Future Healthcare Journal 13 (2026) 100559

Contents lists available at ScienceDirect

Future Healthcare Journal

journal homepage: www.sciencedirect.com/journal/future-healthcare-journal

Opinion

The emerging use of synthetic medical data to support clinical trials

Anmol Arora, MBBChir[a,b,*], Dave Chuter[c], Mark Gaze, MD[a]

[a] Department of Oncology, University College London Hospitals NHS Foundation Trust, London, UK
[b] School of Clinical Medicine, University of Cambridge, Cambridge, UK
[c] Independent Patient Advocate

ARTICLE INFO

Keywords:
Generative artificial intelligence
Randomised clinical trial
Regulation
Patient and public involvement
Synthetic control arms

ABSTRACT

Synthetic data is emerging as a potential tool in healthcare, with possible applications in data sharing and synthetic control arms (SCAs). Interest in synthetic data is growing, but use in a clinical trial setting remains uncertain both methodologically and from a regulatory perspective. This opinion article discusses the use cases of synthetic data in trials and highlights key barriers to implementation, including validation, bias, governance, and public trust.

Synthetic data is emerging as a potential tool in healthcare, with possible applications in data sharing and synthetic control arms (SCAs). Interest in synthetic data is growing, but use in a clinical trial setting remains uncertain both methodologically and from a regulatory perspective. This Opinion discusses the use cases of synthetic data in trials and highlights key barriers to implementation, including validation, bias, governance, and public trust.

Randomised clinical trials are the foundation of evidence-based medicine, yet they are expensive to run and slow to complete. Control arms are essential in randomised trials to demonstrate whether there is a significant difference in measured outcomes, for example disease control or toxicity, between patients who received the experimental treatment and those who did not. However, in small patient populations, for example rare cancers, the need for a control group can be burdensome and a barrier to running a sufficiently powered randomised controlled trial. Where a gold-standard randomised controlled trial is not feasible, investigators have often used historical or external data as a control arm, although these are generally considered less reliable and more open to bias than randomisation.

External controls use data from patients outside the trial, whereas synthetic control arms use statistical or machine-learning methods to algorithmically construct a comparator from external, historical, real-world, or synthetic data. Augmentation with historical or external data has been used successfully but is susceptible to bias and is typically reserved for settings where randomisation is impractical. Synthetic control arms (SCAs), however, have proved technically challenging and have not consistently met regulatory expectations.[1] More recently, advances in synthetic data have reignited enthusiasm that machine learning may enable SCAs, not only in population-limited studies but in routine practice. 'Synthetic data' usually means computer-generated records designed to resemble real datasets without any individual record directly linkable back to the original dataset. In medicine, synthetic data has already started to be used for protecting patient privacy and enabling more open data sharing. The Information Commissioner's Office describes synthetic data as a privacy-enhancing technology, a category of innovation that also includes federated learning, homomorphic encryption and trusted research environments. These technologies are increasingly emerging as tools to provide technical infrastructure for secure data sharing.[2] Consensus-based governance frameworks have emerged in the UK to promote accountability, secure data handling and compliance with UK GDPR.[3–5] It has been suggested that using synthetic data to augment control arms in clinical trials could help accelerate trial development and increase access to experimental therapies by reducing the number receiving control products. SCAs need not be built from synthetic data but there is growing interest in the idea that synthetic data may also help trial teams share and analyse otherwise inaccessible patient-level data to refine SCAs. Ultimately, if credible comparator data could be assembled without recruiting as many contemporaneous controls, trials might become faster, cheaper, and more acceptable to some patients, particularly where enrolment is difficult.

Over the past 50 years, the permissibility of using historical data in clinical trials has gradually expanded. In 1976, Stuart Pocock proposed strict criteria for the acceptability of using historical clinical trial data

This article reflects the opinions of the author(s) and should not be taken to represent the policy of the Royal College of Physicians unless specifically stated.

* Corresponding author at: Department of Oncology, University College London Hospitals NHS Foundation Trust, London, UK.
E-mail address: anmol.arora@ucl.ac.uk (A. Arora).

https://doi.org/10.1016/j.fhj.2026.100559
Received 16 June 2026; Received in revised form 11 August 2026; Accepted 24 August 2026
2514-6645/© 2026 The Author(s). Published by Elsevier Ltd on behalf of Royal College of Physicians. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

to generate control arms, including that the data must have been collected at the same institution.[6] Since then, advances in clinical trial methodologies have iteratively stretched beyond these confines, initially by sourcing data either from existing databases or from concurrent clinical trials under different conditions. In recent years, real-world data use in clinical trials has advanced, including through Bayesian borrowing, which is an established method of combining evidence from related trial populations.[7] Synthetic data may enable increased access to real-world data and could even involve the creation of new datapoints to match participants in the control arm, rather than simply reusing real datapoints from historical trials or real-world cases.

