ScienceDirect

Contents lists available at sciencedirect.com
Journal homepage: www.elsevier.com/locate/jval

Themed Section: Artificial Intelligence in Health Economics and Outcomes Research

Roles of Artificial Intelligence–Based Synthetic Data in Health Economics and Outcomes Research

Tim C. Lai, BPharm, Surachat Ngorsuraches, PhD

ABSTRACT

Objectives: We aim to raise awareness of potential applications of synthetic data within the health economics and outcomes research (HEOR) community.

Methods: We provide a concise overview of synthetic data, including data generation and types. We then discuss 3 major data-associated challenges and how synthetic data may be used to address them. Finally, we discuss data utility, privacy protection, potential concerns of its applicability, and future research direction.

Results: The use of synthetic data is an alternative privacy protection technique to enhance data availability, strengthen the robustness of findings for underrepresented populations, and alleviate data insufficiency issues in rare disease research. More studies are needed to explore synthetic data use and address data challenges in HEOR studies. Furthermore, the development of an evaluation framework is encouraged to better support the integration of synthetic data into the HEOR field.

Conclusions: Synthetic data provide a unique opportunity to overcome data-related challenges in HEOR.

Keywords: data privacy, generative artificial intelligence, privacy enhancing technology, synthetic data.

VALUE HEALTH. 2025; 28(11):1690–1695

Highlights

- The article provides an overview of synthetic data and its potential applications in health economics and outcomes research.
- Synthetic data can improve data availability, strengthen the robustness of findings for underrepresented populations, and alleviate data insufficiency issues in rare disease research. Developing an evaluation framework may facilitate synthetic data adoption in health economics and outcomes research.
- Integrating synthetic data into the evidence-generation process leads to more robust value assessment and timely decision making.

Introduction

Traditionally, health technology assessment (HTA) organizations have considered data from randomized controlled trials as the gold standard because of their robust internal validity.[1] To address the concern of generalizability (ie, external validity), HTA bodies have used real-world evidence to supplement randomized controlled trials.[1,2] However, a recent survey concluded that data insufficiency or unavailability remains a primary challenge during value assessments, particularly for newer therapies or rare diseases.[3] Because the number of health technologies is rapidly growing, there is an urgent need to find approaches to improve data availability and quality for robust and timely evaluations.[3]

The availability of medical data is restricted by data protection regulations, such as the US Health Insurance Portability and Accountability Act[4] and the European General Data Protection Regulation,[5] because of privacy concerns. A potential way to address privacy concerns while enabling the sharing of health-related data beyond its initial collection is the use of synthetic data. The term synthetic data refers to artificially generated data that preserve statistical properties and mimic real data structures without endangering individual privacy.[6–8] The use of synthetic data to address scientific inquiries is not new. Researchers have proposed[9] and applied[10] synthetic data in population-based surveys such as the US Census Bureau's American Community Survey[11] to protect sample privacy before making data publicly available. Classical approaches to generating synthetic data involve estimating the data distribution and deriving new samples. These methods mainly rely on parametric estimation with prior assumptions about the shape of data distribution, which may pose challenges and limit their use in estimating data with complex correlational structures, as commonly seen in medical records.[12]

The utility of synthetic data has become more promising with the emergence of artificial intelligence and machine learning (AI/ML). The AI/ML models can capture more complex data structures and generate synthetic data with high fidelity.[13] Although the health economics and outcomes research (HEOR) community has widely discussed the roles of AI/ML models in terms of their utilization of outcome prediction, feature selection, and economic modeling,[14] we found little discussion about AI/ML models in generating synthetic data for HEOR. After the release of a large language model—ChatGPT (Generative Pre-trained Transformer) from OpenAI in late 2022, the applications of AI have marked a paradigm shift in which scholars start to explore AI/ML models' generative capabilities. In HEOR, a recent report by the ISPOR Working Group discussed the potential utility of generative AI in HTA.[15] However, the report primarily focuses on using large language models to facilitate literature review, unstructured clinical notes summary, and economic model development. We argue that the advanced models used in generative AI also hold great

potential for generating synthetic data that can be used to address major challenges encountered in HEOR.

This article aims to raise awareness of synthetic data and encourage our fellow researchers to explore its potential roles and applications in HEOR. We structure this article as follows. First, we briefly overview the synthetic data generation methods (not limited to AI/ML models) and synthetic data type. Then, we explore the potential applications of synthetic data in addressing some current challenges in HEOR. Finally, we discuss data utility and privacy protection and provide recommendations for future research.

