

Viewpoint

Addressing the Challenges in Using Synthetic Data for Health Research: Application to Cardiology

Louise Baschet¹, MSc; Sebastien Marque¹, PhD; Stéphane Locret², MD; Jens Jenssen³, MD; Vanessa Barbet¹, MSc; Patrick Jourdain⁴, MSc, MBA, MD

¹Horiana, Bordeaux, France

²Research Department, Ramsay Santé, Paris, France

³Department of Clinical Science and Education, Södersjukhuset, Karolinska Institutet and Capho St Görans Hospital, Stockholm, Sweden

⁴Medical Department, Ramsay Sante, Paris Saclay University, Paris, France

Corresponding Author:

Louise Baschet, MSc

Horiana

80 bis rue Paul Camelle

Bordeaux 33100

France

Phone: 33 687270347

Email: louise.baschet@horiana.com

Abstract

Synthetic data offer significant potential for cardiology research by enabling data sharing, preserving privacy, and supporting machine learning model development. By generating artificial patient records that reflect real-world distributions, synthetic data can accelerate clinical research, improve model performance for rare cardiovascular conditions, and facilitate transnational collaborations that would otherwise be restricted by data-sharing barriers. Despite these advantages, the increasing use of synthetic data raises important ethical, regulatory, and methodological concerns that remain insufficiently addressed. Key challenges include assessing the validity and generalizability of synthetic datasets, understanding their limitations in representing complex and heterogeneous patient populations, and preventing the amplification of existing biases in cardiovascular care. Current regulatory frameworks, including the General Data Protection Regulation (GDPR) and Health Insurance Portability and Accountability Act (HIPAA), do not fully address emerging risks such as reidentification and data leakage, and there is no harmonized guidance to govern the use of synthetic data as stand-alone evidence for medical device evaluation or therapeutic research. In this viewpoint, we argue that responsible integration of synthetic data in cardiology requires, first, clear differentiation between synthetic data as a privacy-preserving distributional substitute and synthetic data as a counterfactual simulation tool, and, second, fit-for-purpose governance frameworks that pair rigorous utility and fidelity testing with explicit, adversary-aware privacy evaluation before synthetic cohorts are accepted as evidence in research or product evaluation. A prerequisite for that governance is conceptual clarity about what synthetic data are being used for. Synthetic data in health care serve 2 fundamentally distinct roles that carry entirely different validity requirements, failure modes, and regulatory implications, yet they are routinely conflated. The first role is as a privacy-preserving distributional substitute: the goal is statistical fidelity to the real data distribution, so that analyses of the synthetic dataset yield results equivalent to those of the original. The second role is as a tool for counterfactual simulation: the goal is to generate data that could not have been observed, such as rare conditions, hypothetical interventions, or extrapolations to new populations. These 2 roles are methodologically distinct. A dataset that accurately reflects real-world distributions may be inadequate for extrapolating findings to underrepresented subgroups. Conversely, a simulator optimized for novel scenario generation may systematically diverge from real-world distributions. This distinction informs every subsequent discussion of validity, bias, and regulation in this viewpoint and our proposed 4 concrete actions for the cardiology research community, including mandatory 3-layer (fidelity, utility, and privacy) validation, systematic subgroup reporting, explicit intended-use scoping, and domain-specific acceptability thresholds for synthetic data-based evidence.

JMIR Cardio 2026;10:e92930; doi: [10.2196/92930](https://doi.org/10.2196/92930)

Keywords: health care; machine learning; synthetic data; virtual patients; cardiology; medical research

Introduction: Synthetic Data as an Emerging Tool in Cardiology

Used for many decades in other fields, synthetic data were adopted more recently in health care and have been shown to offer significant advancements. Specifically, synthetic data address challenges related to data privacy, progressively optimize care, and improve clinical research [1-3]. In oncology, models trained on synthetic data have been shown to predict survival more accurately than models trained on original data [4]. In addition, use of massive data can accelerate research, reducing costs and inclusion periods associated with clinical trials and data analysis. However, while the benefits are now becoming well known, the limitations, especially legal framework-related limitations, ethics, and responsibilities, are less well known [5-7]. In this context of growth, understanding the relevance of synthetic data, their compliance with regulations, and their potential to address the lack of patient data in various scenarios is essential.

In this viewpoint, we provide an overview of the main applications and roles of synthetic data in cardiology, highlight the conceptual distinction between distributional and counterfactual uses, and examine the associated methodological, ethical, and regulatory challenges. We begin by providing definitions for the key terms used in the literature, and we outline the methods, benefits, and challenges of using synthetic data in cardiology research. Our central argument is that synthetic data in cardiology should not be viewed merely as a privacy-preserving technical workaround, but as a methodologically distinct form of evidence that requires dedicated governance, validation standards, and interpretability safeguards. The absence of such standards is itself a major risk: without them, synthetic data risk being adopted inconsistently, reinforcing the blind spots they are meant to address. We conclude by underlining the potential impact of synthetic data for transnational research use without compromising patient rights to privacy and by proposing 4 concrete actions for the cardiology research community to operationalize this governance agenda.

