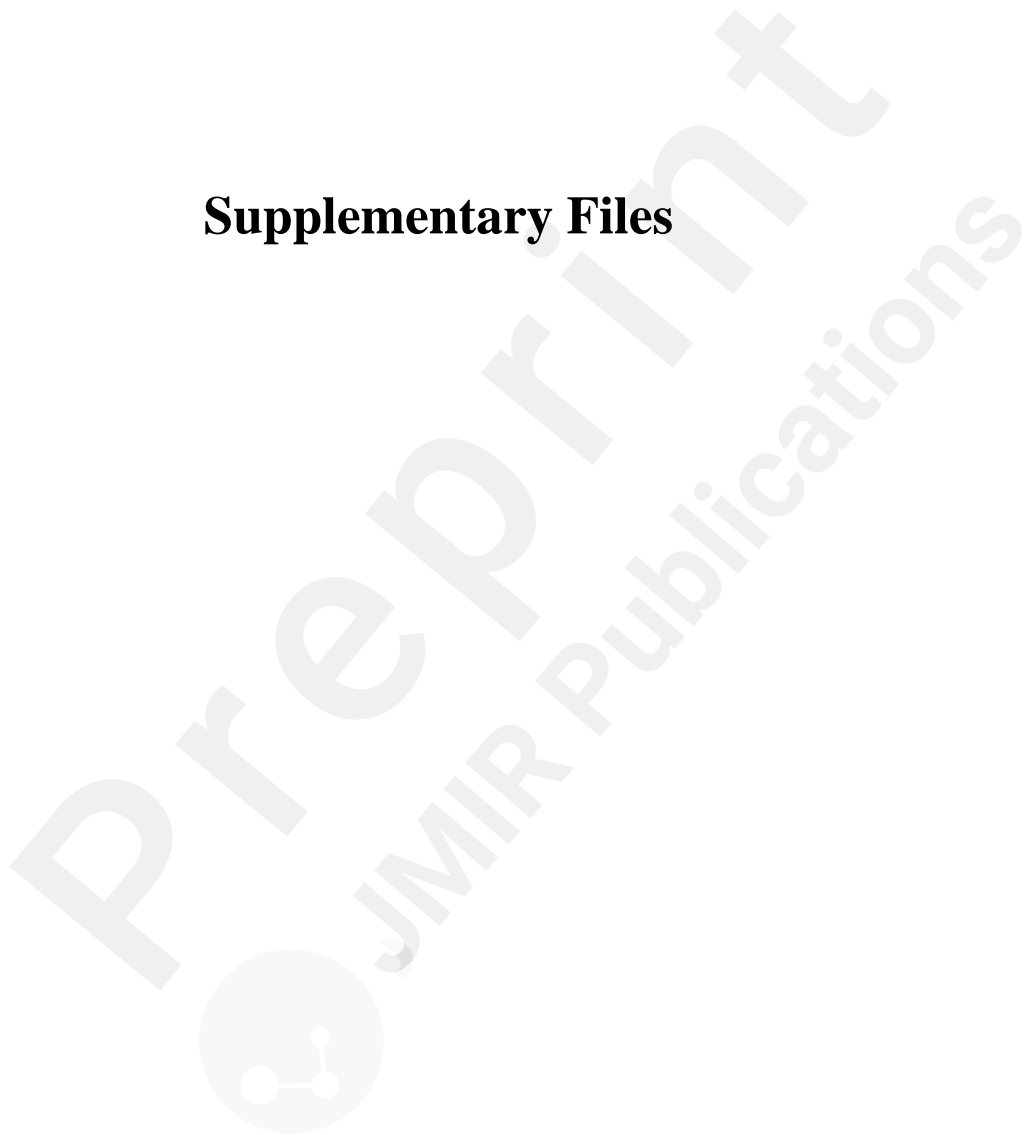
25. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*. 2019 2019/01/01;25(1):44-56. doi: 10.1038/s41591-018-0300-7.

26. Kaabachi B, Despraz J, Meurers T, Otte K, Halilovic M, Kulynych B, et al. A scoping review of privacy and utility metrics in medical synthetic data. *npj Digital Medicine*. 2025 2025/01/27;8(1):60. doi: 10.1038/s41746-024-01359-3.

27. Liaw ST, Guo JGN, Ansari S, Jonnagaddala J, Godinho MA, Borelli AJ, et al. Quality assessment of real-world data repositories across the data life cycle: A literature review. *J Am Med Inform Assoc*. 2021 Jul 14;28(7):1591-9. PMID: 33496785. doi: 10.1093/jamia/ocaa340.

28. Pilgram L, Ko H, Tung A, El Emam K. Protecting patient privacy in tabular synthetic health data: a regulatory perspective. *npj Digital Medicine*. 2025 2025/11/28;8(1):732. doi: 10.1038/s41746-025-02112-0.

Supplementary Files



Multimedia Appendixes

Overview of all HealthData4EU projects, their objectives, data modalities, use cases, and methodological challenges.
URL: <http://asset.jmir.pub/assets/8437d456752fc17aeb87051d57ee2181.docx>