Synthetic Data

1. Overview **2**
1.1 Executive Summary 2
1.2 Summary of key proposals 5
2. Context **6**
2.1 The opportunity **6**
2.2 The challenge 7
3. Synthetic Data: Introduction **8**
3.1 What is synthetic data? 8
3.2 Types of synthetic data 9
4. Publicly-sourced synthetic data **9**
4.1 What is publicly-sourced synthetic data? 9
4.3 Applications and benefits of publicly-sourced synthetic data 11
5. Privately-sourced synthetic data **14**
5.1 What is privately-sourced synthetic data? 14
5.2 How privately-sourced synthetic data is generated 14
5.3 Applications and benefits of privately-sourced synthetic data 17
6. Regulatory and legal guidance on synthetic data **21**
7. Proposed high-level two-tier approach **22**
8. Publicly-sourced synthetic data: Proposed principles, policies and procedures **24**
8.1 Overview 24
8.2 Principles 24
8.3 Policies and procedures 24
9. Privately-sourced synthetic data: Proposed principles, policies and procedures **25**
9.1 Overview 25
9.2 Scope 26
9.4 Roles and responsibilities 27
9.5 Policies and procedures 28
9.6 Phased implementation 33
10. Conclusion **33**
11. Annexes **34**
Annex 1: Survey of synthetic data providers 34
Annex 2: Differential privacy 36
Annex 3: Evaluation methods for privately-sourced synthetic datasets 37
Annex 4: UK Data Service: Access Levels and Conditions 38
Annex 5: Example UCLH script for patient information video about synthetic data 39
Annex 6: Potential use cases for phase 1 rollout of privately-sourced synthetic data 40
12. Glossary **41**
13. References, Interviews, PPIs **44**

1. Overview

1.1 Executive Summary

The rapid digitisation of medical data and the development of powerful computational tools hold significant promise for scientific discovery and innovation. They also carry considerable risk for patient privacy. Access to healthcare data is therefore restricted: best practice currently is to anonymise or de-identify patient data and store it within a Trusted Research Environment (TRE) – a highly secure computing environment that provides access to accredited and approved researchers, with rigorous approval procedures and strict governance of ingress and egress.[1] Anonymisation is technically challenging, and requires deep expertise, especially where a combination of idiosyncratic values can make it possible to identify a patient, even when Personally Identifiable Information (PII) is removed. Meanwhile, the controls inherent in the TRE and the surrounding governance, whilst crucial enablers of research, slow data access and restrict collaborative working in an era of unprecedented demand for local and global healthcare data sharing.[2]

Synthetic data offers a compelling "third way" to improve access to health data whilst safeguarding patient privacy. The anonymised data approach is to collect real data from events, perform anonymisation on this data; the output being real patient records with key identifying information removed. By contrast, synthetic data is created artificially by studying patterns in real patient records and then generating new records that follow these patterns. Synthetic data outputs are entirely new records; they resemble real records but, with a suitable process, will have little or no traceable relation to the original underlying data.

In this paper, we set out three broad types of synthetic data: **structural synthetic data**, **publicly-sourced synthetic data** and **privately-sourced synthetic data**. These terms are deliberately descriptive, intended to clarify the source of the synthetic data, which has implications for the associated privacy risk and appropriate governance approach.[3]

Structural synthetic data uses the structure of a real dataset, without retaining any of its statistical information. Whilst there may be intellectual property (IP) and licensing implications for releasing the structure of a database, there is no learning from the underlying data itself. Therefore there is no disclosure risk (the risk of inappropriate attribution of personal information

[1] HDR UK, "What is a TRE?". Trusted Research Environments are also known as Data Safe Havens. TREs may hold identifiable data depending on the governance and the task, in which case they involve greater levels of access control and more stringent governance. An example of a TRE is the CHIMERA database, which holds anonymised UCLH and GOSH acute care data to be used by authorised individuals for research on human physiology co-sponsored by UCLH and UCL. Data in a TRE is sometimes identifiable.