Cardiology is a particularly important test case for this argument. The high-frequency, time-series nature of key cardiac signals (electrocardiograms [ECGs], photoplethysmogram [PPG], and phonocardiograms) imposes strict fidelity requirements on generative models that are not encountered in the same way in other medical specialties. The spatial complexity of cardiac imaging modalities (magnetic resonance imaging [MRI] and computed tomography [CT]) demands higher-dimensional generative capacity, and the prevalence of rare arrhythmias and structural anomalies create chronic data scarcity for training predictive models. Additionally, the long-term disease trajectories, characteristic of chronic heart failure and coronary artery disease, challenge generative models that capture cross-sectional distributions but cannot capture longitudinal dynamics. These domain-spe-

cific constraints make the governance agenda outlined in this viewpoint particularly relevant and timely.

Clarifying the Landscape: Key Terms in Synthetic and Virtual Patient Data

Synthetic Data

The seminal work on synthetic data by the Royal Society and The Alan Turing Institute stresses functional and intentional aspects of synthetic data that have “been generated using a purpose-built mathematical model or algorithm, with the aim of solving a (set of) data science task(s)” [8]. Synthetic data is artificial data generated from original real data and a model that is trained to reproduce the characteristics and structure of the original data [9]. Synthetic and original databases should deliver very similar results when subjected to the same statistical analysis despite different cell values. The accuracy of synthetic data as a proxy for the original data reflects the *utility* of the method and the model.

Data Augmentation

In computer vision and signal processing, augmentation typically involves transforming existing data (eg, rotation, scaling, and noise addition), while preserving labels [10]. For structured clinical or electronic health record data, augmentation more often relies on generating new patient records using statistical or machine learning (ML) models trained on real cohorts, resulting in datasets that mix real and synthetic patients [11].

Virtual Patients

Virtual patients are simulated patients used in medical training, providing a risk-free environment for health care professionals to refine their skills [6,12]. For example, a simulator was developed for patients living with heart failure to improve care delivery. It combines comprehensive hemodynamic and energetic data for personalized prognostication, such as identifying the need for heart rhythm therapy and forecasting response to device and medical therapies [13]. This is particularly useful for patients with complex clinical situations and/or drugs with a narrow therapeutic margin. However, the model can only reproduce existing elements and cannot identify new sources of interactions.

Synthetic Patients

Synthetic patients mimic real patient characteristics without using actual patient information. They are generated by statistical methods or AI and are represented in datasets, primarily for privacy-preserving research purposes [11].

Digital or Virtual Twins

Digital or virtual twins are real-time virtual replicas of physical or biological systems used to simulate patient behaviors and optimize treatments in cardiology [6,14]. These twins receive continuous updates of multiple data types to predict disease progression. Typically, twinning is partial to preserve privacy. Current innovation is increasing in

cardiovascular digital twin research [15], including growing industry and academic patent activity [6,14]. Digital twins could also help plan complex surgeries, rhythm-based procedures, or to train junior cardiologists.

Taken together, these three categories exist on a spectrum: (1) virtual patients are mechanistic simulation constructs primarily used for training and procedural planning, without direct derivation from a specific individual's records; (2) synthetic patients are population-level statistical surrogates generated to preserve privacy or augment data; and (3) digital twins are individual-level, continuously updated dynamic models. Understanding this spectrum is essential, as each category carries different validity requirements, failure modes, and regulatory implications.

How Synthetic Data Is Generated: Methods and Underlying Models

Artificial data mimicking real patient statistics can be generated from 3 sources: [11] by creating new data points similar to one patient, by modifying data from one patient using the nearest neighbors, or by characterizing the patient population globally to create synthetic patients representing shared traits. Focusing on the last global approach, synthetic data generation can use multiple algorithms [5]. Two types of models are described in McDuff et al's review [16].

First, physical models require detailed parameterization, offering interpretability and known constraints, making them a safe choice for data generation. However, they tend to be task-specific, may not generalize well across domains, and often demand significant computational resources. Complex real-world systems can lead to model mismatches, with unknown factors posing risks to practical use.

Second, statistical generative models, such as generative adversarial networks (GANs) [17] and variational autoencoders (VAEs) [18], replicate dataset distributions and can produce realistic, high-resolution data. These models are generally easier to create than physical simulators but may lack interpretability and control. Their reliability in medical applications heavily depends on the quality of the original data (used for training) and the use of supervised learning for better regulation of output [16].

These models use explicit or implicit data generation methods [11]. Explicit approaches are used when the relationship between the variables is known or at least assumed. In these explicit approaches, it is possible to interpret the strength of association between variables (eg, covariates, outcomes, or interactions).

The first and most frequent explicit approach is mechanistic models, which are typically based on systems of ordinary or partial differential equations describing physiological processes and their interactions over time under known constraints. The second explicit approach is Bayesian networks, which can be used to generate new synthetic data through probabilistic graphical models that represent knowledge about variables and their dependencies. Bayesian

networks can be used to make inferences about specific variables of interest in the network, depending on the other variables in the network.

