Duke
MARGOLIS INSTITUTE for Health Policy

June 11, 2025

Synthetic Data Generation Using Generative AI to Support Biomedical Innovation: A Health Policy Perspective

POLICY BRIEF

AUTHORS

Rachele M. Hendricks-Sturrup
Nora Emmott
Maryam Nafie

Acknowledgements

We would like to thank Christina Silcox, Gillian Sanders Schmidler, and Jieyu Zhang as part of the Duke-Margolis Institute for Health Policy, as well as the Duke-Margolis 2023-2025 Real-World Evidence Collaborative, for their contributory thoughts toward the development, editing, and finalization of this policy brief. The authors also wish to thank Molly Shields and Patrick Rodriguez for their support with editorial support.

EXECUTIVE SUMMARY

Regulators and payers globally are exploring the potential of synthetic data as one of many applications of generative artificial intelligence (AI) to support both operations and decision-making in medical product development. This ongoing exploration has highlighted a current need to identify and develop practical considerations associated with synthetic data generation use in this context.[1,2] In this policy brief, we explore these areas through a discussion of current synthetic data management tools and best practices, ethical considerations for the generation and application of synthetic data, and regulatory developments to date. We recommend specific steps that regulatory stakeholders and practitioners may take to develop and describe regulatory fit-for-use synthetic data. Lastly, we offer a risk-based credibility assessment framework that could be helpful to for those managing synthetic data derived from generative AI applications.

BACKGROUND

Synthetic data can be described as artificial data that maintains certain, but not necessarily all, underlying data and statistical distributions, e.g., shape and variance, and structure, e.g., correlations among the attributes, of original datasets while maintaining data de-identification and preventing reidentification.[2,3] Modern approaches to synthetic data generation today involve the use of semi-supervised or unsupervised generative models, which range from powerful, e.g., language learning models or large language model (LLMs), generative adversarial networks (GANs), and variational autoencoders (VAEs), to moderate, e.g., sequential synthesis with decision trees or k-nearest neighbors algorithm, to simple, e.g., basic simulation.[3–10] Each model holds a certain capability to synthesize synthetic images and other digital data, e.g., real world data (RWD) that comprise electronic health records (EHR), medical images, genomic data, audio/image/video data, and clinical trial data, for likeness in underlying variation of original datasets, but without being an exact copy of the original dataset.

Synthetic data generation and use for medical product development purposes is of increasing interest among regulators globally. For instance, the European Union's (EU) European Medicines Agency (EMA) and the United Kingdom's (UK) Medicines and Healthcare products Regulatory Agency (MHRA)'s real-world evidence research service--the Clinical Practice Research Datalink--are seeking to leverage synthetic data as part of their goals to "maximise the generation, interoperability, use and exchange of data to support EU decision-making" and "validate applications of high fidelity synthetic data for sample size boosting and as external control arms," respectively.[11,12] Further, the EMA's draft reflection

paper on the use of AI in the medicinal product life cycle mentions that synthetic data is an instrument to "deploy differential privacy techniques" and for "increasing model performance."[13]

Canada's Drug Agency (CDA) mentions in their position statement on "The Use of Artificial Intelligence in the Generation and Reporting of Evidence" that synthetic data is "artificial data that is generated from original data and a model that is trained to reproduce the characteristics and structure of the original data, and are generated using AI-based methods, including machine learning algorithms and other approaches."[14] The paper also describes the potential of AI approaches to produce synthetic data and generate external control arms when, for example, it is unethical to include a placebo arm in a clinical trial as well as to predict clinical effectiveness in different subgroups.[14]

The United States (U.S.) Food and Drug Agency's (FDA) Digital Health and Artificial Intelligence Glossary defines "synthetic data" as "data that have been created artificially, e.g., through statistical modeling, computer simulation, so that new values and/or data elements are generated," also noting that "synthetic data are artificial data that are intended to mimic the properties and relationships seen in real patient data" and are "partially or fully generated using computational techniques."[1]

Practical Considerations for Synthetic Data Generation and Use

Despite definitions proposed among certain regulators like the FDA, to avoid the risk of misinterpretation, synthetic data should be described and interpreted on a contextual and purpose-driven basis in practice.[1] For instance, although there are similarities among terms that have been used to describe synthetic data, like "de-identified data," "digital twins," "virtual controls clinical trial data," and "synthetic cohorts;:" there are also important distinctions between these terms. For example, the term "synthetic cohorts," initially used to describe external comparator arms with patient-level data, can be conflated with data generated for purposes of modeling versus data that has undergone (possibly modest) privacy protection adjustments.[15] Context is therefore crucial to ensure shared language and nomenclature across various contexts involving specific handlers and end-users of synthetic data. While not every entity will agree on a single definition for synthetic data, we suggest that the term is best defined through its uses and applications to address a specific context-of-use (COU). As others have noted, regulatory pathways are important to stakeholders invested in real-world evidence (RWE) development and policy to ensure they can function within clear legal and practice frameworks.[16]

While not every entity will agree on a single definition for synthetic data, we suggest that the term is best defined through its uses and applications to address a specific context-of-use (COU).

Our assessment of the literature and professional engagement to date on the topic of synthetic data has led to our understanding that synthetic data use may fall into four areas of value within the health research landscape:

- a privacy enhancing technology (PET),
- a data science "sandbox" environment for research training or exploratory purposes, e.g., predictive modeling or to improve algorithms or machine learning (ML) workflows,
- mechanism to navigate legalities around data sharing and/or use (e.g. restrictive data use agreements, legal jurisdictions, etc.), and
- augmenting signals for underrepresented populations within and across datasets without compromising the structure and format of dataset(s).[3,17–22] We discuss these areas below, focusing on ethical and legal implications and subgroup analysis considerations.