[2] COVID-19 has created demand for an unprecedented level of data sharing with researchers, healthcare providers, and public health organisations: El Emam, K., Mosquera, L., Jonker, E., and Sood, E., "Evaluating the utility of synthetic COVID-19 case data", *JAMIA Open*, 2021, Vol. 00, No. 0; doi: 10.1093/jamiaopen/ooab012

[3] In practice, the nomenclature for different levels of synthetic data is variable and inconsistent; different levels of synthetic data are often given the same name by different organisations, and levels are often collapsed together. To avoid confusion, we will use the descriptive terms, "structural", "publicly-sourced" and "privately-sourced" throughout this document, with the understanding that different organisations will apply their own naming conventions to the different levels of synthetic data described.

to an individual without their approval) precluding the synthetic data's open release. This level of synthetic data has the lowest fidelity (the closeness with which the synthetic data resembles the underlying data), and, correspondingly, the lowest utility for analysis and decision-making; however, structural synthetic data can be useful for researchers to understand the relations and formats of datasets when developing and testing analysis pipelines. For the purposes of this paper, however, we will focus predominantly on the two types of synthetic data which have greater fidelity and utility: **publicly-sourced synthetic data** and **privately-sourced synthetic data**.

Publicly-sourced synthetic data takes not only the structure of a real dataset; it also uses publicly available statistical information. This statistical information is either already in the public domain or has been approved for public release using a standard disclosure control procedure for publishing health statistics. The disclosure risk of publicly-sourced synthetic data is commensurate with the risk carried by information openly published for clinical governance and public health reporting – i.e., provided an appropriate statistical disclosure method is followed, the risk is zero. Publicly-sourced synthetic data is used for tasks where the structure and type of data–as well as some understanding of content–is important; for example in the development and testing of code prior to accessing real data or in the evaluation of study feasibility.[4]

Privately-sourced synthetic data is synthetic data that has been generated from a real, private dataset in order to represent specific aspects of the underlying dataset. A synthetic data generator accesses and learns the statistical properties of the real dataset and uses these to produce artificial values. The direct exposure of the generator to real, private patient records means that there is a risk of disclosure. Nonetheless, the generation of privately-sourced synthetic data can be made quantifiably safer than anonymisation, and privacy controls can therefore be adjusted accordingly, with confidence. Privately-sourced synthetic data is used for tasks where some clinical or biological inference is required, for example in research hypothesis development, clinical trials, and in some limited cases, the training and validation of medical algorithms.

The practical applications of synthetic healthcare data – structural, publicly-sourced and privately-sourced – are powerful. Synthetic data expedites research and innovation by enabling researchers to conduct preparatory work on models and code while the real patient data goes through critical but time-consuming safeguards and approvals. Because it is more readily available for disclosure, it can be used to support scientific peer review; to train and validate algorithms; to educate cohorts of healthcare students in handling big data; to facilitate cross-organisational sharing of analytical tools and outputs; and to share information with the public on health issues and improvements in care.

[4] NHS England defines artificial data as follows: "Artificial data sets provide users with large volumes of data that share some of the characteristics of real data while protecting patient confidentiality. They are designed to model the structure of real data but are completely artificial – they do not contain any actual patient records."

The development and release of synthetic datasets is a way to show researchers, higher education institutions, NHS organisations, students, and the public what patient datasets broadly contain, without giving them access to real data. As such, it represents significant benefits for research, education, cross-NHS collaboration, and public health. Patient focus groups have responded positively to the idea of using synthetic data as a way to identify promising research areas and facilitate exploratory research, whilst ensuring that individual privacy is protected.[5]

While there is increasing energy behind publishing publicly-sourced synthetic datasets (for example via NHS Digital's Artificial Data Pilot),[6] there are no clear NHS-wide guidelines for the generation, use and release of different types of synthetic healthcare data – with their varying disclosure risks – to support research, collaboration, and education. Although synthetic data is broadly defined by the Information Commissioner's Office (ICO) as "non-personal",[7] this is an evolving technological, regulatory and legal area. Recent regulatory guidance recommends a risk-based approach to the processing of synthetic data, based on the level of risk that an individual can be identified from a dataset.[8]

In keeping with this regulatory guidance, this paper sets out a simple, risk-based framework for the classification and release of synthetic data products, based on the nature of the underlying source data, their relationship with this data, and their attendant privacy risk – i.e., based on whether they are structural, publicly-sourced or privately-sourced synthetic data. This approach enables the harnessing of the value of synthetic data whilst applying proportionate controls to protect patient privacy. Specifically, this paper proposes that:

- **Structural synthetic data** can always be publicly shared, provided that it is clearly labelled as such and that there are no commercial or IP concerns.