Implicit approaches are used where there is no need to know how the variables are linked, as the model will determine this by itself, and there is no need to interpret the structure of the model. GANs and VAEs, and denoising diffusion models are all implicit approaches. GAN-based methods have significantly enhanced data analysis capabilities [11]. In health care, GANs have been successfully applied to diverse imaging modalities in cardiology, such as MRI and CT, yielding promising results [19-21]. However, a notable limitation is that most studies have relied on large training datasets or low-dimensional data. While initially less popular than GANs for data augmentation, VAEs are gaining traction in medical applications, particularly for classification and segmentation tasks [22,23]. The main limitation of VAEs is that they can produce imprecise samples when not properly tuned, especially with limited training data. Diffusion models represent an emerging class with promising fidelity for imaging tasks but remain limited in cardiology-specific applications.

To link this taxonomy to application, physical and mechanistic models are preferred when physiological interpretability and regulatory transparency are required (eg, hemodynamic simulation for device testing), at the cost of generalizability. GANs are well-suited for high-dimensional imaging data but are particularly prone to bias amplification when training datasets are small or demographically skewed. VAEs are better suited to lower-dimensional structured data and classification tasks. Bayesian networks are appropriate when variable relationships are partially known and interpretability is a priority. Choosing the right model for the intended use case is a prerequisite for the validation process described below.

Applications and Benefits for Cardiology Research

Overview

Synthetic data enables ML with large, labeled datasets that are otherwise difficult or expensive to obtain due to privacy or access constraints. Synthetic data also accelerates software testing and model pretraining and is widely used in health care to support quicker, smaller trials with fewer privacy risks. For example, synthetic data generation can reduce the time required to set up a cardiovascular data platform. In 1 reported case, synthetic data generation reduced the time required to set up a cardiovascular data platform from 1.5 years to 1 month [24], illustrating the potential for efficiency gains, though systematic evidence on typical effect sizes and the conditions under which such benefits are realized remains limited. AI models trained with synthetic or mixed data (real and synthetic data) often match or outperform those trained with real data only [25]. For example, when cancer survival prediction models were trained on synthetic data, they showed improved performance of >0.1% and,

for extremely small datasets, outperformed those using only original datasets [4].

Synthetic data can significantly enhance the predictive power of AI models and optimize clinical research in cardiology through several key mechanisms.

Augmenting Data Volume and Simplifying Exportability of Health Care Datasets

Synthetic data can bolster the volume of available data, enabling more robust and comprehensive analyses [8,26]. This is particularly valuable in cardiology, where large datasets are crucial for training accurate predictive models. However, this benefit is conditional: augmenting data volume only improves downstream model performance when the generative model has successfully captured the true underlying distribution of the real data, a condition that must be verified through explicit utility and fidelity testing before the synthetic dataset is used.

Predicting Clinical Trial Results

In silico trials represent an experimental application used to expand testing methods. For example, in silico trials enable results to be extrapolated to populations with different demographics or characteristics [27], facilitating the inclusion of additional comparators, and they enable the extension of short-term study results to long-term projections [28].

Addressing Data Scarcity

When real patient data are limited, synthetic data can fill gaps and improve performance of predictive models. Synthetic data can be used to simulate rare cardiovascular conditions, allowing AI models to learn from a larger number of cases than would be available in real-world datasets [29]. For rare events, simulators can be used to create synthetic examples and thereby reduce the lack of sensitivity of a model trained with class imbalance. Examples include detecting arrhythmia from PPG or ECG [16,26].

Enhancing Data Diversity

When intentionally designed with diversity objectives and appropriate constraints, synthetic data can introduce greater variability into datasets [6]. However, this benefit is not automatic. Generative models trained on biased data will reproduce those biases. Achieving genuine diversity enhancement requires deliberate methodological choices, including stratified sampling, conditional generation targeting underrepresented groups, and rigorous subgroup validation.

Preserving Patient Privacy

By using synthetic data generated from real patient records, researchers can develop and test AI models without compromising sensitive health care–related patient information [6, 24].

