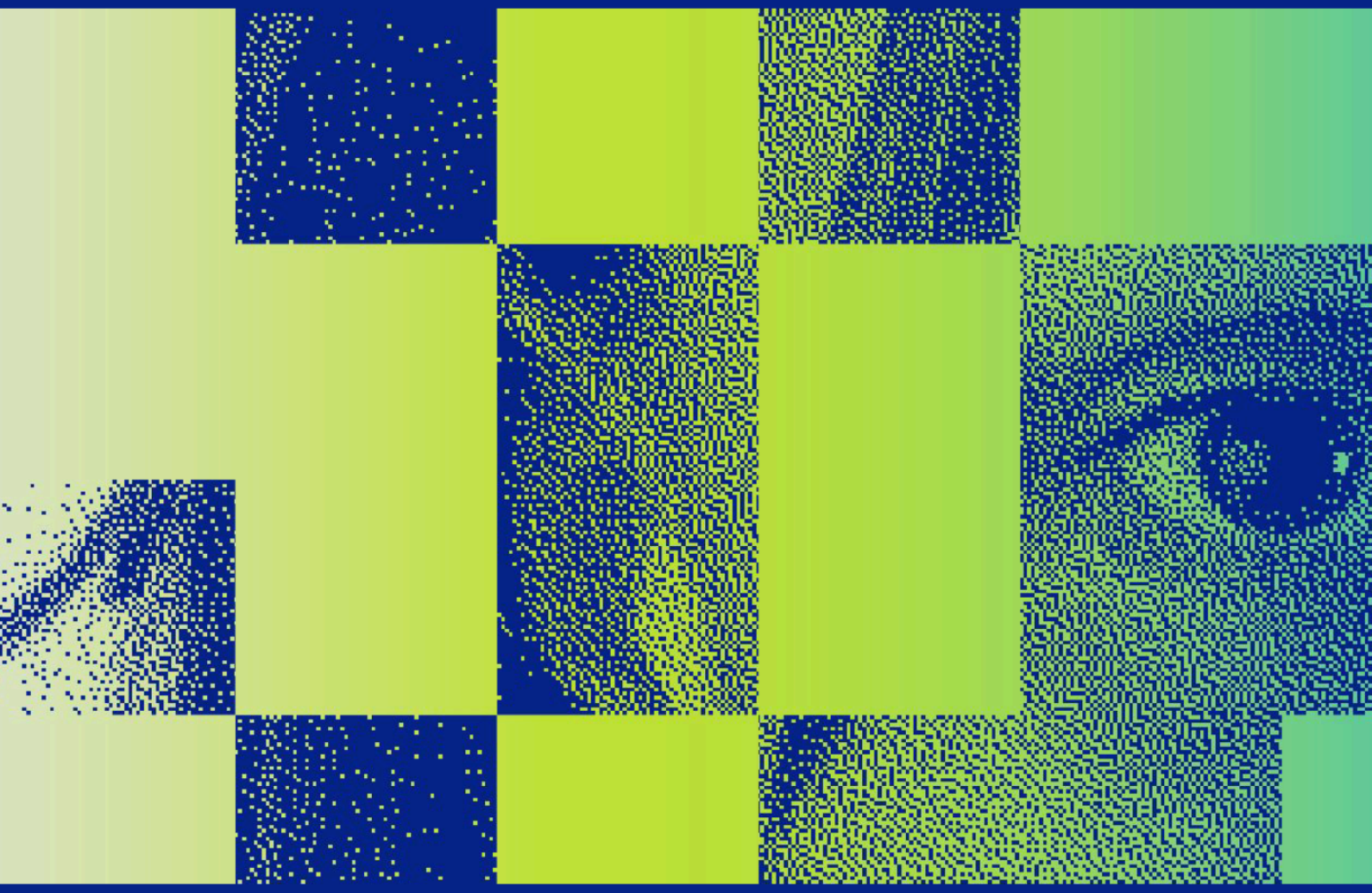


Towards the ethical use of synthetic data in health research

ODI Research
ADVANCING TRUST IN DATA

August 2025

Contents

Contents 1
About 1
Executive summary 2
Introduction 3
Methodology 5
Background 6
- Synthetic data 6
- Synthetic data as a means of preserving privacy 7
- Synthetic data as a means of increasing the availability of data 8
- Existing guidance on the creation and use of synthetic data 9
  - Existing guidance within the UK 10
  - Guidance from outside the UK 12
- Ethical approaches to synthetic data in health research settings 14

Findings 15
- Theme 1: Define use cases 15
- Theme 2: Challenges 17
- Theme 3: Ethical and legal frameworks 19

Discussion 20
- Fidelity 20
- Decision-making 21
- Regulators, ethical guidance, and the provision of data 22
- Patients, consent, and reputational risk 24
- Broader ethical concerns 25
- Future developments 26

Conclusion 27
- Recommendations 27

Acknowledgements 29

About

This report has been researched and produced by the Open Data Institute, and published in August 2025. Its authors were Calum Inverarity and Ben Snaith. If you want to share feedback by email or would like to get in touch, contact the privacy-enhancing technologies programme at pets@theodi.org.

Executive summary

Synthetic data has been gaining traction in recent years as a means by which to increase the availability of data that, while deemed valuable, may be sensitive in nature and/or in limited supply. Within the context of health research, the use of synthetic data presents potentially transformative opportunities, given that patient data is highly sensitive and access can therefore be subject to more stringent safeguards. This raises questions on whether it is ethical to use synthetic data in specific circumstances, and how such assessments can be carried out.

In our research, we explored current practices and attitudes towards the use of synthetic data in health research in the United Kingdom. We sought to determine where using synthetic data holds the greatest potential, while acknowledging where using synthetic data may raise ethical questions with no clearly defined answers.

Using both desk-based and primary qualitative research, we engaged with those working within the health research sector to identify current opportunities for the use of synthetic data, including those that are presently feasible and those that require further research and guidance to be realised.

While we found that the potential risks that surround the use of synthetic data are generally well documented and understood by researchers within the health research sector, guidance on its ethical use is currently underexplored. This is reflected by the emphasis currently placed on managing the privacy/utility trade-off when generating synthetic data, while comparatively less attention has been paid to possible downstream effects of using synthetic data.

Introduction

Synthetic data is part of a suite of tools and practices often referred to as privacy-enhancing technologies (PETs).[1] These technologies can be used to conduct analyses on sensitive data that can be difficult for organisations to share, or would otherwise need to be kept closed.

These technologies can therefore be particularly helpful in the context of conducting health research, where sensitive patient data can be valuable, but also subject to greater protections. Some examples in which PETs have proven useful for conducting health research include the OpenSAFELY[2] and SAIL Databank[3] trusted research environments (TREs), which have enabled researchers to access health and social care records for the purpose of conducting research projects. These TREs are useful illustrations of how a combination of PETs can be used, as well as the analyses that they can enable when configured and deployed correctly. OpenSAFELY has enabled more than 180 approved research projects[4] and SAIL Databank has facilitated more than 700 published papers in medical journals at the time of writing.[5]

Recently, the possibility of using synthetic data in health research settings has attracted more attention, given that it can potentially help increase the availability of training data, while also preserving data privacy. Examples of specific issues that synthetic data has been shown to hold potential in overcoming include limited availability of clinical trial data[6][7] and scaling training datasets for the purpose of designing and assessing X-ray-based intelligence systems.[8] While these promising studies demonstrate the potential of using synthetic data to address data challenges within health research, there remain technical and ethical limitations to synthetic data that currently limit its broader use in certain contexts.