Ethical and Legal Implications

Overall optimism around the use of synthetic data can be balanced with an acknowledgment of ethical, legal, and other similar considerations that accompany original RWD sets.[23] Given that synthetic data is considered a PET that obscures the identities or identifiable information about individuals within a dataset, individual-level privacy is one reality that synthetic data addresses.[24,25] However, like other methods of RWD de-identification, synthetic data can neither protect group-level privacy/discretion nor safeguard against group-level discrimination or population profiling, particularly in cases where certain groups of individuals might be unfairly targeted in tumultuous political or social settings.[25] Biomedical researchers, AI/ML users, and developers in the US have reported general interest in greater ethical and legal research on synthetic data and AI/ML.[26] For instance, like any generative model, synthetic data carries inherent risk in amplifying bias(es) within its original dataset(s). Such risks should be carefully balanced with any intended goals to leverage synthetic data to increase the quantity of row-level data within a dataset.

Purpose-driven validation studies, data curation processes to perform edit checks for plausibility, and additional evaluations are important to avoid potentially erroneous and/or discriminatory exclusion of seemingly anomalous biological relationships. These factors are essential in ensuring the accuracy of exposure and outcome measures and ensuring the fidelity of synthetic data against its original RWD set. For instance, there is potential for discordance between an individual's self-reported gender and sex assignment at birth within an original dataset, and synthetic data should preserve this discordance. If a disease under investigation is known to affect individuals of the male sex, e.g., prostate cancer, then synthetic data validation measures should involve the careful inclusion of individuals of any self-reported gender (including non-male genders) and their clinical phenotypes from the original dataset, as well as careful exclusion of individuals with genetically confirmed non-male sex from the original dataset.

Alongside these best practices is the reality that individuals with discordant gender identity and sex assignment at birth may be socially or politically targeted through data profiling, as they are part of a statistically small subgroup. Astute data governance processes and procedures based on scientific observations to date, like those described above, coupled with strong nondiscrimination protections is one way to address this issue or risk. Careful study design is also important, especially in precision medicine research settings that may advertently or inadvertently disclose genotype/phenotype discordance.[27] Researchers also must carefully design longitudinal studies such that synthetic data and prediction models do not miss or misattribute underrepresented subgroups or subpopulations due to low or varied EHR activity/documentation that could be biased or done in error.[28,29]

Data governance standards, best practices or tools are needed to support researchers and/or research ethics or institutional oversight boards, e.g., Institutional Review Boards (IRBs,) as they may encounter and/or oversee the generation, integration, or implementation of synthetic data or data generation tools in regulated and non-regulated health research. While this community has yet to either develop or broadly disseminate such best practices or tools, we propose that key ethical and legal questions below are considered locally and on a case-by-case basis.

Key Ethical and Legal Questions for Best Practices

- Should patients have an opportunity to consent (broadly, legally, or otherwise) to the specific use of their de-identified data for synthetic data generation and use for broad or specific purposes? Will patients understand when consenting to the use of their data to generate and/or become combined with synthetic data?
- How should patients become educated about synthetic data generation? Should patients require education at all if the same legal privacy protections apply to initial RWD collection, regardless of subsequent synthetic data generation and use?
- Is informed consent model language presently available for synthetic data generation and use? Should IRBs evaluate whether participant consent is needed and whether the risks outweigh the benefits to synthetic data generation and/or use?
- Would IRBs reviewing and overseeing observational studies be less concerned about synthetic data use since synthetic data obscures individual-level data?

- What practical assurances can researchers provide to research oversight boards and research participants today around the risk of re-identification, bias, and/or group-level discrimination?
- Is synthetic data generation and use appropriate for observational, non-interventional RWD only? Or might there be additional data sources or types to consider for broad impact as far as our ability to learn about medication benefits and risks?
- What are risks or ethical implications to disregarding potential uses of synthetic data, such as supporting rare disease patient communities that have a vested interest in leveraging their data to sufficiently power clinical studies, despite universal reidentification risks due to low population numbers?

These questions should be addressed prior to synthetic data generation, use, and/or implementation, especially within consequential health care and/or research scenarios. Instances where health care provider organizations or other data stewards generate and share synthetic datasets as a public service, or might apply synthetic data to address cases of missing data, are example scenarios. Engagement with rare disease advocates is highly encouraged to address specific concerns about inherent risks associated with reidentification and the risk-benefit tradeoffs to data disclosure for purposes of treatment development.

Efforts to address these questions should also be put into context with the existing legal landscape. For example, there is currently no specific policy or legal authority in the U.S. to supervise or ensure data protection and safe synthetic data use/processing and no national privacy standard or law to protect individuals contributing all forms of RWD, e.g., US Health Insurance Portability and Accountability Act (HIPAA) covered and non-covered data combined, used to generate robust synthetic datasets.[30–32] Until an authority or standard is codified, legal complaints that arise due to unanticipated hardships or harm caused by synthetic data use will be solely addressed in the court system and per existing laws or rules, such as the US Department of Health and Human Services' Office of Civil Rights rule concerning the nondiscrimination provision in Section 1557 of the Affordable Care Act (ACA).[33] This particular Section of the ACA could impose penalties on covered entities who rely on algorithm-enabled tools, which would include synthetic data generation tools, that result in discriminatory harms.