Overview of Synthetic Data

In the healthcare context, data can be categorized into either structured, such as claims data (ie, with rows and columns), or unstructured, such as physician notes or images; our following discussions primarily refer to the synthesis of structured data.

Synthetic data generation approaches

Based on sources of input, synthetic data generation approaches can be classified into (1) knowledge-driven, (2) data-driven, and (3) hybrid approaches.[16,17]

Knowledge-driven. Data are generated through simulations of disease progression or rules of care pathways based on known theory or expert knowledge. A prominent example developed in this category is Synthea, an open-source software designed for generating electronic health records.[18] Similar to microsimulation models familiar to the HEOR community, the parameters of Synthea models are typically sourced from domain experts or published studies.[19] However, Synthea is primarily designed to generate individual-level health records instead of outcome estimations, as a microsimulation study usually does.[3] Although this approach could be especially useful when real data are difficult to access, a key limitation is that it requires considerable manual effort to process and consolidate meaningful information to formulate models.[12]

Data-driven. It requires the acquisition of real data to train statistical or AI/ML models, and then the models can be used to generate synthesized data. Multiple imputation is one of the data-driven approaches that create synthetic data based on the statistical properties of real data.[9,10] Assumptions of parametric distributions are usually required when considering statistical properties for model training. However, it becomes challenging to consider the complex correlational structure of variables, which is commonly observed in the medical field, limiting generative models' widespread adoption of this approach.[12] On the other hand, AI/ML models can capture more complex relationships between variables, resulting in a higher resemblance of real data than models generated simply based on statistical properties.[7] The application of AI/ML, especially using advanced generative models, such as Recurrent Neural Network[20] or Generative Adversarial Network (GAN),[21] to generate synthetic data is currently an active reach area. Although data-driven methods can reduce considerable manual efforts compared with the knowledge-driven approach and achieve high resemblance with advanced AI/ML techniques, the approach is restricted by the availability of real data, and the synthesized data may inherit or amplify bias originating from the input data.[12]

Hybrid. This approach leverages both real data and expert knowledge to improve model performance. Specifically, models learn the structure of real data, whereas expert knowledge helps refine the details for better realism. This combination is especially useful when real data are limited or the data set is small. For instance, 1 study[22] examined whether incorporating human advice could improve the quality of data generated by a GAN model. The results showed that adding human guidance enhanced the data quality and allowed the model to learn effectively from a small data set. The synthetic data generated through this method outperformed other GAN-based generative models. However, the authors also cautioned that bad advice from humans could negatively affect the quality of synthetic data.

Types of synthetic data

Figure 1 illustrates data types according to data composition. In general, synthetic data sets can be classified into 3 main categories[23]: (1) Fully synthesis: data is created entirely de novo without containing any real data. (2) Partial or semi-synthesis: only variables with a high risk of identity exposure are replaced by synthetic data. (3) Hybrid: data contains both original and synthetic data, in which original data are paired with the nearest synthetic data according to some mathematical distance (eg, Hellinger distance).[24]

Potential Applications of Synthetic Data to Address HEOR Challenges

Although the roles and applications of synthetic data in HEOR are yet to be explored, we discuss some urgent challenges encountered in HEOR and provide examples of how synthetic data may help address these challenges.