For synthetic data to address these challenges, it must be to some extent similar to, yet different from, the original data it is based on. There are three key attributes when considering synthetic data generation:

---

[1] ODI (2023), 'Privacy enhancing technologies (PETs)'
[2] University of Oxford for the Bennett Institute for Applied Data Science (2025), 'About OpenSAFELY'
[3] ODI (2024), 'PETs in Practice 2'
[4] University of Oxford for the Bennett Institute for Applied Data Science (2025), 'Approved Projects'
[5] SAIL Databank (2023), 'Publications'
[6] Azizi, Z., et al. (2021), 'Can Synthetic Data Be a Proxy for Real Clinical Trial Data? A Validation Study'
[7] Austin, J.A., et al. (2024), 'Decades in the Making: The Evolution of Digital Health Research Infrastructure Through Synthetic Data, Common Data Models, and Federated Learning'
[8] Gao, C., et al. (2023), 'Synthetic Data Accelerates the Development of Generalizable Learning-Based Algorithms for X-Ray Image Analysis'

1. **Utility:** often determined by the usefulness of the new data for its intended task. Defining this will often involve comparison of models run on real vs synthetic data.[9]

2. **Fidelity:** often considered the 'realism' of the synthetic data, or how closely it resembles the original data in terms of aspects such as statistical distributions of variables.[10]

3. **Privacy:** the extent to which the synthetic data reveals information about the original data it is based on.[11]

Authors such as Jordon et al.[12] note that utility and fidelity can often be conflated. However, these are distinct attributes. This is significant, as trade-offs are often necessary when generating synthetic data, such as prioritising fidelity of the data over privacy, or vice versa. For those responsible for generating the synthetic data, this is inherently an ethical consideration of whether it is imperative that the data is as lifelike as possible, even if this potentially decreases the overall privacy of the data, thus exposing the underlying data subjects to increased risks of reidentification. This presents a pressing problem in the use of synthetic data, in which different interpretations and justifications can be made for whether prioritising one attribute in a synthetic dataset over another is necessary.

This research focused on these limitations in an attempt to identify opportunities in which synthetic data could be used more widely to enable health research through the provision of additional guidance and/or policy. We took an ethics-based approach to consider current prevailing attitudes and efforts towards using synthetic data within health research, to determine what can be considered as optimal in terms of responsible use, in absence of definitive guidance.

This report was made possible by the generous support of the Responsible AI team at GSK.

[9] Jordon, J., et al. (2022), 'Synthetic Data – What, Why and How?'
[10] Adams, T. et al. (2024), 'On the Trade-Off between Fidelity, Utility and Privacy of Synthetic Patient Data'
[11] Jordon, J., et al. (2022), ibid
[12] ibid.

Methodology

We started by conducting desk research in which we looked at current guidance from UK government bodies on the use of synthetic data. This also involved consideration of related frameworks relevant to synthetic data, such as data and AI frameworks. Throughout the course of this initial period, we also engaged with government officials and fellow researchers to test hypotheses.

The desk research then informed the first of our data collection activities, which consisted of six interviews with subject matter experts. The interviewees were a mixture of individuals working within health research capacities and from relevant public sector bodies, including government departments and the NHS. These interviews provided the opportunity to gather data on current opportunities and challenges to the use of synthetic health data, where additional guidance on synthetic data is necessary, and in which settings the use of synthetic data in health research is most desirable and practical.

The second data collection activity was a Data Ethics Canvas workshop with invited participants from government departments and agencies, industry and civil society. This workshop explored the ethical considerations necessary when using synthetic data within specific health research use cases. We sought to develop a fuller understanding of current views on how synthetic data is, and can be, used in health research, including the potential benefits of the use of synthetic data, plus the ethical and technical concerns currently limiting broader use, and how these may be overcome.

All views expressed in these data collection activities were personal and unrepresentative of the participants' organisations.

Through these activities, we collected qualitative data, upon which we conducted thematic analysis, which is included within the following sections of the report.

Background

At the ODI, we have been looking into how PETs can enable safe and secure data use for positive economic, social and environmental impact.[13] We have considered the contextual factors that contribute to the successful implementation of a PET,[14] as well as the sociotechnical aspects that can influence willingness to adopt them.[15] Through our efforts, we have aimed to increase understanding of these technologies – particularly among those who are unfamiliar with them – and the challenges that they can help address, as well as their current limitations. Part of these considerations draw upon our wider thinking on responsible data governance, posing questions beyond simply whether a PET maintains privacy, such as: what value is created through the adoption of a PET, and who stands to benefit from its use? Ethical questions such as these go beyond consideration of whether it is possible for PETs to overcome issues of data access and availability to examine whether it is appropriate for the PET to be used, and the potential consequences of doing so.

The use of synthetic data in health research settings provides a rich opportunity to explore some of these ethical questions, given the sensitivity of the data in question and the potential benefits that can be gained from increasing access to this data while preserving privacy.

Synthetic data

The concept of synthetic data is by no means new. Since it was first raised as a potential way to mitigate sensitive data disclosure risks as far back as 1993,[16] technological advances have led to the new methods of synthetic data generation beyond the use of traditional statistical models, to more recent use of advanced techniques based on deep-learning models.[17] It is now acknowledged that it is possible to create highly privacy-preserving, high-utility synthetic datasets that can aid the development of healthcare AI applications.[18]

While the theory and potential applications of synthetic data are increasingly well-established, concerns remain that synthetic data can be poorly generated,

[13] ODI (2023), 'Privacy enhancing technologies (PETs)'
[14] ODI (2024), 'Fostering a thriving PETs ecosystem'
[15] ODI (2025), 'Trust and transparency in privacy-enhancing technologies'
[16] Rubin, D.B. (1993), 'Discussion: Statistical disclosure limitation' Journal of Official Statistics, 9:462–468
[17] A full list of methods of synthetic data generation can be found in: Hao, S., et al. (2024), 'Synthetic Data in AI: Challenges, Applications, and Ethical Implications'
[18] I-Hsien Kuo, N. et al. (2023), 'Generating Synthetic Clinical Data That Capture Class Imbalanced Distributions with Generative Adversarial Networks: Example Using Antiretroviral Therapy for HIV'

leading to possible risks of reidentification of underlying data subjects,[19] or the perpetuation of biases within original datasets, rather than amelioration.[20] We will cover the remaining challenges with synthetic data in greater detail further on in this report. However, it is necessary at this stage to note that generating and using synthetic data are not yet entirely straightforward processes.

Interest in synthetic data has increased significantly in the past decade in response to a number of developments, most notably the greater attention now placed on data protection and the present focus on artificial intelligence. These two factors speak to two of the critical functions of synthetic data: to preserve privacy and increase the availability of data.