Efforts to govern synthetic data should also consider datatypes that may be inherently subject to privacy risks at either or both individual and group levels. For instance, with respect to genomic data, the probability that synthetic data can fully preserve the privacy of genomic data highly depends on both the technical nature, e.g., single nucleotide polymorphisms (SNPs), gene copy number, biomarkers, whole genome or exome sequencing data, RNA, viral genes, etc., and presentation, e.g., aggregated versus individual-level, of the data. Also, while synthetic data might be useful to preserve individual privacy and maintain statistical inferences regarding specific gene-disease associations, it may not ensure group-level privacy for individuals within an aggregated dataset who share genomic features, e.g., SNPs implicated in gene-disease associations. While it may be possible to sufficiently de-identify genomic data to accomplish research goals while ensuring data privacy, protecting individuals from potentially negative downstream effects following decisions made based on analyses of population-level genomic data continues to be a challenge in real-world settings.[34]

Subgroup Analysis Considerations

Some researchers have described the value of generating or using synthetic data across several clinical practice domains, like ophthalmology, palliative care, cardiology, endocrinology, and radiology (including cancer radiology) to better understand or predict disease onset and progression.[35–41] Yet, given the current reality that many individuals may struggle to receive diagnoses for either or both rare and common diseases, the result is often datasets with documentation biases and/or errors for certain patient subgroups, resulting in their under-representation within an analytical dataset.[29] Specifically, to avoid associated statistical impacts, patient subgroups showing as under-represented categories within an analytical dataset are often censored, leading to inaccurate predictions and likely unfair or biased assessments about those subgroups.

Synthetic data augmentation using diverse RWD that not only represent patient subgroups, but also contain annotations to document their diagnostic and treatment journeys within and outside of the health system, e.g., observational, non-interventional natural history, patient registry, and/or patient-generated data, is a potentially viable strategy to improve the reliability of synthetic data and reliably amplify underrepresented patient signals within a dataset.[42] Techniques like the Synthetic Minority Over-sampling Technique (SMOTE) or Synthetic Minority Augmentation with sequential synthesis using decision trees also can be useful to multiply an existing data pool or cohort, with the latter showing to be more effective than SMOTE to mitigate bias, improve effect estimation, and model performance.[9,43] Both techniques warrant further investigation to ensure that either or both the original RWD dataset or synthetic dataset are not limited, restricted, or costly, e.g., monetary cost, privacy cost, time constraint cost, perpetuating social/legal risk, etc., to improve the reliability of synthetic data concerning underrepresented or small patient subgroups.

Synthetic data can be used as a PET, allowing researchers to publicly share data without compromising individuals' privacy or data sharing agreements. The U.S. FDA recently sought to understand the full scope and potential of synthetic data as a PET for special populations, e.g., pregnant and lactating persons, as evidenced within a 2023 workshop co-convened by the FDA and Duke-Margolis.[44–46] During this workshop, participants shared that synthetic data generation and use carries inherent risks and tradeoffs, e.g., capacity to exacerbate health disparities and/or health care access issues due to algorithmic bias, that should be balanced, based on privacy community consensus metrics, with its propensity to serve as a privacy-promoting practice. Therefore, research partnerships and engagement with patient subgroups, with the intent to understand, contextualize, and convey engagement proceedings on the appropriateness and value of synthetic data generation and use is critical. Also, it will be important to establish scenarios in which synthetic data augmentation might be an inappropriate alternative to foregoing direct recruitment of individuals belonging to underrepresented or censored subgroups to amplify their statistical representation within a dataset.

Synthetic Data Management Tools & Best Practices and Target End-User Developments

Data Management Best Practices and Recommendations to Create Fit-for-Use Synthetic Data

RWD quality assurance frameworks and related tools, like the HARmonized Protocol Template to Enhance Reproducibility (HARPER) and ISPOR SUITABILITY Checklist, could be helpful starting points alongside other synthetic data quality assurance measures.[47,48] The HARPER framework provides a structured, standardized protocol for researchers focused on documentation and methodological transparency. The ISPOR SUITABILITY Checklist offers a set of criteria for assessing data fitness for purpose. Likewise, an assessment of the value, fidelity, representativeness, generalization, and resemblance of synthetic data relative to its original RWD set, as well as a quality evaluation of the synthetic data generation process itself, could be part of the quality evaluation criteria or analysis for specific models.[49] In cases where a synthetic dataset derived from RWD could be used to conduct hypothesis-generating studies, such studies should also require subsequent replication or validation using an RWD source that was not used to initially generate the synthetic dataset. This best practice measure can help ensure that all data correlations have been preserved and fidelity aspects met across original, transformed, and referenced datasets. Often, synthetic data can be integrated with real-world data, leading to data mixing which can make it difficult for researchers to identify synthetic data elements, leading to potential bias or errors in analysis. To help address this issue, we encourage transparency in data usage and analysis. This can be done through labeling processes that tag synthetic data in datasets.[16] Data cards, structured summaries that provide important information about a dataset including provenance, composition and intended use can also enhance transparency and clarity of datasets containing synthetic data.[50]

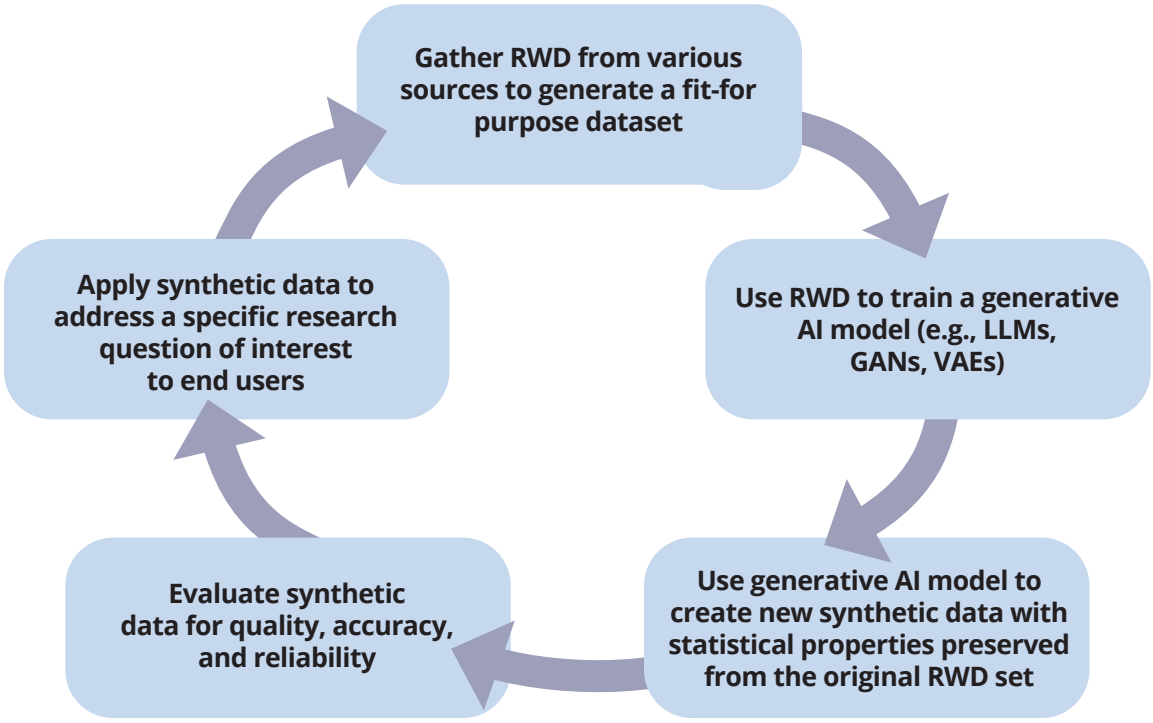
Alloza et al. 2023, recently added that multiple variables should be considered in the process of creating a reliable synthetic dataset including, but not limited to, the level at which data are synthesized, e.g., patient subgroups, intervention location, etc.[51] Failure to do so might impact the overall quality of the synthesized data.[51] We add that, in addition to first ensuring the accuracy and reliability of