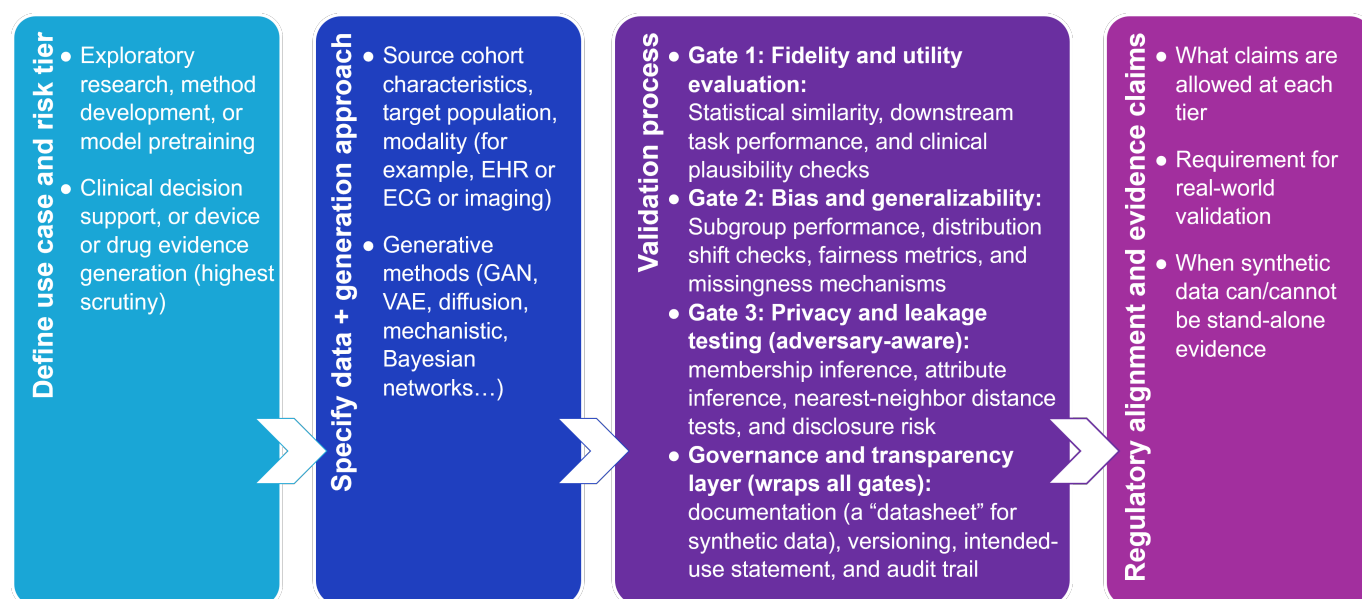
Improved Accuracy of Diagnosis

Synthetic patient data have been widely adopted for the processing of physiological signals and downstream classification in physiological measurement [16]. In particular, researchers have investigated the synthesis of cardiac signals, namely ECGs, phonocardiograms, and PPG [16]. Models trained with real ECGs augmented with synthetic ECGs have shown improved accuracy of arrhythmia diagnosis [30].

Challenges and Limitations in Applying Synthetic Data to Cardiology

While synthetic patients offer promising avenues for research and clinical applications, Figure 1 proposes a fit-for-purpose evaluation and governance workflow that links technical performance (utility and fidelity), generalizability and bias, privacy risk, and regulatory expectations to the intended use of a synthetic cohort in cardiology.

Figure 1. Proposed fit-for-purpose workflow for the evaluation and governance of synthetic cohorts. ECG: electrocardiogram; EHR: electronic health record; GAN: generative adversarial network; VAE: variational autoencoder.



Data Validity and Reliability (Fidelity)

One of the primary concerns with synthetic data is their validity. If the algorithms used to generate synthetic data do not accurately reflect the complexities of real-world patient populations, the conclusions drawn from studies using these data may be misleading. In cardiology, where patient responses can vary significantly, this is a critical issue. Addressing this systematically requires 3 complementary validation layers, as reflected in the governance workflow shown in Figure 1. First, fidelity validation involves assessing statistical similarity between synthetic and real distributions, covering both marginal and joint distributions and, for cardiology, domain-specific signal properties such as QRS morphology and RR-interval dynamics in ECG data. Second, utility validation involves testing whether models trained on synthetic data perform equivalently to those trained on real data for the downstream task of interest, using a “train-on-synthetic, test-on-real” paradigm. Third, privacy validation applies adversary-aware tests, including membership inference attacks (which test whether an adversary can identify individuals from the training set) and attribute inference evaluation (which tests whether sensitive attributes can be reconstructed), rather than relying on the absence of direct identifiers alone. All 3 layers should be treated as minimum reporting standards for any cardiology study using synthetic data.

Generalizability

Synthetic data may not generalize well to diverse or to highly specific populations [16]. Cardiology encompasses a wide range of conditions and patient demographics, and synthetic models may fail to capture the nuances of different patient groups, leading to biased results. Learned generative models produce statistics relative to the underlying data distribution, and large-parameter models have been found to amplify bias [31]. AI or ML models extract or adapt their parameters from training data, making them vulnerable to bias [32].

Ethical Considerations

The use of synthetic data raises ethical questions, particularly regarding consent and the representation of marginalized populations. Ensuring that synthetic data does not perpetuate existing biases in health care is essential for ethical research practices [7,33].

Integration With Clinical Practice

Integrating synthetic patient models into clinical workflows poses challenges. Health care providers may be hesitant to rely on models that are not validated against real-world outcomes, which can hinder the adoption of these technologies in practice.

Regulatory Blind Spots

There are currently no harmonized recommendations from agencies or regulatory authorities defining criteria for the acceptability of synthetic patient cohorts for the evaluation of health products or medical devices, and the very concept of using synthetic patient cohorts for this purpose remains without consensus [7,9,11,34,35]. By providing realistic datasets without exposing sensitive patient information, synthetic data enables health care organizations to comply with privacy regulations, such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR); however, there is still no clear legislation specifically governing the use of synthetic data in health care. Current data protection regulations are limited in their ability to address all the potential risks associated with synthetic data, such as reidentification or data leakage [34,36]. The rise in the use of synthetic data has created an expanding industry of companies offering synthetic data solutions and enabling cross-border data sharing, sometimes beyond the confines of traditional data protection legislation. Regulatory agencies are actively exploring the use of synthetic data, but as of 2025, synthetic

data is not yet broadly trusted as stand-alone evidence for regulatory approval, and findings derived from synthetic data are required to be validated on real data.