Synthetic data as a means of preserving privacy

Throughout the past decade, legislation such as the General Data Protection Regulation (GDPR) has been implemented, partly to strengthen individuals' fundamental rights in the digital age.[21] GDPR – in part – aims to ensure greater protection of personal data, which is certain data that relates to an individual. Furthermore, extra protections are granted to what is called 'special category data', which includes data such as 'genetic data; biometric data (where used for identification purposes); [and] data concerning health'.[22]

As synthetic data seeks to represent the structure, properties and relationships of actual data, while not containing any real information about individuals, it can help organisations comply with data protection legislation, while getting utility from the sensitive data. Furthermore, the UK's Information Commissioner's Office (ICO) notes that the use of synthetic data – where possible – instead of collecting more personal information can also aid compliance with the data minimisation principle, as it reduces the processing of personal information.[23] Therefore using synthetic data can actually be advisable for organisations, as it can potentially reduce risks associated with sharing data, if handled appropriately.

One illustration of how synthetic data can be used to preserve privacy while increasing utility of the underlying data is the synthetic dataset created by the Office for National Statistics (ONS) to support the onboarding of users and tooling of its Integrated Data Service (IDS).[24] The ONS has made this synthetic dataset available to prospective users of its data at an earlier stage in the accreditation process, which would be made available to researchers within the

[19] Giuffrè, M., and Shung, D.L. (2023), 'Harnessing the Power of Synthetic Data in Healthcare: Innovation, Application, and Privacy'
[20] Wyllie, S., et al. (2024), 'Fairness Feedback Loops: Training on Synthetic Data Amplifies Bias'
[21] European Commission (nd), 'Legal framework of EU data protection'
[22] Information Commissioner's Office (2024), 'What is special category data?'
[23] Information Commissioner's Office (2023), 'Privacy-enhancing technologies (PETs)'
[24] Office for National Statistics (2023), 'Enabling Data Access through Privacy Preserving Synthetic Data'

ONS's trusted research environment (TRE) once they had received full accreditation and granted access, which can be a time and resource intensive process for both the ONS and applicant. The generated synthetic dataset therefore enables the researcher to get a sense of the underlying data in advance of approval, while allowing them to familiarise themselves with the TRE in which they would conduct their analysis. This example demonstrates how sensitive data can be made available to accelerate existing research practices through the use of privacy-preserving synthetic data.

It should, however, be noted that in spite of the demonstrable privacy-preserving attributes of synthetic data, many regulators and lawyers remain cautious in their approach to synthetic data, as there are many factors that must be considered to ensure that the privacy of the original data is not compromised.[25] This current lack of consensus could potentially be viewed as stymying the broader use of synthetic data, both within health research and beyond.

Synthetic data as a means of increasing the availability of data

More recently, interest has grown substantially in synthetic data due to its potential use in the training of machine-learning and AI models.[2627] This is partly due to the aforementioned privacy-preserving attributes of synthetic data, and to the fact that synthetic data can be generated to address issues relating to data quality and/or quantity, such as where data is unbalanced, scarce, or too costly or challenging to collect.[28] These are similar to the challenges currently faced in the development of medical AI and health research, where obtaining informed consent can be challenging, certain populations can be underrepresented, and it can be unethical to perform experimentation.[29]

Researchers are already exploring the potential of using synthetic data to tackle these barriers in health research. Specifically, in the case of the ethics of experimentation, synthetic data could potentially be used as a proxy for an empirical control group in instances of trials with a low number of patients.[30] The authors of this research acknowledge, however, that there are uncertainties in their proposed method that could create a risk of original data being disclosed. Further efforts are therefore required to understand and mitigate this risk.

[25] Mitchell, C. and Redrup Hill, E. (2023), 'Are synthetic health data 'personal data'? - PHG Foundation'
[26] Kohli, S. (2021), 'Why Synthetic Data Could Be Useful for a Government Department'
[27] Jordon, J., et al. (2022), 'Synthetic Data – What, Why and How?'
[28] ODI (2024), 'Policy intervention 4: Ensuring broad access to data for training AI models'
[29] Carey, S., Pang, A. and de Kamps, M. (2024), 'Fairness in AI for Healthcare'
[30] Elvatun, S, et al. (2025), 'Synthetic Data as External Control Arms in Scarce Single-Arm Clinical Trials'

In a related vein, organisations such as the Tony Blair Institute for Global Change (TBI) have proposed that synthetic data can help the government to deliver its proposed National Data Library. Specifically, the TBI suggests, 'by generating AI-ready datasets that preserve statistical accuracy without exposing personal records, synthetic data will enable advanced analytics while adhering to strict governance and privacy standards'.[31] In practice, this could take a similar form to the ONS's use of synthetic data as part of its IDS, or, if less sensitive data was made available, this could be provided to organisations for research and development purposes. While the TBI does not identify specific types of data that synthetic alternatives could be generated to mimic, it acknowledges that synthetic data is not without risks and therefore requires continued technological development and learning from the lessons of existing synthetic data safeguards used by TREs.

Therefore, despite burgeoning interest in the prospective uses of synthetic data – both to promote innovation and to preserve privacy – obstacles remain that serve to provide friction to wider use. In the following section, we explore some of these themes based on available guidance and frameworks on the use of synthetic data.

Existing guidance on the creation and use of synthetic data

The attention on synthetic data in the preceding decades has led to the creation of a number of definitions.[32] To remove ambiguity for the purposes of this report, we use the United States Food & Drug Administration's (FDA) definition, which is considered a point of reference for many organisations, including the UK's Medicines and Healthcare products Regulatory Agency (MHRA). This definition considers synthetic data:

[31] Tony Blair Institute for Global Change and the Entrepreneurs Network (2025), 'Governing in the Age of AI: Building Britain's National Data Library'
[32] For a full account, see: Myles, P. et al. (2023), 'Are Synthetic Health Data 'Personal Data'?'

> *"Data that have been created artificially (e.g., through statistical computer simulation) so that new values and/or data elements generated. Generally, synthetic data are intended to represent structure, properties and relationships seen in actual patient da that they do not contain any real or specific information about i*
>
> — US Food & Drug Administration[33]

This definition has been adapted from a number of others to cover both historic methods of construction, such as through modelling, as well as increasingly popular newer methods, such as generative adversarial networks (GANs).[34] This definition also touches on two of the attributes noted earlier: fidelity – relating to the structure, properties and relationships in actual patient data – and privacy – they do not contain sensitive personal data.

There have also been numerous attempts to produce guidance on the creation and use of synthetic data. This includes general guidance, as well as guidance specific to the health data and research domain. The following section serves as a summary of these pre-existing resources.

Existing guidance within the UK

The focus of our research was concentrated primarily on how synthetic data is currently being approached within the context of the regulatory environment of healthcare and research settings in the UK. Through our desk research, we identified a number of sources from government entities and associated organisations that include consideration of the use of synthetic data, including best practices and responses to common questions around the use of synthetic data in health research contexts.

From these resources, we identified common themes – included in the table below – that provide an indication of the common considerations that organisations are dealing with when using synthetic data.