original RWD sets, studying and exploring relationships within and between synthetic datasets are necessary to demonstrate, to prospective or target end-users, consistency in preserving statistical properties across multiple synthetic data transformations and controls against generative AI model drift. Also, to communicate assurances around the reliability of a synthetic dataset, it would be critical to provide end-users with pertinent metadata for an analytical RWD set and/or clinical trial datasets, as well as initially or primarily transformed synthetic data when requested. Lastly, we recommend that synthetic data generators become deeply familiar with any inherent characteristics of and limitations to an original RWD set, e.g. data missingness, lack of data standardization amongst different RWD sources, etc., and clinical trial data to ensure the reliability of the synthetic data derivative.[52]

Figure 1 summarizes general steps that can be taken to create a fit-for-use synthetic dataset using generative AI mechanisms. We acknowledge, however, that these steps are heavily dependent on transparency in the form of access to, or ability to, transfer originally linked or unlinked RWD sets. This transparency may prove a difficult endeavor, especially where rare disease and/or small subgroup datasets are either purposely censored and/or original data transfers are either impossible or prohibited, e.g., licensed access to data platforms). Therefore, whenever possible, data curators should implement strong data governance and provenance measures to convey both the accuracy and traceability, and thus reliability, of both original RWD sets and synthetic data derivatives. This could be in the form of, for example, standard operating procedures for data warehouses that involve documenting and preserving RWD linkage details across distributed data networks and obtaining mechanisms to request local access to original RWD across each data network. Such measures could be considered a best practice to support the uptake of synthetic data, as regulators like the FDA recommend:

- data curators consider potential linkages across data sources or additional collection to capture important confounders that are either unmeasured or imperfectly measured within any original sources
- verifying data against its original source.[53]

Figure 1 | Using Generative AI to Create a Fit-for-Use Synthetic Dataset



From a policy best practice standpoint, transparency in model-specific evaluation criteria or analysis methods is vitally important, especially in cases where synthetic data could become part of an evidentiary package for end-user (regulator, HTA, payer, researcher, etc.) review, validation, and use. From a technical standpoint, transparency in the use of either or both validated or unvalidated metrics for different synthetic data generation models would be critical to evaluate the validity of study results or evidence generated through the inclusion of synthetic data. Also, bias assessment tools like APPRAISE could be useful to assess RWD sets and their synthetic derivatives for bias due to inappropriate study design, and future work can confirm the utility and reliability of APPRAISE in such cases.[54] Clarification around the validity and utility of point estimates derived from synthetic datasets would also be critical to support and inform end-users of synthetic data.

Practical Considerations for Synthetic Data Generation and Use

Amid the regulatory acceptability of RWD being largely exploratory, an interesting development to date is the FDA's regulatory science research pathway that was created for the purposes of exploring the value of synthetic data based on its current definition.[1,55] For instance, the FDA's AI program within their Center for Devices and Radiological Health (CDRH) is presently examining "the possibilities and limitations of supplementing medical patient datasets with synthetic data, for example, artificial data that has been partially or fully generated using computational techniques." The AI program currently supports four distinct projects demonstrating how "real patient datasets can be supplemented by creating realistic digital object models, digital replicas of acquisition devices, and resulting large-scale synthetic datasets."[55] Through this AI program, CDRH has acknowledged that generative models and the resulting synthetic data warrant rapid development and policy assessment to better serve patients. Other divisions within the FDA, like the Center for Drug Evaluation (CDER) and Research and Center for Biologics Evaluation and Research (CBER), could potentially follow suit to help drive policy science innovation within their respective divisions.