Existing partial frameworks offer some orientation but leave critical gaps. The European Medicines Agency's 2024 *Reflection Paper on the Use of Artificial Intelligence (AI) in the medicinal product lifecycle* [32] acknowledges simulation and synthetic data as emerging tools but stops short of defining acceptability criteria for their use as stand-alone evidence. The Food and Drug Administration's evolving guidance on AI and ML-based software as a medical device similarly does not yet address synthetic patient cohorts as a distinct evidence category [37]. Any future harmonized framework would need to specify at minimum: (1) the intended-use scope within which a synthetic dataset has been validated and may legitimately be used, (2) minimum fidelity thresholds appropriate to the application risk tier, (3) mandatory adversary-aware privacy testing before regulatory submission, and (4) a requirement that findings derived from synthetic data be validated on real-world data before supporting regulatory decisions. Until such a framework exists, researchers and sponsors should proactively document their datasets against these criteria to support transparency and anticipate future expectations. Recent work underscores that the core barrier is not only regulatory uncertainty but also the lack of agreed, quantitative ways to demonstrate that a synthetic dataset is both useful and safe. In their scoping review, Kaabachi et al [5] identified dozens of distinct approaches to assess synthetic data utility and only a small set of approaches used to evaluate privacy risk, concluding that many studies do not assess privacy protection at all and that, when they do, residual risks are often underestimated. This supports the call for harmonized evaluation frameworks that pair utility metrics with explicit, adversary-aware privacy testing before synthetic cohorts are considered acceptable evidence in cardiology research or product evaluation.

Privacy

Although synthetic data theoretically offer patient privacy protection, there remains a risk that synthetic data generation may leak personal information [8,16]. Strictly, it is not possible to determine whether synthetic data are sufficiently different from the original samples to be considered anonymized [36]. There are no robust and objective methods of determining whether a synthetic dataset is sufficiently different from the original real dataset to be classified as truly anonymous, and recent reviews highlight inconsistent practice and the frequent absence of privacy-risk evaluation in published synthetic data studies [5]. This is a fundamental, not peripheral, problem: without a reliable method to certify sufficient dissimilarity from training data, the use of synthetic datasets in research settings remains legally ambiguous under both GDPR and HIPAA, regardless of scientific utility. In practice, privacy assurance requires adversary-aware testing rather than reliance on the absence of direct identifiers. Two tests are particularly relevant: membership inference attacks, which evaluate whether an adversary can determine that a specific individual's record was used in training; and attribute inference evaluation, which tests whether sensitive attributes

(such as diagnosis or treatment history) can be reconstructed from synthetic records. Kaabachi et al [5] found that most published synthetic data studies do not perform privacy-risk evaluation at all, and that when they do, adversarial attack methods are rarely applied and residual risks are frequently underestimated. The cardiology field should treat adversary-aware privacy validation as a nonnegotiable component of any synthetic data reporting standard, consistent with the privacy gate in Figure 1.

Technical Limitations

Algorithms used to generate synthetic data can have limitations in terms of complexity and accuracy. In cardiology, where patient data can be highly variable and influenced by numerous factors, these limitations can impact the effectiveness of synthetic data for modeling patient outcomes.

Conclusions

Overall, while synthetic patient research and synthetic data hold substantial promise for advancing cardiovascular research, careful consideration of the associated challenges is necessary to ensure that these tools are used effectively in clinical practice. Synthetic datasets and in silico studies could transform the future of cardiovascular research, allowing the use and transfer of reliable datasets without compromising patient privacy and addressing data scarcity. Realizing this potential depends on recognizing and governing the 2 fundamentally different roles that synthetic data can play, either as a privacy-preserving distributional substitute for existing data or as a counterfactual simulator for rare events and hypothetical scenarios, and on applying validation criteria appropriate to each role. Within strict boundaries and under specific conditions to limit potential bias, synthetic data can uphold privacy, equity, and safety, but care must be taken to avoid introducing flaws, blind spots, and propagating or exaggerating biases. Synthetic datasets represent a fundamentally different form of evidence, and their value depends on how rigorously their limitations, sources of bias, and intended uses are defined and governed. In the absence of shared methodological standards and explicit regulatory expectations, synthetic data risks being adopted inconsistently, potentially reinforcing blind spots rather than mitigating them. We argue that progress in this field now requires a collective shift from exploring what synthetic data *can* do to agreeing on when, how, and for which purposes they *should* be used in cardiology research. Within clearly defined boundaries and under transparent validation and governance conditions, synthetic data can support innovation, data sharing, and equity without compromising scientific credibility or patient trust. Importantly, the challenges associated with synthetic data are amplified in cardiology, where clinically relevant information is often encoded in high-frequency signals, complex cardiac anatomy, and long-term disease trajectories.