Challenge 1: Limited data accessibility for decision making

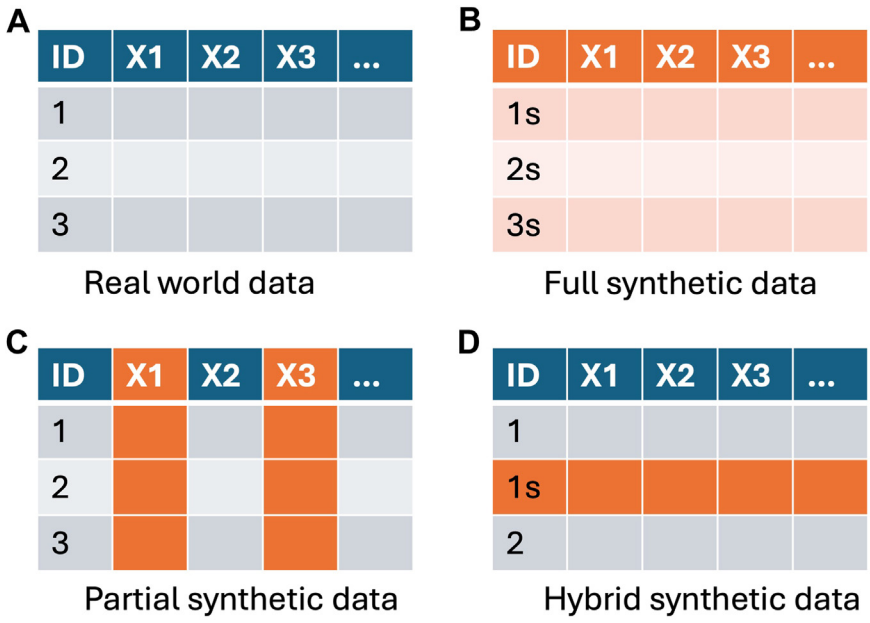
HTA organizations find that the root cause of many challenges for value assessment is a lack of available data.[3] Several efforts have been made to enhance data accessibility. For example, regulators in Europe[25] and Canada[26] require pharmaceutical companies to make the health data submitted for drug approval publicly available so that other researchers can replicate the studies or potentially lead to new insights. Also, medical journals strongly encourage researchers to make their data publicly available for other researchers to replicate the studies.[27] However, real-world individual data are often subject to strict privacy regulations; hence, data should be deidentified before releasing to the public. Generally, data deidentification methods usually err on the side of caution to prevent reidentification risk, significantly eroding data utility.[28]

Synthetic data, by design, aims to mimic real data without revealing individual information, making it possible to simultaneously address privacy concerns while providing publicly available data with analytical value. This creates opportunities for researchers to access a broader range of data for more comprehensive analyses and inform policymakers for more robust decision making. For instance, the National Disease and Registration Service at NHS England has generated the synthetic cancer registry data, Simulacrum,[29] which allow researchers to conduct hypothesis testing and cost-effectiveness research for various types of cancer.[28]

Challenge 2: Lack of subgroup information to support equity-focused value assessment

Identifying variety in treatment effects among subgroup populations is crucial for conducting equity-focus value assessments, such as the distributional cost-effectiveness analysis (DCEA). The DCEA is used to evaluate healthcare interventions' equity impact through the expansion of the traditional CEA framework, examining how costs and health benefits are distributed among

Figure 1. Illustration of synthetic data type.



different subgroups within a population.[30] One of the major obstacles to the consistent application of DCEA is the lack of sufficient subgroup data to conduct robust analyses.[31] Specifically, DCEA requires data stratified by equity-relevant variables, such as socioeconomic deprivation, age, sex, and race/ethnicity. However, clinical trials usually have strict inclusion and exclusion criteria, and data are often underrepresentative of certain demographic subgroups, such as women, older adults, and racial/ethnic minorities.[32,33] This lack of data diversity can limit the generalizability of study findings.

Synthetic data may fill the gap by augmenting data to reduce data bias from underrepresentation, enabling a more comprehensive understanding of the equity implications of healthcare resource allocation. Specifically, by generating synthetic data that augment underrepresentative groups, researchers can better understand the robustness of their findings and evaluate whether study results apply to more diverse populations. For instance, Juwara et al[34] proposed the synthetic minority augmentation (SMA) approach to generate synthetic individuals in the minority groups, aiming to rebalance samples to allow the results of evaluations to be more generalizable to target populations. The results showed that SMA could successfully reconstruct low- to median-biased data sets to produce results close to the ground truth. However, we should be cautious that synthetic data generation approaches, such as SMA, should not be taken as alternatives to the continuous efforts on diversifying clinical trial enrollment[35] because Juwara et al[34] provided evidence that the advantage of utilizing SMA or other generative models to augment underrepresentative data is not apparent when the original sample is highly biased (ie, missing high proportion of underrepresentative subgroups).

Challenge 3: Sparce data and scarce evidence in the rare disease area

Value assessment in rare disease areas faces significant challenges because of the data paucity, stemming from small patient populations and the inherent heterogeneity of disease conditions.[36,37] Outcome evaluation for rare diseases is also hampered by difficulties in recruiting sufficient clinical trial participants or retrieving real-world data. Although numerous challenges exist, sustainable solutions are rarely identified.[37] Scholars have recommended that there is a need for a more flexible and transparent evaluation framework and the necessity to conduct more studies to address publication bias.[37]