Likewise, the UK's MHRA provides access to high-fidelity and medium-fidelity synthetic datasets under a non-negotiable data sharing agreement, noting that the data could be used for "training purposes or to improve algorithms or machine learning workflows."[17] The UK's NHS England's National Disease Registration Service also offers Simulacrum, a "dataset that contains artificial patient-like cancer data to help researchers gain insights,"[19] which was used recently to compare real-world endpoints, e.g., real-world overall survival, within, or gleaned from, the dataset to oncology data from patients in the U.S. who participated in the Friends of Cancer Research Real-World Evidence Pilot Project 1.0. This intended to evaluate outcomes among patients with immunotherapy-treated advanced non–small-cell lung cancer (a rare form of cancer). This method can inform similar activities focused on immunotherapy development.[56] Additionally, in 2022, the EMA released a favorable qualification opinion to provide a regulatory framework for the application of PROCOVA™, a prognostic digital twin (or arguably synthetic data) solution to support Phase 2 and 3 clinical trials.[57]

The FDA recently published a special communication discussing possible requirements for "flexible mechanisms to keep up with the pace of change in AI across biomedicine and health care" and "proficiency in evaluating the use of AI in premarket development."[58] The FDA has authorized over a thousand AI/ML-enabled medical devices as of December 2024.[59] However, the FDA and most regulators have published no specific recommendations for synthetic data. Therefore, in addition to **Figure 1,** we also highlight the National Institute of Standards and Technology (NIST) Artificial Intelligence Risk Management Framework that offers a structured approach to assessing and managing risks associated with AI systems.[60] The framework is organized around four core functions to assess AI risk: govern, map, measure, and manage. In line with recommendations from researchers and stakeholders working with synthetic data generated by generative AI, the NIST "Govern" function emphasizes the importance of understanding,

managing, and documenting the legal and regulatory requirements associated with AI use. As a best practice, we recommend stakeholders engage with the NIST framework and assess risk of synthetic data generated through generative AI to align guidance standards with synthetic data use. **With the NIST framework and regulatory guidance and developments in mind, we recommend that practitioners conducting risk assessments of synthetic data generation using generative AI take the following seven steps.**

These proposed steps were adapted from the FDA's Risk-Based Credibility Assessment Framework, as outlined in the draft guidance "Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products."[61] Therefore, we believe this proposed risk-based credibility assessment framework provides a structured and regulator-aligned process for establishing and evaluating the credibility of synthetic data generated using generative AI within a specific COU.

Steps for Risk Assessment of Synthetic Data Generation Using Generative AI

Step 1
Define the question of interest that will be addressed by the synthetic dataset generated through generative AI.

Step 2
Define the COU for the synthetic dataset generated through generative AI.

Step 3
Assess the risk, relevance, reliability, and quality of synthetic data generated through generative AI. Risk can be assessed using the NIST Artificial Intelligence Risk Management Framework. Relevance, reliability, and quality can be assessed using RWD/E assessments, including determining fitness for use.

Step 4
Develop a plan to establish the relevance, reliability, and quality of synthetic data generated through generative AI output within the COU.

Step 5
Execute the plan.

Step 6
Document the results of the relevance, reliability, and quality assessment plan and discuss deviations from the plan.

Step 7
Determine the adequacy of the synthetic data generated through generative AI for the COU.

CONCLUSION

Regulators outside of the U.S., such as the EMA and MHRA, who each have explicit goals and interests to leverage synthetic data should take interest in sharing the results of their explorations broadly with the regulatory community to help avoid siloed thinking across regulatory settings globally. Synthetic data acceptability requires a culture of learning and transparency among not only regulators, but also end-users of synthetic data and those participating in the synthetic data generation and exchange pipeline, e.g., sponsors, patients, IRBs, health care providers, etc. Members of industry, academia, and government alike should openly and ongoingly share both successes and challenges in using synthetic data to drive clinical research and foster precompetitive approaches toward unified research framework. Last but not least, alongside scientific best practices, e.g., bias assessments, transparency, etc., close considerations of ethical and legal implications are key to understanding both the benefits and risks associated with synthetic data generation and use. Where synthetic data generated using generative AI is concerned, we recommend that principles from current RWD and AI best practice frameworks and guidance documents be applied or adapted, as proposed herein and when appropriate, to address potential risk within a given COU.

References

[1] Office of the Commissioner. FDA Digital Health and Artificial Intelligence Glossary – Educational Resource. FDA Digital Health and Artificial Intelligence Glossary-Educational Resource. September 26, 2024. Accessed October 17, 2024. https://www.fda.gov/science-research/artificial-intelligence-and-medical-products/fda-digital-health-and-artificial-intelligence-glossary-educational-resource

[2] Goyal M, Mahmoud QH. A Systematic Review of Synthetic Data Generation Techniques Using Generative AI. *Electronics.* 2024;13(17):3509. doi:10.3390/electronics13173509

[3] van Breugel B, Liu T, Oglic D, van der Schaar M. Synthetic data in biomedicine via generative artificial intelligence. *Nat Rev Bioeng.* 2024;2(12):991-1004. doi:10.1038/s44222-024-00245-7

[4] Reddy S. Generative AI in healthcare: an implementation science informed translational path on application, integration and governance. *Implement Sci.* 2024;19(1):27. doi:10.1186/s13012-024-01357-9

[5] Arora A, Arora A. Generative adversarial networks and synthetic patient data: current challenges and future perspectives. *Future Healthc J.* 2022;9(2):190-193. doi:10.7861/fhj.2022-0013

[6] Akkem Y, Biswas SK, Varanasi A. A comprehensive review of synthetic data generation in smart farming by using variational autoencoder and generative adversarial network. *Eng Appl Artif Intell.* 2024;131:107881. doi:10.1016/j.engappai.2024.107881

[7] Emam KE, Mosquera L, Zheng C. Optimizing the synthesis of clinical trial data using sequential trees. *J Am Med Inform Assoc.* 2021;28(1):3-13. doi:10.1093/jamia/ocaa249

[8] El Kababji S, Mitsakakis N, Fang X, et al. Evaluating the Utility and Privacy of Synthetic Breast Cancer Clinical Trial Data Sets. *JCO Clin Cancer Inform.* 2023;(7):e2300116. doi:10.1200/CCI.23.00116

[9] Rodriguez L. Synthetic Data Helps Counter Lack of Diversity in Data | CBIIT. Cancer Data Science Pulse Blog. October 10, 2024. Accessed February 21, 2025. https://datascience.cancer.gov/news-events/blog/synthetic-data-helps-counter-lack-diversity-data

[10] Melo CM de, Torralba A, Guibas L, DiCarlo J, Chellappa R, Hodgins J. Next-generation deep learning based on simulators and synthetic data. *Trends Cogn Sci.* 2022;26(2):174-187. doi:10.1016/j.tics.2021.11.008