[33] U.S. Food & Drug Administration (n.d.), 'FDA Digital Health and Artificial Intelligence Glossary – Educational Resource: Synthetic Data'
[34] Hao, S., et al. (2024), 'Synthetic Data in AI: Challenges, Applications, and Ethical Implications'

Theme	Description
Balancing utility and privacy	Synthetic data is seen as a method to unlock research potential while respecting individuals' privacy and confidentiality obligations. Higher-fidelity synthetic data has better analytical utility but may pose higher reidentification risks, necessitating careful privacy-protection measures. Guidance from the NHS[35] and Health Research Authority[36] (HRA) reinforces the importance of minimising identifiable data use wherever possible, emphasising that privacy and ethical considerations must drive data choices rather than merely technical convenience.
Data fidelity and quality	Fidelity spans from low- to high-fidelity synthetic data, influencing the extent to which complex relationships are preserved. Poorly generated synthetic data can introduce unrealistic patterns and distributions within the data.[37] Synthetic data can achieve near real-data performance in predictive modelling while acknowledging that synthetic data may not capture every subtlety of original datasets. Evaluations of synthetic data (using metrics like Maximum Mean Discrepancy or k-anonymity) can ensure that data is both useful and non-identifiable.
Transparency and accountability	Clear documentation of synthetic data generation methods, intended uses, limitations, and associated risks, seems critical for trust.
Public trust and acceptability	Building trust with patients and the public involves communicating the concept of synthetic data clearly, emphasising that no individual-level identification is possible if data is properly anonymised. Public confidence can be strengthened by showing that synthetic data accelerates research without undermining privacy.
Legal and regulatory compliance	Determining when synthetic data counts as personal data is complex. Creation processes (akin to anonymisation) are subject to data protection laws and common law duties of confidentiality when derived from patient data. For example, the HRA further clarifies that if identifiable data usage is essential, additional legal steps (such as Section 251 support) must be considered, often prolonging project timelines.[38]
Technical capacity and knowledge gaps	Administrative Data Research (ADR) UK,[39] PHG Foundation[40] and the UK Statistics Authority (UKSA)[41] all point out that technical expertise is required to create, evaluate and manage synthetic data safely. They highlight the importance of organisational capabilities, ensuring staff understand and follow good practice (such as test vs. live environments). Better data planning and staff education can streamline compliance and reduce reliance on complex approvals.
Bias and representativeness	Synthetic data can reproduce biases from the source data, potentially affecting research fairness and generalisability. Demographic limitations in source data (such as certain racial or age groups) mean synthetic outputs may not be suitable for all contexts. Addressing representativeness and bias ensures that synthetic datasets support equitable and reliable research findings.

[35] Warren, G. (2024), 'Governance Considerations for the Use of Synthetic Data in Health and Care Research'
[36] NHS Health Research Authority (2022), 'Step 1: Scoping – What Are the Data Requirements for Your Project?'
[37] Government Digital Service (GDS) (2025), 'AI Insights: Synthetic Data (HTML)'
[38] NHS Health Research Authority (2022), ibid
[39] Calcraft, P., and Sutherland, A., (2022), 'Accelerating Public Policy Research with Easier, Safer Synthetic Data'
[40] Mitchell, C. and Redrup Hill, E. (2023), 'Are synthetic health data 'personal data'? - PHG Foundation'
[41] UK Statistics Authority (2022), 'Ethical Considerations Relating to the Creation and Use of Synthetic Data'

As can be seen from the themes, a combination of technical and governance issues are readily covered by existing guidance on using synthetic data in the UK. This speaks to the relatively established understanding of some of the issues that can arise, such as the trade-off between privacy and utility, which is also a known consideration when using other methods, such as differential privacy.[42]

There is currently less clarity within existing guidance in areas such as legal and regulatory compliance and issues relating to the publication of, and communication around, synthetic datasets. The latter were frequently discussed among our workshop participants, as will be covered later in this report.

From our desk research into existing guidance on synthetic data, we found that many practical aspects appear to have been sufficiently documented. However, how to account for the possible downstream consequences of creating, publishing and using synthetic data is less well documented.

Guidance from outside the UK

Similar guidance on synthetic data can be found in other jurisdictions outside the UK. Some of the most developed guidance on the use of synthetic data can be found from Singapore's Personal Data Protection Commission (PDPC) and Agency for Science, Technology and Research (A*STAR).

These organisations' recent guidance[43] is perhaps the most comprehensive and authoritative. In addition to providing an overview of what synthetic data is and where it can be useful, additional resources are included, such as real-world case studies and a suggested walkthrough for organisations contemplating generating synthetic data. This resource includes a thorough account of the steps involved in generating synthetic data and relevant considerations for organisations to bear in mind. That said, it is worth noting that the suggested instructions, as with many other guides, place emphasis on more practical matters, such as managing reidentification risks.

While it is encouraging that the guidance recommends that organisations take a proactive approach to considering these possible consequences, the suggested considerations are parsed primarily in privacy/utility trade-off terms. For example, in 'Step 1: Know Your Data' of the PDPC guidance, these include:

[42] Nanayakkara, P., et al. (2022), 'Visualizing Privacy-Utility Trade-Offs in Differentially Private Data Releases'
[43] Personal Data Protection Commission, and Agency for Science, Technology and Research (2024), 'Proposed Guide on Synthetic Data Generation'

- Where general trends/insights of source data are sensitive, organisations should take note that the use of synthetic data will not offer any protection to the trends/insights since they will be replicated in the synthetic data.[44]
- Where the synthetic data is intended to be released publicly, organisations may have to prioritise data protection over data utility in such circumstances.
- Where relevant, organisations should also put in place proper contractual obligations on recipients of synthetic data where necessary to prevent reidentification attacks on the data.

While these practical suggestions are a good starting point for organisations looking to generate synthetic datasets, they could be elaborated upon, to encourage organisations to identify and mitigate potential consequences rather than make a decision to prioritise data protection over utility. It should, however, be noted that this guidance provides a thorough general basis for organisations as a starting point to explore the possibility of generating synthetic data. It should be considered as a useful reference document on which to add further contextual considerations as necessary.

The United Nations Economic Commission for Europe (UNECE)[45] offers similarly comprehensive guidance, but this is intended primarily for those working within national statistical offices and is therefore intended for prospective purposes such as the release of census data. On this basis, the guidance is tailored quite specifically with emphasis placed on the challenges involved in large, public disclosures of data.

Finally, there are ongoing international efforts within standards bodies looking to increase understanding and use of synthetic data. TheInternational Organization for Standardization's (ISO) work on ISO/IEC AWI TR 42103 will provide an overview of synthetic data in the context of AI systems.[46] Similarly, the Institute of Electrical and Electronics Engineers (IEEE) Standards Association has an ongoing activity group looking into dimensions of synthetic data and plans to create an AI-generated synthetic data privacy and accuracy standard proposal.[47] However, both of these efforts are currently within the early stages of development and it is unclear when these standards will be published.