Synthetic data can be generated to reflect a variety of scenarios, patient profiles, and outcomes, even when real-world data are limited or unavailable. A pragmatic example is conducted by Sliman et al,[38] in which the authors developed a synthetic data generation framework for uveitis, a rare disease in ophthalmology, using the hybrid approach. Specifically, the authors first generated an original data set based on known statistical properties of disease. Second, the original data were validated by ophthalmologists (experts' input) to serve as base data to be used for generative model training; finally, a GAN-based model, MedWGAN, is used to generate the synthetic data. The synthetic data set is openly available and expected to serve as a foundation for future model improvement. Researchers are also encouraged to compare the synthetic data with real data, if available, to generate more reliable data sets. Another application is demonstrated by a research group from Italy[39] that trained a conditional GAN (cGAN) using data from patients with a rare type of blood cancer—myelodysplastic syndromes. The study showed that synthetic data could (1) retain statistical properties and complex interactions between features, (2) replicate estimates of treatment effects, and (3) overcome the lack or imbalance of information of real data. Additionally, the research group developed a website that allows scholars or clinicians to generate cohorts of up to 10 000 synthetic patients for results replication or other research purposes without compromising the privacy of real patients. The 2 pragmatic examples provide potential directions for researchers to address data scarcity, improve more transparent evaluation, and mitigate publication bias for rare disease research.

Discussion

Although the capability of generative AI is still evolving, the potential applications of synthetic data have been widely discussed

in clinical and data science communities. Additionally, the US FDA is actively investigating the potential use of synthetic data as supporting evidence (eg, synthetic control arms) for regulatory approval of medical interventions.[40] The recent report from the ISPOR Working Group[15] only briefly discussed using synthetic data to mitigate model bias and privacy protection, and we believe its role in HEOR is underestimated and yet to be explored. Hence, we explore 3 major challenges widely discussed in the HEOR community and provide pragmatic examples to demonstrate how synthetic data may help address these challenges.

Although the primary goal of using synthetic data is to improve data availability without compromising privacy, it is crucial to ensure that the synthetic data fulfill their intended purpose: serving as a reliable proxy for real data. However, the data synthesis process often involves a trade-off between data utility and the strength of privacy protection that inherently reduces utility.[7] Unfortunately, there is an absence of standardized frameworks or guidance throughout this process. Researchers may use various measures depending on the methodologies used and the specific goal of their study.[12]

Data Utility

Current literature often uses "utility[13,24]" and "fidelity[12,41,42]" interchangeably to describe the extent to which a synthetic data set can act as a proxy for real data, focusing on the statistical resemblance between the synthetic and original data sets. Specifically, data utility (or fidelity) is typically assessed through statistical properties, such as univariate measures (eg, means and variances), bivariate correlations, or multivariate distribution shapes that mirror the characteristics of the original data.[12] However, some scholars distinguish "utility," which we refer to as practical utility hereafter to avoid confusion, as a subjective measure tied to the data's usefulness in practical applications[12](Fig. 2). For instance, a synthetic data set might show moderate fidelity in replicating statistical patterns of real data but offer high practical utility for tasks such as ML model training.[13] The way practical utility is measured can vary depending on the use case. For example, when synthetic data are used for ML model training, performance metrics, such as the F1 score, could be used to compare models trained on real versus synthetic data. In contrast, if the data are intended for hypothesis generation, such as synthetic cancer registry data, statistical estimates might be more appropriate for practical utility evaluation.

Privacy (Reidentification Risk) Protection

Privacy protection receives little attention, and there is currently no formal guidance governing synthetic data,[43] potentially because researchers generally presume that synthetic data have inherent privacy assurances.[12] However, a study showed that if a model is overfitted to the original data, a synthetic record can still be linked to a real person.[44] Privacy risks encompass identity disclosure, membership disclosure, and attribute disclosure. Identity disclosure occurs when an individual's identity in synthetic data is directly linked to the real data,[45] membership disclosure reveals whether an individual was part of the original data set used for synthetic data training,[46] and attribute disclosure infers sensitive information about an individual based on other attributes.[45] It is recommended to perform disclosure risk assessment on a regular basis on synthetic data to ensure that the generative models do not overfit.[45,46] Data generators may integrate privacy protection risk measurement as a loss function[45] or differential privacy mechanism[47,48] in generative models to mitigate privacy disclosure risk. Additionally, it is recommended to establish explicit guidelines and regulations for synthetic data sharing and utilization to ensure that it is handled with the same level of care and protection as real-world data.

Practical Concerns and Recommendations for Future Work

It is worth noticing that regardless of the level of fidelity, synthetic data may not be a standalone solution but a tool that

Figure 2. Illustration of fidelity versus (practical) utility.