[11] European Medicines Agency. *Seizing Opportunities in a Changing Medicines Landscape | European Medicines Agency (EMA).;* 2024. Accessed October 17, 2024. https://www.ema.europa.eu/en/news/seizing-opportunities-changing-medicines-landscape

[12] Medicines and Healthcare products Regulatory Agency. *MHRA Data Strategy 2024 - 2027.;* 2024. Accessed October 17, 2024. https://www.gov.uk/government/publications/mhra-data-strategy-2024-2027

[13] European Medicines Agency. *Reflection Paper on the Use of Artificial Intelligence in the Lifecycle of Medicines* | European Medicines Agency (EMA).; 2023. Accessed May 6, 2025. https://www.ema.europa.eu/en/news/reflection-paper-use-artificial-intelligence-lifecycle-medicines

[14] Canada's Drug Agency. *Canada's Drug Agency Position Statement on the Use of Artificial Intelligence in the Generation and Reporting of Evidence.*

[15] Thorlund K, Dron L, Park JJ, Mills EJ. Synthetic and External Controls in Clinical Trials- A Primer for Researchers. *Clin Epidemiol.* 2020;12:457-467. doi:10.2147/CLEP.S242097

[16] Pasculli G, Virgolin M, Myles P, et al. Synthetic Data in Healthcare and Drug Development: Definitions, Regulatory Frameworks, Issues. *CPT Pharmacomet Syst Pharmacol.* n/a(n/a). doi:10.1002/psp4.70021

[17] Clinical Practice Research Datalink. Synthetic data. August 22, 2024. Accessed October 17, 2024. https://www.cprd.com/synthetic-data

[18] Giuffrè M, Shung DL. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *Npj Digit Med*. 2023;6(1):1-8. doi:10.1038/s41746-023-00927-3

[19] NHS England's National Disease Registration Service. The Simulacrum. Accessed October 17, 2024. https://healthdatainsight.org.uk/project/the-simulacrum/

[20] Shen X, Liu Y, Shen R. Boosting Data Analytics With Synthetic Volume Expansion. Published online March 10, 2024. doi:10.48550/arXiv.2310.17848

[21] Wang P, Loignon AC, Shrestha S, Banks GC, Oswald FL. Advancing Organizational Science Through Synthetic Data: A Path to Enhanced Data Sharing and Collaboration. *J Bus Psychol.* Published online December 6, 2024. doi:10.1007/s10869-024-09997-w

[22] Agency for Science, Technology and Research, Personal Data Protection Commission Singapore (PDPC). *Privacy Enhancing Technology (PET): Proposed Guide on Synthetic Data Generation.*

[23] Rajotte JF, Bergen R, Buckeridge DL, Emam KE, Ng R, Strome E. Synthetic data as an enabler for machine learning applications in medicine. *iScience.* 2022;25(11):105331. doi:10.1016/j.isci.2022.105331

[24] Innovative Health Initiative, EFPIA. *Data Sharing Playbook Can Unlock Health Partnerships.* Accessed October 17, 2024. http://www.ihi.europa.eu/news-events/newsroom/data-sharing-playbook-can-unlock-health-partnerships

[25] Jordan S, Fontaine C, Hendricks-Sturrup R. Selecting Privacy-Enhancing Technologies for Managing Health Data Use. *Front Public Health.* 2022;10. doi:10.3389/fpubh.2022.814163

[26] Wagner JK, Cabrera LY, Gerke S, Susser D. Synthetic data and ELSI-focused computational checklists—A survey of biomedical professionals' views. *PLOS Digit Health.* 2024;3(11):e0000666. doi:10.1371/journal.pdig.0000666

[27] Corona LE, Lee VS, Weisman AG, et al. Mixed Gonadal Dysgenesis: A Narrative Literature Review and Clinical Primer for the Urologist. *J Urol.* 2024;212(5):660. doi:10.1097/JU.0000000000004137

[28] Yan C, Zhang X, Yang Y, et al. Differences in Health Professionals' Engagement With Electronic Health Records Based on Inpatient Race and Ethnicity. *JAMA Netw Open.* 2023;6(10):e2336383. doi:10.1001/jamanetworkopen.2023.36383

[29] Akgün KM, Feder SL. Whom Should We Regard as Responsible for Health Record Inaccuracies That Hinder Population-Based Fact Finding? *AMA J Ethics.* 2025;27(1):6-13. doi:10.1001/amajethics.2025.6

[30] *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.;* 2023. Accessed October 17, 2024. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

[31] Center for Devices and Radiological Health (CDRH). Artificial Intelligence and Machine Learning in Software as a Medical Device. FDA. Published online September 26, 2024. Accessed October 17, 2024. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device

[32] US State Privacy Legislation Tracker. Accessed October 17, 2024. https://iapp.org/resources/article/us-state-privacy-legislation-tracker/

[33] American Medical Association. *Affordable Care Act Section 1557 Fact Sheet.* https://www.ama-assn.org/sites/ama-assn.org/files/corp/media-browser/public/ama-fact-sheet-section-1557.pdf

[34] Tenenbaum JD, Goodman KW. Beyond the Genetic Information Nondiscrimination Act: Ethical and Economic Implications of the Exclusion of Disability, Long-Term Care and Life Insurance. *Pers Med.* Published online March 1, 2017. doi:10.2217/pme-2016-0078

[35] Das SK, Roy P, Kumar Mishra A. Analysis of Synthetic Data Generation Techniques in Diabetes Prediction. In: Borah MD, Laiphrakpam DS, Auluck N, Balas VE, eds. *Big Data, Machine Learning, and Applications.* Springer Nature Singapore; 2024:587-599.