To operationalize this governance agenda, we propose 4 concrete actions for the cardiology research community. First, mandatory three-layer validation for any publication using

synthetic cardiology data: (1) fidelity validation (statistical similarity, including signal-specific properties such as QRS morphology for ECG data), (2) utility validation (train-on-synthetic, test-on-real performance equivalence for the downstream task), and (3) privacy validation through adversary-aware testing, including membership inference and attribute inference evaluation. Second, mandatory subgroup reporting requires performance metrics should be disaggregated by relevant demographic subgroups, rather than reported

only at the population level. Third, intended-use scoping dictating that any synthetic dataset should document the specific purpose for which it was validated, consistent with the fit-for-purpose principle illustrated in [Figure 1](#). Fourth, engagement by cardiology societies and regulatory agencies to define domain-specific acceptability thresholds for synthetic evidence, beginning with lower-stakes applications such as model pretraining and simulation before moving to higher-stakes regulatory submissions.

Acknowledgments

The authors thank Pippa McKelvie-Sebileau, PhD, from Horiana for editing and proofreading. AI tools (Claude) were used to suggest language improvements within the manuscript.

Funding

No external funding was received for this study.

Data Availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during this study.

Authors' Contributions

Conceptualization: LB, SM, SL, PJ

Investigation: LB, SL, JJ, PJ

Methodology: LB, SM, VB

Supervision: PJ

Validation: SM

Visualization: LB

Writing—original draft: LB

Writing—review and editing: LB, SM, SL, JJ, VB, PJ

All authors approved the final version for submission.

Conflicts of Interest

None declared.

References

1. Rujas M, Martín Gómez Del Moral Herranz R, Fico G, Merino-Barbancho B. Synthetic data generation in healthcare: a scoping review of reviews on domains, motivations, and future applications. *Int J Med Inform*. Mar 2025;195:105763. [doi: [10.1016/j.ijmedinf.2024.105763](https://doi.org/10.1016/j.ijmedinf.2024.105763)] [Medline: [39719743](https://pubmed.ncbi.nlm.nih.gov/39719743/)]
2. Kokosi T, Harron K. Synthetic data in medical research. *BMJ Med*. 2022;1(1):e000167. [doi: [10.1136/bmjmed-2022-000167](https://doi.org/10.1136/bmjmed-2022-000167)] [Medline: [36936569](https://pubmed.ncbi.nlm.nih.gov/36936569/)]
3. Gonzales A, Guruswamy G, Smith SR. Synthetic data in health care: a narrative review. *PLOS Digit Health*. Jan 2023;2(1):e0000082. [doi: [10.1371/journal.pdig.0000082](https://doi.org/10.1371/journal.pdig.0000082)] [Medline: [36812604](https://pubmed.ncbi.nlm.nih.gov/36812604/)]
4. Kim H, Jang WS, Sim WS, et al. Synthetic data improve survival status prediction models in early-onset colorectal cancer. *JCO Clin Cancer Inform*. Jan 2024;8:e2300201. [doi: [10.1200/CCI.23.00201](https://doi.org/10.1200/CCI.23.00201)] [Medline: [38271642](https://pubmed.ncbi.nlm.nih.gov/38271642/)]
5. Kaabachi B, Despraz J, Meurers T, et al. A scoping review of privacy and utility metrics in medical synthetic data. *NPJ Digit Med*. Jan 27, 2025;8(1):60. [doi: [10.1038/s41746-024-01359-3](https://doi.org/10.1038/s41746-024-01359-3)] [Medline: [39870798](https://pubmed.ncbi.nlm.nih.gov/39870798/)]
6. Giuffrè M, Shung DL. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *NPJ Digit Med*. Oct 9, 2023;6(1):186. [doi: [10.1038/s41746-023-00927-3](https://doi.org/10.1038/s41746-023-00927-3)] [Medline: [37813960](https://pubmed.ncbi.nlm.nih.gov/37813960/)]
7. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med*. Jul 6, 2023;6(1):120. [doi: [10.1038/s41746-023-00873-0](https://doi.org/10.1038/s41746-023-00873-0)] [Medline: [37414860](https://pubmed.ncbi.nlm.nih.gov/37414860/)]
8. Jordon J, Szpruch L, Houssiau F, et al. Synthetic data - what, why and how? The Royal Society; 2022. URL: https://royalsociety.org/-/media/policy/projects/privacy-enhancing-technologies/Synthetic_Data_Survey-24.pdf [Accessed 2026-07-18]
9. Synthetic data. European Data Protection Supervisor. URL: https://www.edps.europa.eu/press-publications/publications/techsonar/synthetic-data_en [Accessed 2026-07-18]
10. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. *J Big Data*. 2019;6:60. [doi: [10.1186/s40537-019-0197-0](https://doi.org/10.1186/s40537-019-0197-0)]