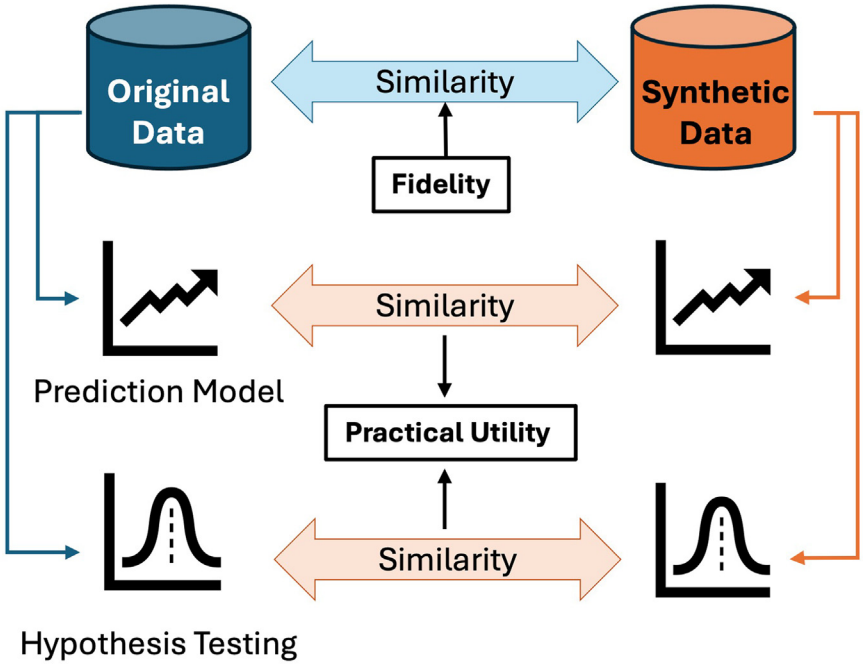


Table 1. Example of framework.

Stakeholder considerations		PET approaches			
Criteria	Weight	HIPAA*	Full synthetic	Partial synthetic	Hybrid
Privacy					
Utility					
Cost					
Trust					

Note. Framework adopted from El Emam et al.[7]
HIPAA indicates US Health Insurance Portability and Accountability Act; PET, privacy-enhancing technologies.
*General Data Protection Regulation in the European context, representing the protection mechanism for real-world data.

requires careful application. That is, synthetic data may be considered as a complementary rather than a primary source of evidence. Studies have shown that naively treating a single synthetic data set as a direct substitute for real data cannot consistently yield reliable effect estimates.[49,50] Because only data owners can validate results using the original data, it is recommended to provide data users with multiple synthetic data sets to enable the replication and cross-verification of findings.[49,50] For data users, it would be a good practice to investigate data validation studies for the intended utilization.

Given the lack of formal guidance for evaluating and adopting synthetic data, developing a framework to determine its viability and the extent to which it can address data-related challenges in HEOR is warranted. El Emam et al[7] indicate 4 key factors to be considered when implementing privacy-enhancing technologies: (1) privacy, (2) utility, (3) cost, and (4) trust. (Table 1)

We have already explored privacy and utility in previous discussions. According to El Emam et al,[7] Cost encompasses 2 aspects: (1) implementation cost, which refers to the expenses associated with adopting privacy-enhancing technologies, such as US Health Insurance Portability and Accountability Act compliance or synthetic data generation, and (2) operational cost, which includes the expenses tied to infrastructure and data processing. Trust plays a critical role in determining whether individuals are willing to continuously engage within the system. For example, in healthcare settings, when patients are skeptical about how their information might be used, they may avoid care, resort to self-medication, or withhold important details during interactions with providers. Evaluating these criteria will not be a straightforward task because stakeholders may assign varying weights to each. Multicriteria decision analysis[51,52] could be valuable in facilitating decision making by accommodating these diverse perspectives.