[44] While the PDPC asserts that synthetic data will not provide any protection due to replication, this can be overcome if the sensitive trend is identified in advance. That said, addressing privacy concerns of data sets can often involve the lowering of the fidelity and, sometimes, utility of the data.
[45] United Nations Economic Commission for Europe (2022), 'Synthetic Data for Official Statistics: A Starter Guide'
[46] ISO (nd), 'ISO/IEC AWI TR 42103'
[47] IEEE Standards Association (2025), 'Synthetic Data'

Ethical approaches to synthetic data in health research settings

As we have seen, there is both significant interest in the use of synthetic data and practical guidance that can help organisations that are considering generating synthetic data to overcome data sharing and access challenges. We have also observed that much of the guidance and current emphasis placed on the use of synthetic data revolves primarily around what is possible with synthetic data, rather than what is appropriate.

In addition to the previously mentioned risks associated with reidentification of underlying subjects that may be benign in nature, it is possible that individuals or organisations may knowingly seek to generate biased data, or create synthetic data, given the ease with which it can now be created, at the expense of efforts to collect traditional data.[48] These considerations should be scrutinised by applying a data ethics lens to data interventions.[49]

In the following sections, we turn to the findings and analysis of the primary data collected through the course of interviews with subject matter experts, and the discussions that took place throughout the course of the data ethics canvas workshop, held in March 2025.

[48] Whitney, C. D. and Norman, J. (2024), 'Real Risks of Fake Data: Synthetic Data, Diversity-Washing and Consent Circumvention'
[49] ODI (n.d.), 'Data Ethics Canvas'

Findings

Our findings relate to three themes we identified through our research:

- **Theme 1**: the broad scope of potential use cases for synthetic data in healthcare settings
- **Theme 2**: the varied challenges with using synthetic data
- **Theme 3**: the status of existing frameworks and guidance for those using synthetic data

Theme 1: Define use cases

A number of proposed use cases surfaced during our research. These use cases are diverse in regards to:

- Applicability to the healthcare setting
- Targeted users
- Likelihood and promise of benefits
- The required fidelity of the synthetic data (see Fidelity discussion)

Use case	Example	Fidelity[50]
Simulation and prediction research	Synthetic data that is close to an external benchmark can be used to assess the impact of policy options[51]	High
Hypothesis, methods, and algorithm testing	Broad agreement that this is a useful and less ethically challenging use case – using synthetic data to test edge cases and identify potential biases in tools and models before implementation/launch. Synthetic data can be used in pilot studies to test hypothesis, code and models[52]	Low
Validation of models	Validate ready-to-implement models before real data is used	High
Generating a synthetic control arm (SCA)	Synthetic data can be used to generate synthetic control groups for experiments[53]	High

[50] the suggested level of fidelity required is generally indicative, based on the review of literature and participants' comments. In operation, the fidelity required for each use case is likely to vary.
[51] Davis, P., Lay-Yee, R. and Pearson, J. (2010), 'Using micro-simulation to create a synthesised data set and test policy options: The case of health service effects under demographic ageing'
[52] Arora, A. et al. (2025), 'The urgent need to accelerate synthetic data privacy frameworks for medical research'
[53] Adams, T. et al. (2024), 'On the Trade-Off between Fidelity, Utility and Privacy of Synthetic Patient Data'

Data augmentation and diversifying research datasets	Synthetic data can fill gaps in original datasets without requiring additional real-world collection	High
Epidemiological study/Public health research	Simulation studies on synthetic populations - for example, to understand the potential spread of disease and the impact of emergency measures[54]	Low
Education and training	Synthetic data can be used to train students in a realistic but safe environment[55]	Medium
Public release of datasets	Synthetic data can be released as a proxy for real data, when sharing the real data is inappropriate[56]	Low

Accessibility of research data	Researchers can demonstrate the potential of their research using synthetic data and use it as part of their justification for accessing real datasets	Low
Linking data	Synthetic data can be used to evaluate and test the linkage quality of an algorithm to link records before applying it to real-world administrative data[57]	Low
Global collaboration	Synthetic data can plug gaps in global collaborations where there are inconsistencies with data access and quality from partners across different geographies, technologies, regulatory environments[58]	Low

[54] Enanoria, W.T.A. et al. (2016), 'The Effect of Contact Investigations and Public Health Interventions in the Control and Prevention of Measles Transmission: A Simulation Study'
[55] Laderas, T. et al. (2017), 'Teaching data science fundamentals through realistic synthetic clinical cardiovascular data'
[56] Kokosi, T. and Harron, K. (2022), 'Synthetic data in medical research'; Artificial data pilot (2025) NHS England Digital
[57] Harron, K. et al. (2016), 'Linking Data for Mothers and Babies in De-Identified Electronic Health Data'
[58] From interviews

Theme 2: Challenges

Through our research, we have identified a number of challenges relating to the creation and use of synthetic data. These challenges apply to all uses, regardless of sector, but may become more critical in a health research setting. We have put these challenges into three categories: technical, ethical and governance.

	Challenge	Details
Technical	Validation	There are challenges with validating synthetic data, as there is currently no benchmark or standard for validating synthetic data[59]
	Fidelity	There is a trade-off between fidelity, utility and privacy with any use of synthetic data. Often the higher fidelity data is more useful for research and testing purposes, but has increased costs and privacy risks connected to it.[60] It is worth differentiating between fidelity and utility, as a dataset may still prove useful for the training of a model, even when the fidelity of the dataset has been reduced. This can potentially contribute towards increasing data privacy
	Accuracy	While studies have shown that synthetic data can provide close estimates to real data results, accuracy can vary depending on the size of the dataset and the complexity of the analysis[61]
	Evaluation	The evaluation of synthetic data quality evaluation is crucial once synthetic data is used in real world applications to ensure that sensitive information is protected and machine learning models are performing as expected[6263]
	Overfitting	Particularly for the use of synthetic in generative models, there is a risk that models memorise the artificial patterns from synthetic training samples that are less applicable to real, more complex data
	Model collapse	Seddik et al. show that training exclusively on synthetic data can lead to model collapse, though this can potentially be mitigated by incorporating real data alongside synthetic data.[64] Meanwhile, Feng et al. suggest using feedback on synthesised data as a preventive measure against collapse[65]

[59] Gonzales, A., Guruswamy, G. and Smith, S.R. (2023), 'Synthetic data in health care: A narrative review'
[60] Adams, T. et al. (2024), 'On the Trade-Off between Fidelity, Utility and Privacy of Synthetic Patient Data'
[61] Benaim, A.R. et al. (2020), 'Analyzing Medical Research Results Based on Synthetic Data and Their Relation to Real Data Results: Systematic Comparison From Five Observational Studies'; Murtaza, H. et al. (2023), 'Synthetic data generation: State of the art in health care domain'
[62] Livieris, I.E. et al. (2024), 'An evaluation framework for synthetic data generation models'
[63] Arnold, C. and Neunhoeffer, M. (2021), 'Really Useful Synthetic Data -- A Framework to Evaluate the Quality of Differentially Private Synthetic Data'
[64] This is in the context of developing Large Language Models.
[65] Feng, Y. et al. (2024), 'Beyond Model Collapse: Scaling Up with Synthesized Data Requires Verification'