[36] Chen A. A novel graph methodology for analyzing disease risk factor distribution using synthetic patient data. *Healthc Anal.* 2022;2:100084. doi:10.1016/j.health.2022.100084

[37] Halfpenny W, Baxter SL. Towards effective data sharing in ophthalmology: data standardization and data privacy. *Curr Opin Ophthalmol.* 2022;33(5):418. doi:10.1097/ICU.0000000000000878

[38] Hahn W, Schütte K, Schultz K, et al. Contribution of Synthetic Data Generation towards an Improved Patient Stratification in Palliative Care. *J Pers Med.* 2022;12(8):1278. doi:10.3390/jpm12081278

[39] Skandarani Y, Lalande A, Afilalo J, Jodoin PM. Generative Adversarial Networks in Cardiology. *Can J Cardiol.* 2022;38(2):196-203. doi:10.1016/j.cjca.2021.11.003

[40] Candemir S, Nguyen XV, Folio LR, Prevedello LM. Training Strategies for Radiology Deep Learning Models in Data-limited Scenarios. *Radiol Artif Intell.* 2021;3(6):e210014. doi:10.1148/ryai.2021210014

[41] Johnson C, Price G, Khalifa J, et al. A method to combine target volume data from 3D and 4D planned thoracic radiotherapy patient cohorts for machine learning applications. *Radiother Oncol J Eur Soc Ther Radiol Oncol.* 2018;126(2):355-361. doi:10.1016/j.radonc.2017.11.015

[42] Nora Emmott, Maryam Nafie, Neha Shaw, Rachele Hendricks-Sturrup. *Regulatory Fit for Purpose Considerations for Patient-Generated Health Data.* Duke-Margolis Institute for Health Policy; 2024. https://healthpolicy.duke.edu/sites/default/files/2024-06/Regulatory%20Fit-for-Purpose%20Considerations%20for%20Patient-Generated%20Health%20Data.pdf

[43] Juwara L, El-Hussuna A, Emam KE. An evaluation of synthetic data augmentation for mitigating covariate bias in health data. *Patterns.* 2024;5(4). doi:10.1016/j.patter.2024.100946

[44] Food and Drug Administration, Duke-Margolis Institute for Health Policy. Optimizing the Use of Postapproval Pregnancy Safety Studies. Presented at: Accessed October 17, 2024. https://healthpolicy.duke.edu/events/optimizing-use-postapproval-pregnancy-safety-studies

[45] Food and Drug Administration, Duke-Margolis Institute for Health Policy. Emerging Best Practices and Future Directions in Data Privacy and Security. Presented at: Accessed October 17, 2024. https://healthpolicy.duke.edu/events/emerging-best-practices-and-future-directions-data-privacy-and-security

[46] Office of the Commissioner. Pregnancy Exposure Registry Information. FDA. August 9, 2024. Accessed October 17, 2024. https://www.fda.gov/consumers/pregnancy-exposure-registries/pregnancy-exposure-registry-information

[47] Wang SV, Pottegård A, Crown W, et al. HARmonized Protocol Template to Enhance Reproducibility of Hypothesis Evaluating Real-World Evidence Studies on Treatment Effects: A Good Practices Report of a Joint ISPE/ISPOR Task Force. Value Health. 2022;25(10):1663-1672. doi:10.1016/j.jval.2022.09.001

[48] Fleurence RL, Kent S, Adamson B, et al. Assessing Real-World Data From Electronic Health Records for Health Technology Assessment: The SUITABILITY Checklist: A Good Practices Report of an ISPOR Task Force. *Value Health J Int Soc Pharmacoeconomics Outcomes Res.* 2024;27(6):692-701. doi:10.1016/j.jval.2024.01.019

[49] Alaa AM, Breugel B van, Saveliev E, Schaar M van der. How Faithful is your Synthetic Data? Sample-level Metrics for Evaluating and Auditing Generative Models. Published online July 13, 2022. Accessed October 17, 2024. http://arxiv.org/abs/2102.08921

[50] Pushkarna M, Zaldivar A, Kjartansson O. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency.* FAccT '22. Association for Computing Machinery; 2022:1776-1826. doi:10.1145/3531146.3533231

[51] Alloza C, Knox B, Raad H, et al. A Case for Synthetic Data in Regulatory Decision-Making in Europe. *Clin Pharmacol Ther.* 2023;114(4):795-801. doi:10.1002/cpt.3001

[52] Mahendraratnam N, Silcox C, Mercon, et al. Determining Real-World Data's Fitness for Use and the Role of Reliability. Published online September 26, 2019. https://healthpolicy.duke.edu/publications/determining-real-world-datas-fitness-use-and-role-reliability

[53] Center for Drug Evaluation and Research (CDER). *Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products.* FDA; 2024. Accessed September 20, 2024. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory

[54] OSF | APPRAISE: a tool for appraising potential for bias in real-world evidence studies on medication effectiveness or safety. Accessed October 17, 2024. https://osf.io/a4nhd/

[55] Center for Devices and Radiological Health (CDRH). Addressing the Limitations of Medical Data in AI. FDA. June 10, 2024. Accessed October 17, 2024. https://www.fda.gov/medical-devices/medical-device-regulatory-science-research-programs-conducted-osel/addressing-limitations-medical-data-ai

[56] Horvat P, Gray CM, Lambova A, et al. Comparing Findings From a Friends of Cancer Research Exploratory Analysis of Real-World End Points With the Cancer Analysis System in England. *JCO Clin Cancer Inform.* 2021;(5):1155-1168. doi:10.1200/CCI.21.00013

[57] European Medicines Agency. *Qualification Opinion for Prognostic Covariate Adjustment (PROCOVATM).;* 2022.

[58] Warraich HJ, Tazbaz T, Califf RM. FDA Perspective on the Regulation of Artificial Intelligence in Health Care and Biomedicine. *JAMA.* Published online October 15, 2024. doi:10.1001/jama.2024.21451

[59] Center for Devices and Radiological Health, US Food and Drug Administration (FDA). Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. Published online March 25, 2025. Accessed May 6, 2025. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices

[60] National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework (AI RMF 1.0).* US Department of Commerce; 2023.