Limitations

There are several limitations in this article. First, although acknowledging that there are different types of structured data, such as cross-sectional, longitudinal, or time-series data, we did not explicitly distinguish between them in our discussions. Differentiating these data types could be significant during the synthetic data generation process because a generative model may be suited to only specific structures.[12] Nevertheless, the core concepts we focused on remain broadly applicable across these variations. Second, because our aim is to introduce synthetic data to readers unfamiliar with the topic, we intentionally excluded complex mathematical details in sections addressing data generation, utility, and privacy risk assessments. Readers seeking methodological depth or guidance on operationalizing data generation and quality assessment can consult referenced articles for further exploration. Finally, we highlighted only a selection of use cases and generative models to illustrate our discussion points without aiming to provide an exhaustive list. Readers who want to explore more can seek studies that offer comprehensive reviews,[12] use cases,[6] and tools.[53]

Conclusions

Synthetic data provide a unique opportunity to overcome data-related challenges in HEOR. It can enhance data availability for decision making, strengthen the robustness of the findings for underrepresented populations, and alleviate data insufficiency issues in rare disease research. More studies to explore the use of synthetic data to address data challenges in HEOR studies are needed. Also, to better support the integration of synthetic data into the HEOR field, the development of evaluation framework is encouraged. As a result, synthetic data have the potential to become a valuable tool for decision making.

Author Disclosures

Author disclosure forms can be accessed below in the Supplemental Material section.

Supplemental Material

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.jval.2025.04.2157.

Article and Author Information

Accepted for Publication: April 15, 2025

Published Online: May 29, 2025

doi: https://doi.org/10.1016/j.jval.2025.04.2157

Author Affiliations: Health Outcomes Research and Policy, Harrison College of Pharmacy, Auburn University, Auburn, Alabama, USA (Lai, Ngorsuraches).

Correspondence: Surachat Ngorsuraches, PhD, Health Outcomes Research and Policy, Harrison College of Pharmacy, Auburn University, Auburn, AL, USA. Email: szn0053@auburn.edu

Authorship Confirmation: All authors certify that they meet the ICMJE criteria for authorship.

Funding/Support: Tim C. Lai receives the Predoctoral Fellowship from the Pharmaceutical Research and Manufacturers of America (PhRMA) Foundation. Tim C. Lai had received Health Equity Research Fellowship from the Center for Innovation and Value Research (formerly Innovation and Value Initiative). Surachat Ngorsuraches receives grants from the

PhRMA Foundation, the Patient-Centered Outcomes Research Institute, and the National Institutes of Health.

Role of the Funder/Sponsor: The funders had no role in the preparation, review, approval of the manuscript, and decision to submit the manuscript for publication.

Acknowledgment: The authors greatly appreciate the 3 anonymous reviewers providing invaluable suggestions and guidance to improve this work.

Disclosure for Using AI Tool: The authors used Grammarly to check grammar and Microsoft Copilot to modify sentences to enhance readability.