11. Allasonniere S, Fraysse JL. Données de santé artificielles: analyse et pistes de réflexion [Article in French]. BOTdesign; 2024. URL: https://static.botdesign.net/docs/VF_Livre_blanco_Donn%C3%A9es_de_sant%C3%A9_artificielles-250424.pdf [Accessed 2026-07-18]
12. Cook DA, Triola MM. Virtual patients: a critical literature review and proposed next steps. *Med Educ*. Apr 2009;43(4):303-311. [doi: [10.1111/j.1365-2923.2008.03286.x](https://doi.org/10.1111/j.1365-2923.2008.03286.x)] [Medline: [19335571](https://pubmed.ncbi.nlm.nih.gov/19335571/)]
13. Grinstein J, Blanco PJ, Torii R, Goswami R. The virtual patient simulator: using hemodynamics and the cardioenergetic profile to optimize heart failure care. *JACC Basic Transl Sci*. Mar 2025;10(3):259-261. [doi: [10.1016/j.jacbts.2025.01.009](https://doi.org/10.1016/j.jacbts.2025.01.009)] [Medline: [40139863](https://pubmed.ncbi.nlm.nih.gov/40139863/)]
14. Venkatesh KP, Raza MM, Kvedar JC. Health digital twins as tools for precision medicine: considerations for computation, implementation, and regulation. *NPJ Digit Med*. Sep 22, 2022;5(1):150. [doi: [10.1038/s41746-022-00694-7](https://doi.org/10.1038/s41746-022-00694-7)] [Medline: [36138125](https://pubmed.ncbi.nlm.nih.gov/36138125/)]
15. Coorey G, Figtree GA, Fletcher DF, et al. The health digital twin to tackle cardiovascular disease-a review of an emerging interdisciplinary field. *NPJ Digit Med*. Aug 26, 2022;5(1):126. [doi: [10.1038/s41746-022-00640-7](https://doi.org/10.1038/s41746-022-00640-7)] [Medline: [36028526](https://pubmed.ncbi.nlm.nih.gov/36028526/)]
16. McDuff D, Curran T, Kadambi A. Synthetic data in healthcare. *arXiv*. Preprint posted online on Apr 6, 2023. [doi: [10.48550/arXiv.2304.03243](https://doi.org/10.48550/arXiv.2304.03243)]
17. Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. *arXiv*. Preprint posted online on Jun 10, 2014. [doi: [10.48550/arXiv.1406.2661](https://doi.org/10.48550/arXiv.1406.2661)]
18. Kingma DP, Welling M. Auto-encoding variational Bayes. *arXiv*. Preprint posted online on Dec 20, 2013. [doi: [10.48550/arXiv.1312.6114](https://doi.org/10.48550/arXiv.1312.6114)]
19. Peirlinck M, Costabal FS, Yao J, et al. Precision medicine in human heart modeling: perspectives, challenges, and opportunities. *Biomech Model Mechanobiol*. Jun 2021;20(3):803-831. [doi: [10.1007/s10237-021-01421-z](https://doi.org/10.1007/s10237-021-01421-z)] [Medline: [33580313](https://pubmed.ncbi.nlm.nih.gov/33580313/)]
20. Merle M, Collot F, Castelneau J, et al. MUSIC: cardiac imaging, modelling and visualisation software for diagnosis and therapy. *Appl Sci*. 2022;12(12):6145. [doi: [10.3390/app12126145](https://doi.org/10.3390/app12126145)]
21. Nakatani Y, Nuñez-García M, Cheniti G, et al. Preoperative personalization of atrial fibrillation ablation strategy to prevent esophageal injury: impact of changes in esophageal position. *J Cardiovasc Electrophysiol*. May 2022;33(5):908-916. [doi: [10.1111/jce.15447](https://doi.org/10.1111/jce.15447)] [Medline: [35274776](https://pubmed.ncbi.nlm.nih.gov/35274776/)]
22. Baowaly MK, Lin CC, Liu CL, Chen KT. Synthesizing electronic health records using improved generative adversarial networks. *J Am Med Inform Assoc*. Mar 1, 2019;26(3):228-241. [doi: [10.1093/jamia/ocy142](https://doi.org/10.1093/jamia/ocy142)] [Medline: [30535151](https://pubmed.ncbi.nlm.nih.gov/30535151/)]
23. Diaz-Pinto A, Alle S, Nath V, et al. MONAI Label: a framework for AI-assisted interactive labeling of 3D medical images. *Med Image Anal*. Jul 2024;95:103207. [doi: [10.1016/j.media.2024.103207](https://doi.org/10.1016/j.media.2024.103207)] [Medline: [38776843](https://pubmed.ncbi.nlm.nih.gov/38776843/)]
24. Azizi Z, Lindner S, Shiba Y, et al. A comparison of synthetic data generation and federated analysis for enabling international evaluations of cardiovascular health. *Sci Rep*. Jul 17, 2023;13(1):11540. [doi: [10.1038/s41598-023-38457-3](https://doi.org/10.1038/s41598-023-38457-3)] [Medline: [37460705](https://pubmed.ncbi.nlm.nih.gov/37460705/)]
25. McDuff D, Hernández J, Wood E, Liu X, Baltrusaitis T. Advancing non-contact vital sign measurement using synthetic avatars. *arXiv*. Preprint posted online on Oct 24, 2020. [doi: [10.48550/arXiv.2010.12949](https://doi.org/10.48550/arXiv.2010.12949)]
26. Williams MC, MacArthur JA, Forsyth R, Petersen SE. New ways to use imaging data in cardiovascular research: survey of opinions on federated learning and synthetic data. *Eur Heart J Imaging Methods Pract*. Jan 2025;3(1):qyaf012. [doi: [10.1093/ehjimp/qyaf012](https://doi.org/10.1093/ehjimp/qyaf012)] [Medline: [40066149](https://pubmed.ncbi.nlm.nih.gov/40066149/)]
27. Myles P, Ordish J, Tucker A. The potential synergies between synthetic data and in silico trials in relation to generating representative virtual population cohorts. *Prog Biomed Eng*. 2023;5:013001. [doi: [10.1088/2516-1091/acafbf](https://doi.org/10.1088/2516-1091/acafbf)]
28. Angoulvant D, Granjeon-Noriot S, Amarenco P, et al. In-silico trial emulation to predict the cardiovascular protection of new lipid-lowering drugs: an illustration through the design of the SIRIUS programme. *Eur J Prev Cardiol*. Nov 11, 2024;31(15):1820-1830. [doi: [10.1093/eurjpc/zwae254](https://doi.org/10.1093/eurjpc/zwae254)] [Medline: [39101472](https://pubmed.ncbi.nlm.nih.gov/39101472/)]
29. Lutfi H, Spittler T, Ibrahim HM. Using synthetic data and machine learning to predict heart rate and enable lifestyle recommendations. *J Med Artif Intell*. Sep 2024;7. [doi: [10.21037/jmai-24-35](https://doi.org/10.21037/jmai-24-35)]
30. Skandarani Y, Lalande A, Afilalo J, Jodoin PM. Generative adversarial networks in cardiology. *Can J Cardiol*. Feb 2022;38(2):196-203. [doi: [10.1016/j.cjca.2021.11.003](https://doi.org/10.1016/j.cjca.2021.11.003)] [Medline: [34780990](https://pubmed.ncbi.nlm.nih.gov/34780990/)]
31. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? In: *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery; 2021:610-623. [doi: [10.1145/3442188.3445922](https://doi.org/10.1145/3442188.3445922)]
32. Reflection paper on the use of artificial intelligence (AI) in the medicinal product lifecycle. European Medicines Agency. 2024. URL: https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-use-artificial-intelligence-ai-medicinal-product-lifecycle_en.pdf [Accessed 2025-02-28]