- **Publicly-sourced synthetic data** can be publicly shared, provided that:
    - The statistical information used to generate it is either already publicly available or has been approved for public release via the organisation's standard procedure for publishing statistics;
    - The generation mechanism (i.e., algorithm) has been approved by Information Governance, via the sign-off of a Data Protection Impact Assessment (DPIA);[9]
    - There are no commercial or IP concerns.

---

[5] UCLH BRC-led PPI event on use of data in research, March 2021; UCLH BRC-led PPI event on synthetic data, January 2024

[6] NHS Digital is currently piloting a public service for the release of artificial datasets ("Artificial Data Pilot"). GOSH's Data Partnerships Committee (DPC) has approved and implemented a" Dummy Data Policy and SoP" which enables researchers to use low-fidelity synthetic datasets before they have R&D approval to access real anonymised data through TREs

[7] The ICO's current definition of synthetic data broadly considers it as anonymous or non-personal, and therefore out of scope of the UK GDPR or the common law duty of confidentiality: "[t]o the extent that synthetic data cannot be related to identified or identifiable living individuals, it is not personal data and therefore data protection obligations do not apply when you process it."

[8] In its draft guidance on privacy enhancing technologies, the ICO clarifies: "You should consider whether the synthetic data you generate is personal data. You should focus on the extent to which individuals are identified or identifiable in the synthetic data, and what information about them would be revealed if identification is successful." (ICO, *Draft anonymisation, pseudonymisation and privacy enhancing technologies guidance*, Chapter 5: "Privacy-enhancing technologies (PETs)", September 2022

[9] Model operated by GOSH

- **Privately-sourced synthetic data** can be released with moderate user access and governance controls. These controls are lighter-touch than those applied to anonymised and de-identified data held in a TRE, in keeping with the lower privacy risk. Privately-sourced synthetic datasets should be released with appropriate documentation regarding their generation and evaluation (e.g., metrics of similarity to the original dataset), alongside any relevant licensing stipulations.

1.2 Summary of key proposals

Exhibit 1 below summarises this paper's key proposals (detailed further in Sections 6 and 7).

Exhibit 1: Summary of key proposals

Synthetic data level	Description	Access to underlying real data	Privacy risk	Example uses	Key proposed policies & procedures
Structural synthetic data	Mimics only the structure of the real data	None - only exposure is to structure of dataset	Zero	Developing and testing analysis pipelines	• Can always be publicly shared, provided no commercial/ IP concerns • Must be labelled as structural synthetic data
Publicly-sourced synthetic data	Mimics the structure of real data and incorporates aggregated statistical information that is either already in the public domain or has been approved for public release using a standard disclosure control procedure	None - only exposure is to the structure of the dataset and aggregated statistical information about the data that has been approved for release into the public domain	Zero, provided standard statistical disclosure process is followed	Developing and testing code prior to accessing real data	• Can always be publicly released, provided that the mechanism of its generation has been approved by IG and a standard disclosure control method has been applied to the statistical information used (and provided no commercial/IP concerns) • The scope of internal use cases/applications must be defined and documented (for example, it cannot be used for direct clinical care purposes) • Must be clearly labelled as publicly-sourced synthetic data and accompanied by documentation explaining its generation mechanism • A record is kept of the uses of the synthetic data
Privately-sourced synthetic data	Mimics the structure of real data and incorporates private statistical information about the real underlying health data	Mediated/ modelled - the generation tool has access to the real data from which it learns currently private	Low	Generating research hypotheses or training and validating	• Can be released for specific use cases with moderate privacy controls (user access controls and clear digital chain of custody) • Generation and evaluation should be overseen by a qualified Synthetic Data Audit Team

		statistical information about the data.		medical algorithms	• Release must be approved by IG • Must be released with appropriate documentation regarding generation and evaluation and any relevant licensing stipulations

Each type of synthetic data has different privacy disclosure risks (the risk of inappropriate attribution of personal information to an individual without their approval) and different applications.[10]