REFERENCES

1. Monti S, Grosso V, Todoerti M, Caporali R. Randomized controlled trials and real-world data: differences and similarities to untangle literature data. *Rheumatology*. 2018;57(suppl 7):vii54–vii58.
2. Turner AJ, Sammon C, Latimer N, et al. Transporting comparative effectiveness evidence between countries: considerations for health technology assessments. *Pharmacoeconomics*. 2024;42(2):165–176.
3. Hogervorst MA, Vreman RA, Mantel-Teeuwisse AK, Goettsch WG. Reported challenges in health technology assessment of complex health technologies. *Value Health*. 2022;25(6):992–1001.
4. US Department of Health and Human Services. Health information privacy. https://www.hhs.gov/hipaa/for-professionals/index.html; Updated July 19, 2024. Accessed November 25, 2024.
5. European Union. General data protection regulation. https://gdpr-info.eu/; Updated April 5, 2016. Accessed November 25, 2024.
6. James S, Harbron C, Branson J, Sundler M. Synthetic data use: exploring use cases to optimise data utility. *Discov Artif Intell*. 2021;1(1):15.
7. El Emam K, Mosquera L, Hoptroff R. Practical synthetic data generation: balancing privacy and the broad availability of data. O'Reilly Media. https://cdn.ttgtmedia.com/rms/pdf/Practical_Synthetic_Data_Generation.pdf; Published 2020. Accessed February 15, 2025.
8. Drechsler J, Haensch A-C. 30 years of synthetic data. *Stat Sci*. 2024;39(2):221–242.
9. Rubin DB. Statistical disclosure limitation. *J Off Stat*. 1993;9(2):461–468.
10. Raghunathan TE, Reiter JP, Rubin DB. Multiple imputation for statistical disclosure limitation. *J Off Stat*. 2003;19(1):1.
11. Freiman M, Lauger A, Reiter J. Data Synthesis and Perturbation for the American Community Survey at the US Census Bureau. US Census Bureau. https://www2.census.gov/adrm/CED/Papers/CY17/2017-09-FreimanLaugerReiter-ACSSynthesisPerturbation.pdf; Published 2017. Accessed February 15, 2025.
12. Murtaza H, Ahmed M, Khan NF, Murtaza G, Zafar S, Bano A. Synthetic data generation: state of the art in health care domain. *Comput Sci Rev*. 2023;48:100546.
13. Hittmeir M, Ekelhart A, Mayer R. On the utility of synthetic data: an empirical evaluation on machine learning tasks. Paper presented at: Proceedings of the 14th International Conference on Availability, Reliability and Security. Canterbury, UK https://doi.org/10.1145/ 3339252.3339281; August 26–29, 2019.
14. Padula WV, Kreif N, Vanness DJ, et al. Machine learning methods in health economics and outcomes research-The PALISADE checklist: a good practices report of an ISPOR task force. *Value Health*. 2022;25(7):1063–1080.
15. Fleurence RL, Bian J, Wang X, et al. Generative AI for health technology assessment: opportunities, challenges, and policy considerations-an ISPOR working group report. *Value Health*. 2025;28(2):175–183.
16. Surendra H, Mohan H. A review of synthetic data generation methods for privacy preserving data publishing. *Int J Sci Technol Res*. 2017;6(3):95–101.
17. Aggarwal CC, Yu PS. *A General Survey of Privacy-Preserving Data Mining Models and Algorithms*. Berlin, Germany: Springer; 2008.
18. Walonoski J, Hall D, Bates KM, et al. The "coherent data set": combining patient data and imaging in a comprehensive, synthetic health record. *Electronics*. 2022;11(8):1199.
19. The Office of the National Coordinator for Health Information Technology. Synthetic health data generation to accelerate patient-centered outcomes research. https://www.healthit.gov/topic/scientific-initiatives/pcor/synthetic-health-data-generation-accelerate-patient-centered-outcomes; Updated April 28, 2022. Accessed November 25, 2024.
20. Mosquera L, El Emam K, Ding L, et al. A method for generating synthetic longitudinal health data. *BMC Med Res Methodol*. 2023;23(1):67.
21. Qian Z, Callender T, Cebere B, Janes SM, Navani N, van der Schaar M. Synthetic data for privacy-preserving clinical risk prediction. *Sci Rep*. 2024;14(1):25676.
22. Dhami DS, Das M, Natarajan S. Knowledge intensive learning of generative adversarial networks. CEUR Workshop Proceedings, 6 https://ceur-ws.org/Vol-2657/short1.pdf; 2020.
23. Gonzales A, Guruswamy G, Smith SR. Synthetic data in health care: a narrative review. *PLoS Digit Health*. 2023;2(1):e0000082.
24. El Emam K, Mosquera L, Fang X, El-Hussuna A. Utility metrics for evaluating synthetic health data generation methods: validation study. *JMIR Med Inform*. 2022;10(4):e35734.
25. European Medicines Agency. External guidance on the implementation of the European Medicines Agency Policy on the publication of clinical data for medicinal products for human use. version 1.4. https://www.ema.europa.eu/en/clinical-data-publication/support-industry-clinical-data-publication/external-guidance-implementation-european-medicines-agency-policy-publication-clinical-data-medicinal-products-human-use; Updated November 9, 2018. Accessed November 25, 2024.
26. Government of Canada. Guidance document on public release of clinical information. https://www.canada.ca/en/health-canada/services/drug-health-product-review-approval/profile-public-release-clinical-information-guidance.html; Updated March 12, 2019. Accessed November 25, 2024.