33. Boraschi D, van der Schaar M, Costa A, Milne R. Governing synthetic data in medical research: the time is now. *Lancet Digit Health*. Apr 2025;7(4):e233-e234. [doi: [10.1016/j.landig.2025.01.012](https://doi.org/10.1016/j.landig.2025.01.012)] [Medline: [39984419](https://pubmed.ncbi.nlm.nih.gov/39984419/)]
34. Hendricks-Sturup R, Nafie M. 64 Synthetic data generation using generative AI to support biomedical innovation: a health policy perspective. *J Clin Transl Sci*. 2026;10(s1):20. [doi: [10.1017/cts.2026.10290](https://doi.org/10.1017/cts.2026.10290)]
35. Foraker R, Morrow JD, Johnson JA, Wilcox AB, Forster AJ, Payne PR. Understanding synthetic data: artificial datasets for real-world evidence. *BMJ Evid Based Med*. May 21, 2026;31(3):148-151. [doi: [10.1136/bmjebm-2024-113617](https://doi.org/10.1136/bmjebm-2024-113617)] [Medline: [40602833](https://pubmed.ncbi.nlm.nih.gov/40602833/)]
36. Arora A, Arora A. Synthetic patient data in health care: a widening legal loophole. *Lancet*. Apr 23, 2022;399(10335):1601-1602. [doi: [10.1016/S0140-6736\(22\)00232-X](https://doi.org/10.1016/S0140-6736(22)00232-X)] [Medline: [35358423](https://pubmed.ncbi.nlm.nih.gov/35358423/)]
37. Artificial intelligence-enabled device software functions: lifecycle management and marketing submission recommendations. U.S. Food & Drug Administration. 2025. URL: <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/artificial-intelligence-enabled-device-software-functions-lifecycle-management-and-marketing> [Accessed 2026-07-08]

Abbreviations

CT: computed tomography
ECG: electrocardiogram
GAN: generative adversarial network
GDPR: General Data Protection Regulation
HIPAA: Health Insurance Portability and Accountability Act
ML: machine learning
MRI: magnetic resonance imaging
PPG: photoplethysmogram
VAE: variational autoencoder

Edited by Andrew Coristine; peer-reviewed by Anup Menon Nandialath, Saiqing Guan; submitted 05.Feb.2026; final revised version received 18.Jun.2026; accepted 19.Jun.2026; published 31.Jul.2026

Please cite as:

Baschet L, Marque S, Locret S, Jenssen J, Barbet V, Jourdain P
Addressing the Challenges in Using Synthetic Data for Health Research: Application to Cardiology
JMIR Cardio 2026;10:e92930
URL: <https://cardio.jmir.org/2026/1/e92930>
doi: [10.2196/92930](https://doi.org/10.2196/92930)

© Louise Baschet, Sebastien Marque, Stéphane Locret, Jens Jenssen, Vanessa Barbet, Patrick Jourdain. Originally published in *JMIR Cardio* (<https://cardio.jmir.org>), 31.Jul.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in *JMIR Cardio*, is properly cited. The complete bibliographic information, a link to the original publication on <https://cardio.jmir.org>, as well as this copyright and license information must be included.