2. Context

2.1 The opportunity

Explosion of digital healthcare data and technologies

The past decade has seen a global explosion in digital healthcare data and technologies, accelerated by COVID-19. Population-scale Electronic Patient Records (EPRs) are rapidly becoming the standard in developed health systems, and are being progressively rolled out across the NHS.[11] Meanwhile, advances in Artificial Intelligence (AI) and Machine Learning (ML) are transforming our capacity to realise the value of big health data and to evaluate and tackle healthcare problems at the population scale. In his recent report on the state of the NHS in England, Lord Darzi casts technology as a key driver of performance, urging the NHS to take a "major tilt towards technology to unlock productivity": "the extraordinary richness of NHS datasets is largely untapped either in clinical care, service planning, or research".[12]

UCLP is well positioned to spur data-driven improvements to healthcare

UCLPartners (UCLP) comprises some of the first NHS trusts to transform into fully integrated digital hospitals (e.g., UCLH and GOSH, via the implementation of the Epic EPR). It is also at the forefront of global healthcare research and innovation, via its cross-partner collaborations, the Academic Health Science Centre, and its partnership with key research and innovation organisations.[13] As such, UCLP is well-positioned to "deliver timely and impactful health research

[10] From Johns Hopkins: Disclosure is the term typically used to refer to *inappropriate* attribution of information to an individual or organisation without their approval. (Any such approval usually takes the form of the terms within the *consent forms* that participants sign.). Levels of disclosure risk are: identity disclosure (subject can be directly identified -e.g., an individual is part of a trial); attribute disclosure (reveals sensitive information about a subject, e.g., HIV status); and inferential disclosure(released data makes it easier to determine a characteristic of a subject, through links to external information, such as the internet, or because of outlier variables or unusual combinations of identifiers that narrow down to certain individuals).

[11] As of January 31, 2024, 189 NHS trusts in England have electronic patient records (EPRs).

[12] The Rt Hon. Professor the Lord Darzi of Denham, *Independent Investigation of the National Health Service in England*, September 2024; Introduction and p.103

[13] The UCLP Academic Health Science Centre (AHSC) brings together five hospitals and three universities (including four NIHR Biomedical Research Centres) to drive a pipeline of innovative research, to tackle the most pressing health challenges facing society. UCLP's Innovation and Research partners include the Health Innovation Network, NIHR North Thames, North Thames Genomic Medicine Service, NHSE Health Innovation Network, Care City and Med City.

and its translation to offer more effective treatments, track and prevent public health risks, utilising health data to improve and save lives".[14]

Public support for the use of unidentifiable patient data to improve healthcare
Meanwhile, patients and the public are broadly in favour of the use of patient data to improve healthcare, provided that individual privacy is protected. In summarising the findings of a number of studies on the UK public's attitudes to patient data, the research and advocacy organisation, Understanding Patient Data, stated that "people are generally comfortable with anonymised data from medical records being used for improving health, care, and services, provided there is a public benefit."[15] This has been borne out in recent Patient and Public Involvement and Engagement (PPIE) events held by UCLP members. A PPIE event on the use of data in research in 2021 found that ~80% of patients felt it was imperative that the NHS used data to improve healthcare, with some saying it would be "immoral" if it did not, and the majority being happy to consent to unidentifiable data being used.[16]

Emerging public understanding of and support for use of synthetic data
More specifically, initial engagement with the public on the manufacture and release of synthetic data has been positive. At a recent UCLH BRC-led PPIE event focused on synthetic data, public and patient representatives generally endorsed the release of synthetic data, regarding it as a powerful and efficient way to share information for research and education whilst providing better protection of patient confidentiality compared to anonymised data.[17]

The vision: to share healthcare data with the broadest possible research community
Beyond UCLP and the NHS, there is significant utility in sharing health data and collaborating more broadly; scientific discovery and innovation rely upon the collaboration of researchers and the sharing of data.[18] The ultimate vision is a global research community in which "data controllers and individuals should be able to share the largest possible data sets within the broadest possible community, while effectively protecting the identities of individual subjects and, in turn, be able to recapture value from data transactions, fostering network effects to progressively increase the total volume of available data".[19]

2.2 The challenge

New risks to patient privacy

---

[14] *From privacy to partnership: The role of privacy enhancing technologies in data governance and collaborative analysis*, The Royal Society (January 2023); hereafter, *From privacy to partnership*, The Royal Society (2023)