Perhaps one of the biggest issues affecting the use of SCAs is 'data drift', the idea that there may be changes in population characteristics over time, including those of which we are unaware or unable to measure.[8] Algorithms are limited to reproducing characteristics and correlations present in their training data. This dictates that the data used to train the models must be recent, due to potential changes in standards of care, care pathways, concomitant comorbidities and coding practice drift, all of which could be more damaging to a control arm's fidelity than biological drift. When synthetic data is being used to facilitate a synthetic control arm, the risks of drift may be amplified as they could propagate through both the synthetic data generation and then the SCA generation process. If the source data no longer reflects current care, even a very sophisticated algorithm could generate a control group that appears plausible but is quietly encoding outdated practice. Synthetic control arms can be evaluated in three domains:

1. Statistical fidelity, for which methods exist such as scatter plots, correlation matrices, or principal component analysis.[3]
2. Disclosure risk, in certain settings where synthetic data is being used as a privacy enhancing technology to enable the use of real-world evidence.
3. Causal validity, which measures whether the SCA is producing reliable treatment-effect conclusions.

Metrics in all three domains are limited by our current understanding of the data and in the future, we may learn about differences between real and synthetic data that we are unable to measure today. Whilst there have been advancements in measuring the first two domains, whether high-fidelity synthetic datapoints can act as counterfactual patients for the purposes of a clinical trial remains an open question.

There is precedent for FDA approval of drugs which have provided historical controls as a comparative framework in the absence of a traditional control arm. The FDA recognises that conventional double-blind randomised controlled trials are sometimes not feasible and their accelerated approval pathway allows for the early approval of drugs that treat an unmet medical need based on an effect on a surrogate endpoint or an intermediate clinical endpoint. The pathway allows for unconventional trial designs, including for single arm trials to be provided as evidence, where it is impractical to run a control group, although there is still an expectation that post-marketing confirmatory trials would be required post-approval.[9] The methodological bar is rightly high as weak control arms can have long-term effects on prescribing decisions, treatment pathways and expose patients to ineffective or harmful treatments. Furthermore, there is notable variability in the methodology and reporting of SCAs, with a lack of trial-specific reporting standard. The concept of 'data cards' has been suggested, whereby dataset creators would release a standardised set of information dictating the scope of use of the synthetic data, its underlying methodology and limitations.[10]

Whilst SCA research continues to evolve, there are other ways that synthetic data can more immediately help drug development. For example, the privacy-preserving usage of synthetic data may help to advance data sharing between academic institutions, hospital systems and pharmaceutical companies. Synthetic cohorts can also support trial design, including power and sample size exploration under realistic covariate distributions without needing to analyse identifiable patient data.

Despite positive regulatory signals,[11] it can be risky to run an isolated SCA with no concurrent validation against a real arm due to the threat of subsequently learning of data drift when it is too late. In the short-term, perhaps the risk-averse approach would be to concurrently run synthetic and real control arms whilst establishing evidence that they can be effective, particularly in conditions with small population sizes, such as rare disease. The ethical imperative to improve the effectiveness of research studies, coupled with the financial incentives that SCAs offer, suggest that research in this field is likely to continue to grow. The key question is not whether synthetic records can resemble real patients, but whether they can support causal claims with calibrated uncertainty at a standard comparable to randomisation. It is difficult to discern whether SCAs are producing genuinely correct counterfactual outcomes as opposed to just a realistic-looking control distribution, although the advancement of explainable synthetic data generation may help to alleviate these concerns.[12] Whilst there have been examples of data-owners releasing synthetic data in academic contexts, the release of synthetic data to pharmaceutical companies to support synthetic control arm production is less well described. This will require considerations of data sovereignty and evidence of compliance with best-practice standards and data protection legislation.

The ethical case is more complicated than sometimes projected. It is often said that SCAs would allow more participants to receive the experimental treatment rather than placebo or standard care. That may be attractive in some settings, but it is not automatically in patients' best interests. Patients have different views and some prefer the protection of a conventional control arm if the experimental therapy carries uncertainty, additional toxicity, or a longer treatment burden without proven evidence of better outcomes. Involving patients early helps make sure trials are designed in a way that reflects what matters most to them, such as quality of life, physical and emotional well-being, safety, and clear information about risks. Effective patient and public involvement (PPI) can only be achieved by embedding PPI throughout the research process. This means involving patients from the earliest stages of study development and design, through delivery and oversight, to interpretation, dissemination and implementation of the findings. Meaningful PPI is achieved by working in partnership with individual patient advocates and patient organisations, recognising the complementary perspectives and expertise they each bring. This includes working with patients to design and develop the study, explain how SCAs work in simple terms, and involve them in decision-making groups. Trial participants should also be informed when consenting to join a study that the comparator may consist of computer-generated patient trajectories, rather than a group of real patients receiving a control product. It is also important to include a wide range of voices so that the research reflects real-world populations.