	Data quality	Good quality synthetic data relies upon the real data being high quality. Issues regarding diversity of data and a lack of appropriate linked data can hold back more realistic synthetic data being generated
Ethical	Human decision making	Azizi et al. asks whether synthetic data would ever be used for decision making or whether final analyses will always need to be conducted on the original data[66]
	Disclosing data	The generation of synthetic data must be carefully managed to ensure that it does not inadvertently reveal sensitive information about real patients[67]. Assessing privacy risks in synthetic datasets remains a significant challenge, with limited practical examples of how disclosures might occur in real-world scenarios
	Connection to real data	As synthetic health data, while not deidentified, can closely replicate real data, there are questions about whether it should be classified as protected health information and require informed consent and research ethics review[68]
	Informed consent	There are ethical considerations surrounding the generation of synthetic data based on real patient information without explicit permission. This also pertains to the original purpose for which the data was collected, which remains contentious when considered within the use of other privacy-enhancing technologies where raw data is not shared
Governance	Documentation	There is a lack of documentation on how to create and use synthetic data safely and ethically, as this is still a nascent area. Experiments are being conducted without much guidance. There is literature, but little that provides substantive structures
	Governance expertise	Compliance teams cannot be expected to maintain the same level of understanding as technical experts. Translation needs to take place to ensure that different challenges and mitigation strategies are explained appropriately and effective governance decisions can be made

[66] Azizi, Z., et al. (2021), 'Can Synthetic Data Be a Proxy for Real Clinical Trial Data? A Validation Study'
[67] Murtaza, H. et al. (2023), 'Synthetic data generation: State of the art in health care domain'
[68] Tsao, S.-F. et al. (2023), 'Health Synthetic Data to Enable Health Learning System and Innovation: A Scoping Review'

Theme 3: Ethical and legal frameworks

Researchers rely on guidance from the Information Commissioner's Office (ICO), Data Protection Impact Assessments (DPIAs), and the General Data Protection Regulation (GDPR) when exploring synthetic data. Each of these has different levels of relevance to synthetic data. For example, GDPR:

- includes no specific mention of synthetic data
- only provides guidance on lawfulness, transparency, and specific purposes
- lacks clear definitions for anonymisation and pseudonymisation,[69] making it difficult to establish concrete privacy risk thresholds for synthetic data[70]

Researchers use DPIAs to request approval for research from compliance teams. While they provide ample coverage for synthetic data projects, in their current state they may be insufficient to deal with the impacts in the long term. The ICO is exploring the topics of synthetic data[71] and PETs[72], though interviewees noted that questions often remain regarding the clarity that can be provided when synthetic data is sufficiently anonymised. This uncertainty becomes particularly problematic when dealing with edge cases that fall between established categories.

As the field of research for synthetic data is immature, there are not yet any standardised approaches or guidance for what processes to follow – for example, documentation for when to use one privacy preserving technique over another, or what level of fidelity to use. This means we don't yet have consensus from both governance and data science perspectives on privacy, methods, and practical aspects of synthetic data creation and use. It is even less clear for 'edge cases'. In response to these challenges, practitioners have expressed a need for specific 'Terms of Use' guidelines providing clearer direction on appropriate synthetic data applications, helping to fill the governance gap while more comprehensive frameworks develop.

The following section discusses how these shortfalls can be addressed.

[69] Stalla-Bourdillon, S. and Burt, A., (2023), 'The definition of 'anonymization' is changing in the EU: Here's what that means'
[70] Redrup Hill, E. and Mitchell, C. (2024), 'GDPR Brief: when are synthetic health data personal data?'
[71] Information Commissioner's Office (2023), 'Research Paper: Exploring Synthetic Data Validation – Privacy, Utility and Fidelity'
[72] Information Commissioner's Office (2023), 'Privacy-enhancing technologies (PETs)'

Discussion

The engagement sections of our project have exposed a number of ongoing challenges relating to synthetic data that warrant further consideration. We have also included recommendations for each challenge identified in the discussion.

Fidelity

When generating synthetic patient data, there is a risk/reward balance regarding fidelity. Low-fidelity approaches may lack sufficient connection to the patterns present in real medical data. This can undermine their utility for training algorithms or conducting meaningful research. Furthermore, the use of synthetic data with lower fidelity is often perceived as lower quality within certain research circles.[73] As noted previously, this can sometimes be a misconception. However, users of lower fidelity synthetic data may have to provide greater evidence of the utility of the data to demonstrate that this has not been lost. Conversely, high-fidelity approaches capture more authentic clinical relationships but demand sophisticated modelling techniques, domain expertise, and substantial computational resources, and run a greater risk of reidentification.

Striking the optimal balance requires careful consideration of a three-dimensional trade-off between fidelity (how closely the synthetic data mirrors real-world medical complexities), utility (how effectively the data serves its intended analytical or training purpose) and privacy (the degree to which the data protects individual patients from reidentification risks).[74]

One potential solution could be the generation of Risk-Utility scores, which could help guide decision-making,[75] although the generation of these scores requires further research. While there are already a number of methods of evaluating the privacy and utility of synthetic health-related data, there is not yet consensus on a standardised approach, nor on what method is optimal in which scenario.[76] The researchers responsible for this paper have cautioned against 'trusting synthetic data in high-risk scenarios where false positives, missed findings, or privacy breaches could cause harm'.[77] This suggests that further efforts are required to understand the most suitable methods for measuring the utility-fidelity-privacy trade-off in specific circumstances. This research could be undertaken by academia or other research institutes and would benefit from close cooperation with healthcare companies and providers.

[73] From interviews
[74] Adams, T. et al. (2024), 'On the Trade-Off between Fidelity, Utility and Privacy of Synthetic Patient Data'
[75] Taub, J. et al. (2019), 'Creating the Best Risk-Utility Profile: The Synthetic Data Challenge'
[76] Kaabachi, B. et al. (2025), 'A scoping review of privacy and utility metrics in medical synthetic data'
[77] Ibid.

In parallel, researchers have found that creating high-fidelity data for specific use cases often proved more time-consuming than simply obtaining access to the original data. Hence, there is a significant use case for low-fidelity approaches that enable people to decide whether a project is worth pursuing, and supports the justification for higher-fidelity or real data being used.

Recommendation one: Academia and industry should explore technical and non-technical mechanisms to evaluate and understand the fidelity trade-off within specific health-related scenarios

Decision-making

Our research participants broadly agreed that synthetic data should not be used for direct decision-making purposes. Due to the present issues regarding missing data and bias, and validating and evaluating synthetic data, they considered it inappropriate to use synthetic data in situations that could lead to harmful clinical outcomes. They suggested that any decision-making model should be validated on real data and through human-in-the-loop validation for synthetic data usage.[78] Further development of benchmarks to validate a synthetic dataset's likeness to the original data source can also guide those looking to use synthetic data as part of an integrated approach. On this basis, companies looking to use synthetic development in their research and product development cycles should consider taking proactive efforts to be transparent in their use of synthetic data. This could be through consistent labelling and documentation in instances where synthetic data has been used. This could be carried out on an industry or sector basis, which would require collaboration between actors within the sector.