[15] Understanding Patient Data, collation of studies from 2010-2021; link here

[16] UCLH BRC PPI, Co-production of data: *Summary findings of our work with patients and public looking at the use of data in research*, March 2021

[17] UCLH BRC PPI on Use of Synthetic Data, 17 January 2024

[18] *From privacy to partnership*, The Royal Society (2023)

[19] Human Perspectives in Health Sciences and Technology Series, *Personalized Medicine in the Making: Philosophical Perspectives from Biology to Healthcare*, Ed. Chiara Beneduce & Marta Bertolaso: "New Solutions to Biomedical Data Sharing: Secure Computation and Synthetic Data", Edwin Morley-Fletcher; hereafter, Morley-Fletcher

Whilst the boom in digital health data offers significant opportunities for scientific research and innovation, it also creates new and significant risks to patient privacy. As the Royal Society states in its paper, *From Privacy to Partnership*:

> The ever-growing quantity of data collected in contemporary life, coupled with increasing power to compute, is opening new possibilities for data-driven solutions. At the same time, there is unprecedented potential for the misuse of data – whether intentional or unintentional – leading to downstream harms at individual, community, corporate and national scales.[20]

Indeed, the major anxiety that patients and the public have in relation to the secondary use of healthcare data is the extent to which individuals might be identified via the process of analysing data and publishing research.[21]

"Five Safes" to mitigate privacy risks…
To mitigate the risks of privacy breaches and misuse of data, best practice within the UK and elsewhere is to follow the "Five Safes" framework in processing healthcare data: safe data; safe people; safe projects; safe settings; and safe outputs.[22]

.., but privacy protections slow research and impede collaboration
Whilst providing crucial safeguarding of patient confidentiality, the security measures inherent in the Five Safes framework pose significant challenges for researchers. Research application and approval processes ("safe projects") are often lengthy and complicated. The process of anonymising or de-identifying data ("safe data") is expensive, complex, and slow; it is also not without privacy risk, especially for multidimensional data, where a combination of different data fields (e.g., demographic, laboratory measurements, observational data) can make it possible to identify a patient, even when PII (e.g., name, date of birth, date of visit) are removed. The controls in the Trusted Research Environment ("safe settings"), where all activity is audited and monitored, are highly restrictive, inevitably restrict collaborative working; and the governance of outputs ("safe outputs") is resource-intensive.[23] The overall result is that the sharing of healthcare data is limited and characterised by high transaction costs. As Morley-Fletcher notes, " 'although available data is continuously expanding, it largely sits idle' (Finck, 2019), i.e. fragmented in siloes carefully guarded by data controllers to reduce legal exposure".[24] This approach is antithetical to collaborative and reproducible science: locking away data inhibits the validation of the quality and outputs of scientific research. It also hampers efforts to work in

[20] *From privacy to partnership*, The Royal Society (2023)
[21] UCLH BRC PPI, Co-production of data: *Summary findings of our work with patients and the public looking at the use of data in research*, March 2021
[22] The "Five Safes": "Safe People": only trained and specifically accredited researchers can access the data; "Safe Projects": data are only used for ethical, approved research with the potential for clear public benefit; " Safe Settings": access to data is only possible using secure technology systems - the data never leaves the Trusted Research Environment; "Safe Data": researchers only use data that have been de-identified to protect privacy; "Safe Outputs": all research outputs are checked to ensure that they cannot be used to identify subjects. Source: HDR UK
[23] Kokosi, K., De Stavola, B, Mitra, R., Harron, K., at el, "Using synthetic administrative data for research", September 2021
[24] Morley-Fletcher

partnership across organisations; specifically, it prevents NHS Trusts from sharing data with researchers, educational partners and each other in a safe and fast way.

3. Synthetic Data: Introduction

3.1 What is synthetic data?

A third way to improve access to data while protecting privacy
Synthetic data – fabricated data that is manufactured to mimic real health data – offers a compelling "third way" to improve access to health data whilst ensuring the robust protection of patient privacy. Synthetic data is artificially generated data designed to mimic real data, for a specific purpose. It does not contain data directly collected from real patients and hence does not contain personally identifiable information (PII). Synthetic data is a risk minimisation technique to protect privacy whilst maximising the utility and use of data. As Giuffrè and Shung argue: "Synthetic data offers an attractive alternative that addresses privacy concerns, streamlines data utility agreements, protocol submissions, and ethics review approvals, and decreases costs".[25]

3.2 Types of synthetic data

Three broad types of synthetic data , carrying different levels of disclosure risk
We can categorise synthetic data into three broad types, based upon its source: **structural synthetic data**, **publicly-sourced synthetic data** and **privately-sourced synthetic data**. In practice, the nomenclature for different types of synthetic data is variable and inconsistent across organisations; by using purely descriptive terms, we aim to avoid naming confusion and to clarify our proposed governance approach.[26] Each type of synthetic data has different methods of generation, privacy disclosure risks and applications.