At present, there are clear use cases for how the generation of synthetic data can augment existing pipelines. A next logical step would be to generate SCAs for a wide range of randomised trials to run concurrently with real control arms to establish their validity, especially when generated prospectively. Though the timeline for routine use of synthetic control arms appears uncertain, we should not assume that it is impossible, or indeed far away.

CRediT authorship contribution statement

Anmol Arora: Writing – original draft, Validation, Methodology, Conceptualization. **Dave Chuter:** Writing – review & editing, Validation, Methodology, Conceptualization. **Mark Gaze:** Writing – review & editing, Validation, Supervision, Methodology, Conceptualization.

Ethics approval and consent to participate

This project did not require ethical approval.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability

All relevant data is included within the manuscript.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Anmol Arora is an Associate Editor of the Future Healthcare Journal. He had no involvement in the peer review of this article and had no access to information regarding its peer review. Full responsibility for the editorial process for this article was delegated to another journal editor. Anmol Arora is Co-Chair of the UK Synthetic Data Community Group (unpaid). Dave Chuter is a Director of 'use MY data' (unpaid).

References

1. Wang X, Dormont F, Lorenzato C, Latouche A, Hernandez R, Rouzier R. Current perspectives for external control arms in oncology clinical trials: analysis of EMA approvals 2016–2021. *J Cancer Policy*. 2023;35:100403. https://doi.org/10.1016/j.jcpo.2023.100403
2. Konda B, Kasula VK, Yenugula M, Yadulla AR, Addula SR. Homomorphic encryption and federated attribute-based multi-factor access control for secure cloud services in integrated space-ground information networks. *Int J Commun Inf Technol*. 2022;3(2):33–40. https://doi.org/10.33545/2707661X.2022.v3.i2a.103
3. UK Synthetic Data Community Group. VSTAR framework: perspectives & recommendations on the development of synthetic datasets in trusted research environments [Internet]; 2026 [cited 2026 May 27]. Available from: 〈https://portal.dementiasplatform.uk/wp-content/uploads/2026/01/VSTAR-Framework-Report-Final.pdf〉.
4. Arora A, Hotchkiss L, Myles P, et al. *Synthetic Data Release Framework under UK Data Protection Law [Internet]*. Zenodo; 2026https://doi.org/10.5281/zenodo.20535817
5. Privacy-enhancing technologies (PETs) [Internet]. ICO; 2025 [cited 2026 Jul 16]. Available from: 〈https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/privacy-enhancing-technologies/〉.
6. Pocock SJ. The combination of randomized and historical controls in clinical trials. *J Chronic Dis*. 1976;29(3):175–188. https://doi.org/10.1016/0021-9681(76)90044-8
7. Struebing A, McKibbon C, Ruan H, et al. Augmenting external control arms using Bayesian borrowing: a case study in first-line non-small cell lung cancer. *J Comp Eff Res*. 2024;230175(5):e230175. https://doi.org/10.57264/cer-2023-0175 PubMed PMID: 38573331; PubMed Central PMCID: PMC11036906.
8. Chevret S, Timsit JF, Biard L. Challenges of using external data in clinical trials-an illustration in patients with COVID-19. *BMC Med Res Methodol*. 2022;22:321. https://doi.org/10.1186/s12874-022-01769-5 PubMed PMID: 36522698; PubMed Central PMCID: PMC9753019.
9. FDA. Clinical trial considerations to support accelerated approval of oncology therapeutics guidance for industry [Internet]; 2023 [cited 2026 Mar 6]. Available from: 〈https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-trial-considerations-support-accelerated-approval-oncology-therapeutics〉.
10. Pushkarna M, Zaldivar A, Kjartansson O. *Data cards: purposeful and transparent dataset documentation for responsible AI. Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT '22) [Internet]*. New York, NY, USA: Association for Computing Machinery; 2022https://doi.org/10.1145/3531146.3533231 [cited 2026 Mar 6].
11. EMA. Establishing efficacy based on single-arm trials submitted as pivotal evidence in a marketing authorisation [Internet] [cited 2026 Mar 6]. Available from: 〈https://www.ema.europa.eu/en/establishing-efficacy-based-single-arm-trials-submitted-pivotal-evidence-marketing-authorisation〉.
12. Valina L, Teixeira B, Reis A, Vale Z, Pinto T. *Explainable artificial intelligence for deep synthetic data generation models. Proceedings of the IEEE Conference on Artificial Intelligence (CAI) [Internet]*. 2024; 2024:555–556. https://doi.org/10.1109/CAI59869.2024.00109 [cited 2026 Jul 16]. https://ieeexplore.ieee.org/document/10605530.