Consistent labelling of synthetic datasets is necessary, due to the risk that synthetic data could be incidentally used in decision-making. For example, when research or models have been augmented, practitioners may unknowingly rely on synthetic data-enhanced models/datasets without understanding the underlying limitations. The ONS's approach of developing a carefully considered definition of synthetic data can help to maintain a consistent and cautious approach to labelling and using synthetic data in a format that is machine-readable and resilient to deliberate/accidental deletion. Those generating and using synthetic data within health research settings should

[78] Beck, E. A., Tavalire, H. and Searcy, J. (2023), 'Frontline Data Science: Lessons Learned From a Pandemic'

consider adopting the Croissant metadata vocabulary for machine-learning datasets[79] when labelling their synthetic dataset, to ensure that this has been sufficiently and responsibly documented.

Recommendation two: Regulators, such as the ICO and MHRA, should ensure the health sector has clear guidance for instances where synthetic data might be used in decision-making

Recommendation three: Organisations – including companies and research institutes – developing and using synthetic data should strengthen approaches to ensure synthetic data has a clear definition and is clearly labelled in a consistent and machine-readable format

Regulators, ethical guidance, and the provision of data

There are outstanding questions about where the guidance for using synthetic data should originate. Our research participants broadly agreed that:

- The need for firmer guidance is particularly acute in the healthcare setting. There is an expectation that the ICO would perform at least some of this role to set the ground rules for when synthetic data can be used appropriately and to establish the guardrails for how it can be used safely, and that the NHS would provide guidance more specific to the healthcare setting.
- The ICO should also consider issuing clarifications regarding legal obligations, and consider how GDPR applies to synthetic data.

Some of this work should be focused on raising awareness of the possibilities and limitations of synthetic data. It is not hard to imagine that as attention in synthetic data rises, it becomes a field occupied not only by experts who have a strong grasp on the ethics of using a novel technology, and is instead misused, or that mistakes are made.

Researchers have noted how the positive attention that synthetic data is garnering might attract 'bad actors'.[80] As with any emerging technology, there are further risks that limited public understanding can impact the perceptions of the technology, and might affect the risk appetite of organisations to engage

[79] MLCommons (2024), 'Croissant RAI Specification'
[80] Arora, A. *et al.* (2025), 'The urgent need to accelerate synthetic data privacy frameworks for medical research'

with it. Appropriate guidance and public engagement can act jointly in building trust in the technology and the actors looking to utilise it.

To further promote safe and effective uses of synthetic data within UK healthcare, the NHS should play a role in providing both suitable data for synthesising and open access synthetic data that can be used for health research. This work began in 2022, when NHS Skunkworks experimented with creating a synthetic version of Medical Information Mart for Intensive Care III (MIMIC-III) data,[81] as well as three further artificial datasets available from Hospital Episode Statistics (HES).[82] NHS Digital has also openly shared a codebase that can be used to generate synthetic data.[83] The artificial data generator creates random synthetic records using anonymised and aggregated data. Since statistical relationships between columns are not maintained, individual synthetic records do not accurately represent actual individual records from the original dataset.

Our research identified differing ideas on the mechanisms that should be put in place to identify and manage ethical challenges with using synthetic data. For example:

- a tiered system or ethical gating process could support researchers and practitioners. The tiers could relate to different thresholds regarding the size, quality, fidelity or benchmarks that could guide the regulatory decision-making
- A standardised consent model could accelerate approval for synthetic data projects
- A simplified version of a DPIA relevant to synthetic data could minimise confusion relating to synthetic data projects

In pursuing mechanisms such as those outlined above, actors within the private sector could proactively demonstrate responsible approaches to the use and management of synthetic data, which could serve as the basis for future guidance and regulation within the health sector and beyond.

Recommendation four: Regulators, the NHS and researchers – both from industry and academia – should collaborate on a sector-wide ethical framework that can help data scientists and health researchers to create, experiment with, and use synthetic data safely and responsibly

[81] NHS Transformation Directorate (2022), 'Exploring how to create mock patient data (synthetic data) from real patient data'
[82] NHS England Digital (2025), 'Artificial data pilot'
[83] NHS Digital (nd), 'artificial-data-generator'

Patients, consent, and reputational risk

There are outstanding questions regarding the extent to which patients need to be informed about the use of synthetic data. There might also be a struggle to obtain patient consent for data about them to be synthesised.

There are obvious reputational risks if something should actually go wrong with the use of synthetic data – for example, real patient data is exposed, or a synthetic dataset is used in rogue decision-making. Any widespread fear or concern over the use of synthetic data risks tarnishing responsible synthetic data use. As previously explored in ODI research,[84] this runs the risk of inflicting incalculable damage to user confidence in the use and motivations for using this type of technology, thus eroding the social-licence basis for its use.

Additionally, there is also a reputational risk involved in any experimental use of data within a health setting. For example, the NHS's care.data controversy[85] lives large in the memories of people working in this environment. That case demonstrates how narratives can escape the realities of the proposed use of data/technology if the public engagement aspects are not managed correctly. Likewise, present concerns around access to Biobank data could complicate matters regarding patient consent.[86] Controversies such as these can therefore serve to undermine broader confidence, among those who have chosen to participate in these initiatives as well as the wider public, thus potentially jeopardising efforts towards greater data sharing and collaboration.

One consideration, particularly for any organisation looking to create a large-scale synthetic dataset (for example, if exploring a full 'synthetic patient' or patient pathway), is to create a patient consultation group, to enable researchers to understand patient concerns and responses to experimental synthetic data uses. Through this group, ethical frameworks and pilot use cases could also be co-developed in a manner that builds patient trust and understanding in the technology. These insights could be scaled and form part of an open explanation about synthetic data usage in health research. The ODI's research on participatory data providers further guidance on the benefits of engaging with those impacted by data usage.

Finally, a recurrent theme across the groups during the data ethics canvas workshop was the need for greater public engagement work. Amid national efforts to improve use and transparency in the secondary processing of healthcare data, further research is needed into the public acceptability of

[84] ODI (2025), 'Trust and transparency in privacy-enhancing technologies'
[85] Carter, P., Laurie, G.T. and Dixon-Woods, M. (2015), 'The social licence for research: why care.data ran into trouble'
[86] Pegg, D., Devlin, H. and Burgis, T. (2024), 'Concerns raised over access to UK Biobank data after "race scientists" claims'

synthetic data. There is a dearth of literature exploring this.[87] As a possible example to draw upon, Research Data Scotland (RDS) has been engaging with its 'Scotland Talks Data' public panel as part of a project to make synthetic datasets available for research.

Recommendation five: The Department for Health and Social Care should consider exploring patient engagement processes to build trust and understanding of synthetic data

Broader ethical concerns

Researchers had concerns regarding the use of synthetic data to reduce bias, or augment datasets with more diversity. For instance, the publicly-available skin image datasets used to train machine-learning algorithms for skin cancer diagnosis significantly under-represent darker skin types.