27. Taichman DB, Sahni P, Pinborg A, et al. Data sharing statements for clinical trials: a requirement of the International Committee of Medical Journal Editors. *JAMA*. 2017;317(24):2491–2492.
28. El Emam K. Accelerating AI with synthetic data-generating data for AI projects. https://www.nvidia.com/content/dam/en-zz/Solutions/deep-learning/resources/accelerating-ai-with-synthetic-data-ebook/accelerating-ai-with-synthetic-data-nvidia_web.pdf; Published 2020. Accessed November 25, 2024.
29. Health Data Insight. The Simulacrum. https://simulacrum.healthdatainsight.org.uk/. Accessed November 25, 2024.
30. Asaria M, Griffin S, Cookson R. Distributional cost-effectiveness analysis: a tutorial. *Med Decis Mak*. 2016;36(1):8–19.
31. Meunier A, Longworth L, Kowal S, Ramagopalan S, Love-Koh J, Griffin S. Distributional cost-effectiveness analysis of health technologies: data requirements and challenges. *Value Health*. 2023;26(1):60–63.
32. Schwartz AL, Alsan M, Morris AA, Halpern SD. Why diverse clinical trial participation matters. *N Engl J Med*. 2023;388(14):1252–1254.
33. Alegria M, Sud S, Steinberg BE, Gai N, Siddiqui A. Reporting of participant race, sex, and socioeconomic status in randomized clinical trials in general medical journals, 2015 vs 2019. *JAMA Netw Open*. 2021;4(5):e2111516.
34. Juwara L, El-Hussuna A, El Emam K. An evaluation of synthetic data augmentation for mitigating covariate bias in health data. *Patterns*. 2024;5(4):100946.
35. Tackling biases in clinical trials to ensure diverse representation and effective outcomes. *Nat Commun*. 2024;15(1):1407.
36. Nestler-Parr S, Korchagina D, Toumi M, et al. Challenges in research and health technology assessment of rare disease technologies: report of the ISPOR rare disease Special Interest Group. *Value Health*. 2018;21(5):493–500.
37. Grand TS, Ren S, Hall J, Åström DO, Regnier S, Thokala P. Issues, challenges and opportunities for economic evaluations of orphan drugs in rare diseases: an umbrella review. *Pharmacoeconomics*. 2024;42(6):619–631.
38. Sliman H, Megdiche I, Alajramy L, et al. MedWGAN based synthetic dataset generation for uveitis pathology. *Intell Syst Appl*. 2023;18:200223.
39. D'Amico S, Dall'Olio D, Sala C, et al. Synthetic data generation by artificial intelligence to accelerate research and precision medicine in hematology. *JCO Clin Cancer Inform*. 2023;7:e2300021.
40. US Food and Drug Administration. Addressing the limitations of medical data in AI. https://www.fda.gov/medical-devices/medical-device-regulatory-science-research-programs-conducted-osel/addressing-limitations-medical-data-ai. Accessed November 25, 2024.
41. Tucker A, Wang Z, Rotalinti Y, Myles P. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *npj Digit Med*. 2020;3(1):147.
42. Mendelevitch O, Lesh MD. Fidelity and privacy of synthetic medical data. arXiv Preprint ArXiv:210108658. https://arxiv.org/abs/2101.08658; Published 2021. Accessed February 15, 2025.
43. Arora A, Arora A. Synthetic patient data in health care: a widening legal loophole. *Lancet*. 2022;399(10335):1601–1602.
44. Elliot M. Final report on the disclosure risk associated with the synthetic data produced by the SYLLS team. https://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/reports/2015-02%20-Report%20on%20disclosure%20risk%20analysis%20of%20synthpop%20synthetic%20versions%20of%20LCF_%20final.pdf. Accessed February 15, 2025.
45. El Emam K, Mosquera L, Bass J. Evaluating identity disclosure risk in fully synthetic health data: model development and validation. *J Med Internet Res*. 2020;22(11):e23139.
46. El Emam K, Mosquera L, Fang X. Validating a membership disclosure metric for synthetic health data. *JAMIA Open*. 2022;5(4):ooac083.
47. Ficek J, Wang W, Chen H, Dagne G, Daley E. Differential privacy in health research: a scoping review. *J Am Med Inform Assoc*. 2021;28(10):2269–2276.
48. Gong M, Xie Y, Pan K, Feng K, Qin AK. A survey on differentially private machine learning [review article]. *IEEE Comp Intell Mag*. 2020;15(2):49–64.
49. Van Breugel B, Qian Z, Van Der Schaar M. *Synthetic data, real errors: how (not) to publish and use synthetic data*. Honolulu, Hawai: Paper presented at: International Conference on Machine Learning; May 16, 2023.
50. El Emam K, Mosquera L, Fang X, El-Hussuna A. An evaluation of the replicability of analyses using synthetic health data. *Sci Rep*. 2024;14(1):6978.
51. Thokala P, Devlin N, Marsh K, et al. Multiple criteria decision analysis for health care decision making-an introduction: report 1 of the ISPOR MCDA emerging good practices task force. *Value Health*. 2016;19(1):1–13.
52. Marsh K, IJzerman M, Thokala P, et al. Multiple criteria decision analysis for health care decision making—emerging good practices: report 2 of the ISPOR MCDA emerging good practices task force. *Value Health*. 2016;19(2):125–137.
53. Pezoulas VC, Zaridis DI, Mylona E, et al. Synthetic data generation methods in healthcare: a review on open-source tools and methods. *Comp Struct Biotechnol J*. 2024;23:2892–2910.