[61] Center for Drug Evaluation and Research, US Food and Drug Administration. *Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products.* US Food and Drug Administration; 2025. Accessed May 16, 2025. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-artificial-intelligence-support-regulatory-decision-making-drug-and-biological

Appendix A | 2025 RWE Collaborative Advisory Group Roster

This paper was informed by the expert collaborators in the Duke-Margolis 2023-2025 Real-World Evidence Collaborative Advisory Group. The lists below reflect the 2023, 2024 and 2025 Advisory Group rosters, which advised on the development of this policy briefs. Listed member affiliations may not reflect current affiliations.

Jessica Albano
Syneos Health

Ginny Beakes-Read
Janssen Pharmaceuticals

Marc L. Berger
Independent Consultant

Elise Berliner
Oracle Life Sciences

Paul Boyce
PatientsLikeMe

Brian Bradbury
Amgen

Stella Chang
OMNY Health

Mandy Christensen
Walmart

Mark Cziraky
Carelon

Nancy Dreyer
Independent Consultant

Andenet Emiru
University of California

Laura Fabbri
Chiesi Farmaceutici

Nicole Gatto
Aetion

Morgan Hanger
CTTI

Camille Jackson
Flatiron

Ryan Kilpatrick
Abbvie

Grazyana Lieberman
Independent Consultant

Christina Mack
ISPE and IQVIA

Janna Manjelievskaia
Veradigm

Mark Marsico
Merck & Company

Sarah Martin
Eli Lilly & Company

Allison Martin
Sanofi

Lynn McRoy
Pfizer

Anne-Marie Meyer
Independent Consultant

Carsten Moeller
Bayer

Irene Nunes
Genmab

Paul Petraro
Boehringer Ingelheim

Rodrigo Refoios Camejo
GSK

Dan Riskin
Verantos

Debra Schaumberg
Evidera

Lauren Silvis
Tempus

Jaime Smith
Paraxel International

Michael Taylor
Genentech

David Thompson
Independent Consultant

Darren Toh
Havard Pilgrim

Alex Vance
Holmusk

Richard Willke
Independent Consultant

Appendix B | 2024 RWE Collaborative Advisory Group Roster

Jessica Albano
Syneos Health

Marc L. Berger
Independent Consultant

Elise Berliner
Oracle Life Sciences

Barbara Bierer
Brigham and Women's Hospital and Harvard

Mac Bonafede
Veradigm

Brian Bradbury
Amgen

Jeff Brown
TriNetX

Stella Chang
OMNY Health

Mandy Christensen
Walmart

Bill Crown
Brandeis University

Mark Cziraky
Carelon

Riad Dirani
Teva Pharmaceuticals Industries

Nancy Dreyer
Independent Consultant

Andenet Emiru
University of California Office of the President

Omar Escontrias
National Health Council

Laura Fabbri
Chiesi Farmaceutici

Marni Hall
IQVIA

Morgan Hanger
Clinical Trials Transformation Initiative

Stacy Holdsworth
Eli Lilly & Company

Javier Jimenez
Flatiron Health

Ryan Kilpatrick
Abbvie

Grazyana Lieberman
Independent Consultant

Erlyn Macarayan
PatientsLikeMe

Christina Mack
ISPE and IQVIA

Mark Marsico
Merck & Company

Nell Marshall
Evidation Health

Allison Martin
Sanofi

Lynn McRoy
Pfizer

Anne-Marie Meyer
Independent Advisor

Eleanor Perfetto
University of Maryland

Laura Pizzi
ISPOR

Jeremy Rassen
Aetion

Rob Reynolds
GSK

Miriam Saillant
Genmab US

Khaled Sarsour
Janssen Pharmaceuticals

Debra Schaumberg
Evidera

Thomas Seck
Boehringer Ingelheim International

Lauren Silvis
Tempus Labs

Jaime Smith
Parexel International

Montse Soriano Gabarro
Bayer

Michael Taylor
Genentech

David Thompson
Independent Consultant

Darren Toh
Harvard Pilgrim Health Care Institute

Alex Vance
Holmusk

Richard Willke
Independent Consultant

Appendix C | 2023 RWE Collaborative

Marc Berger
Independent Consultant

Elise Berliner
Oracle Life Sciences

Barbara Bierer
Brigham and Women's Hospital

Mac Bonafede
Veradigm

Brian Bradbury
Amgen

Jeffrey Brown
TriNetX

Adrian Cassidy
Novartis

Stella Chang
OMNY Health

William Crown
Brandeis University

Mark Cziraky
Carelon

Riad Dirani
Teva Pharmaceuticals

Nancy Dreyer
Independent Consultant

Andenet Emiru
University of California
Office of the President

Omar Escontrias
National Health Council

John Graham
GSK

Matthew Harker
Evidation

Joe Henk
UnitedHealthCare

Ceri Hirst
Bayer

Stacy Holdsworth
Eli Lilly & Company

Ryan Kilpatrick
Abbvie

Lisa Lavange
University of North Carolina

Grazyna Lieberman
Independent Consultant

Lyn Macarayan
PatientsLikeMe

Christina Mack
IQVIA and ISPE

Megan O'Brien
Merck & Company

Sally Okun
Clinical Trials
Transformation Initiative

Eleanor Perfetto
University of Maryland

Richard Platt
Harvard University

Jeremy Rassen
Aetion

Stephanie Reisinger
Flatiron

Khaled Sarsour
Janssen Pharmaceuticals

Debra Schaumberg
Evidera, within Thermo
Fisher Scientific

Thomas Seck
Boehringer-Ingelheim

Lauren Silvis
Tempus

Michael Taylor
Genentech

David Thompson
Independent Consultant

Alex Vance
Holmusk

Richard Willke
ISPOR

Bob Zambon
Syneos Health