Researchers have begun to experiment with developing synthetic images of darker-skin melanomas using deep-learning approaches.[88] There are, however, a number of concerns and risks associated with these efforts:

- There are implicit assumptions that darker skin images can be created from lighter skin images
- It does not appear that the underlying issues of dataset diversity are being remedied, for example through increased efforts to create under-represented samples
- Models might be created with the synthetic datasets without sufficient human or medical expert validation

Further, there is a risk that synthetic data becomes entangled with ideas for 'quick fixes' for wider issues within data quality and AI development. Synthetic data, due to its artificial nature, is considered cleaner and free from issues that plague real datasets regarding quality, ethics, privacy, diversity and size. Experts have however noted that synthetic data is just as 'deeply entangled with societal injustices as "real" data, despite claims it can be used to do the opposite'.[89]

Nevertheless, reducing bias as far as possible – by establishing or repurposing benchmarks from other technical areas[90] – can support data scientists to

[87] Arora, A. et al. (2025), 'The urgent need to accelerate synthetic data privacy frameworks for medical research'
[88] Rezk, E., Eltorki, M. and El-Dakhakhni, W. (2022), 'Improving Skin Color Diversity in Cancer Detection: Deep Learning Approach'
[89] Lee, F., Hajisharif, S. and Johnson, E. (2025), 'The ontological politics of synthetic data: Normalities, outliers, and intersectional hallucinations'
[90] Zhang, J. et al. (2024) 'VLBiasBench: A Comprehensive Benchmark for Evaluating Bias in Large Vision-Language Model'

evaluate that their synthetic datasets are 'good enough'. As with measuring privacy and utility metrics of synthetic data, there is not yet definitive guidance on the most suitable evaluation methods for specific instances. This could benefit from further research and the development of actionable resources for developers and users of synthetic data.

Further, efforts should be strengthened to diversify original datasets to address systemic under-representation in medical research.[91] Interviewees raised this as a necessary counterpoint to mitigate the prospect of a growing dependency on synthetic data materialising, which is deprioritising efforts to tackle structural issues that underlie the present need and use of synthetic data.

Recommendation six: Organisations should ensure ethical frameworks on the use of synthetic data adequately cover bias and understand systemic issues in health datasets

Future developments

Participants noted a number of upcoming developments that warrant future analysis:

- A research group in China used synthetic data to recreate the RETFound model,[92] which was created by Moorfields Eye Hospital to identify eye diseases - with a fraction of the real data required[93]
- The UK Health Security Agency (UKHSA) is developing a framework for synthetic data to streamline the privacy agreement process and enable multiple copies of synthetic data to be generated without requiring multiple different DPIAs
- Two major European Union initiatives, SEARCH[94] and SYNTHIA[95], are developing multimodal synthetic data for healthcare with significant funding

[91] Boden-Albala, B. (2022) 'Confronting legacies of underrepresentation in clinical trials: The case for greater diversity in research'; Daitch, V. *et al.* (2022) 'Underrepresentation of women in randomized controlled trials: a systematic review and meta-analysis'.
[92] Zhou, Y. et al. (2023), 'A foundation model for generalizable disease detection from retinal images'
[93] Sun, Y. et al. (2025), 'A data-efficient strategy for building high-performing medical foundation models'
[94] Synthetic Healthcare Data Governance Hub (SEARCH) (2024), 'SEARCH'
[95] Synthetic Data Generation framework for integrated validation of use cases and AI healthcare applications (Synthia) (2025), 'Synthia'

Conclusion

There is ample evidence of the potential for synthetic data to be very useful in health research. As has already been demonstrated, it can expedite access to sample sets that can speed up innovation, and also help to develop more robust and generalisable ML/AI models. However, uncertainty remains at the regulatory level, where the tendency is understandably to err on the side of caution. Regulators would do well to consider how organisations can be better equipped to use synthetic data with confidence.

From an ethics perspective, organisations must consider the potential downstream effects of using synthetic data on those who stand to be affected. Concerns remain around unintentional disclosure and the perpetuation of bias, which can be attributed to improper care being taken in the generation of synthetic data.

Furthermore, as evidenced from our interviews, labelling synthetic data is also a risk when considering whether to publish a synthetic dataset that an organisation has created. The potential for misrepresentation of a synthetic dataset has prevented UK institutions from publishing synthetic datasets in the past. This remains an issue and could limit the possible innovative benefits around synthetic data.

At present, the advisable solution appears to be that synthetic datasets based on sensitive data should be subject to additional safeguards, such as being accessible within a TRE, where greater control can be maintained.

Recommendations

Through the course of this research, a number of recommendations have emerged that we believe could be worth pursuing to promote the ethical use of synthetic data in health research.

Recommendation one: Academia and industry should explore technical and non-technical mechanisms to evaluate and understand the fidelity trade-off within specific health-related scenarios

Recommendation two: Regulators, such as the ICO and MHRA, should ensure the health sector has clear guidance for instances where synthetic data might be used in decision-making

Recommendation three: Organisations – including companies and research institutes – developing and using synthetic data should strengthen approaches to ensure synthetic data has a clear definition and is clearly labelled in a consistent and machine-readable format

Recommendation four: Regulators, the NHS and researchers – both from industry and academia – should collaborate on a sector-wide ethical framework that can help data scientists and health researchers to create, experiment with, and use synthetic data safely and responsibly

Recommendation five: The Department for Health and Social Care should consider exploring patient engagement processes to build trust and understanding of synthetic data

Recommendation six: Organisations should ensure ethical frameworks on the use of synthetic data adequately cover bias and understand systemic issues in health datasets

Furthermore, we believe that the above recommendations can be complemented and strengthened when combined by building on ongoing efforts within the sector. For example, Data and Analytics Research Environments UK (DARE UK) has recently established a UK Synthetic Data Community Group.[96] Based on discussions that took place during the data ethics canvas workshop, we believe that it could be useful to explore the possibility of developing a healthcare research-specific working group, either within the Community Group, or to convene alongside it in a complementary fashion. This could ensure that those who are testing and exploring the use of synthetic data in this specific domain are conscious of the ongoing efforts. Such a group could also begin to establish consensus and work towards the development of a framework for the ethical creation and use of synthetic data in healthcare research.

Another initiative that could provide greater confidence to users of synthetic data in healthcare research settings is the establishment of an NHS synthetic data sandbox. This could provide interested innovators with the opportunity to create datasets in a secure environment, while drawing on the experience of regulators. This could also be beneficial for regulators, as it could help to create clearer definitions around suitably anonymised synthetic data.

[96] DARE UK (2025), 'UK Synthetic Data Community Group'

Acknowledgements

We are also incredibly grateful for the time and input provided by the many experts who contributed to the research activities that were undertaken throughout the course of this research. This includes workshop participants, who provided valuable input. These include, but are not limited to:

Sahar Abdulrahman, GSK
Paolo Buscemi, Genomics England
Henry Li, Tony Blair Institute for Global Change
Ed Morrell, GSK
Puja Myles
Ella Shoup, GSK
John Smith, Department of Science, Innovation and Technology (DSIT)
Cassie Smith, Health Data Research UK
Anna Steere
Markus Trengove, GSK
Robert Vandersluis, GSK