Structural synthetic data uses the structure of a real dataset (for example, ensuring valid characters for names, numbers and dates), without retaining any statistical information about the real data. Essentially, this means simply taking the structure of a relational database – its tables, columns and data types, and generating random data in accordance to this structure. There is no direct access to or learning from the underlying data itself, and there is therefore no disclosure risk.

This paper focuses on the two types of synthetic data which have greater fidelity (closeness with which it resembles the underlying data) and utility (usefulness), and which therefore offer compelling healthcare applications: **publicly-sourced synthetic data** and **privately-sourced**

[25] Giuffrè, M., Shung, D.L. *Harnessing the power of synthetic data in healthcare: innovation, application, and privacy*. *npj Digit. Med.* 6, 186 (2023). https://doi.org/10.1038/s41746-023-00927-3

[26] For example, what we define here as "structural synthetic data", is also known as "mock data" or "dummy data" (Data Science Campus). What we define as "publicly-sourced synthetic data" is also known as "dummy data", "artificial data" (NHS England term), "cohort-constrained dummy data", and "cohort-defined dummy data", and in some cases "low-fidelity synthetic data". What we define as "privately-sourced synthetic data" is also known as "synthetic data" and "high-fidelity synthetic data".

synthetic data. Sections 4 and 5 below set out in detail the generation, applications and privacy risks of publicly-sourced and privately-sourced synthetic data.

4. Publicly-sourced synthetic data

4.1 What is publicly-sourced synthetic data?

Publicly sourced and therefore publicly shareable
Publicly-sourced synthetic data (called "artificial data" by NHS England) takes the structure of a real dataset and applies publicly-available statistical information from aggregated, anonymised real data . This statistical information is either already in the public domain (e.g., distribution of patient ages), or has been approved for public release using a standard disclosure control procedure for publishing health information. There is no link back to individual patient records; any attempt to reverse-engineer the publicly-sourced synthetic data would only yield the same aggregate statistics on which the publicly-sourced synthetic data is based. Since publicly-sourced synthetic data is entirely generated from dataset structures and statistical information approved for public release, with no direct access to individual personal data, the disclosure risk it carries is the same as that carried by aggregated information that may be openly published for clinical governance, public health, and freedom of information reporting – i.e., zero. As such it is not considered "personal data" and does not require safeguarding under GDPR.

4.2 How publicly-sourced synthetic data is generated

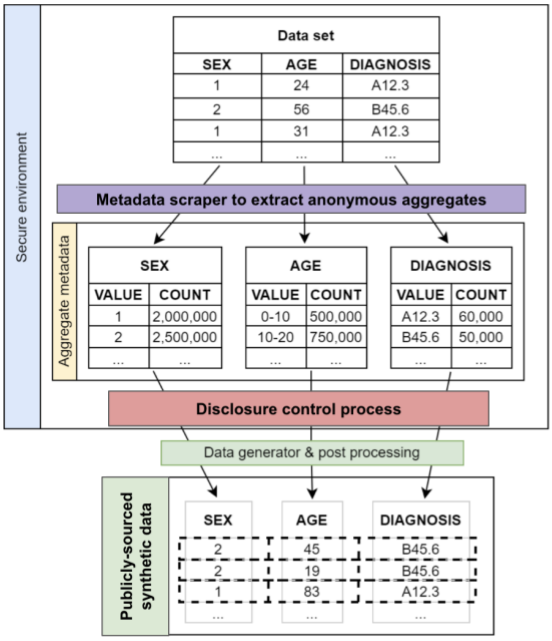
Publicly-sourced synthetic data is derived from aggregated metadata
There are three steps required to generate publicly-sourced synthetic data:

1. A "metadata scraper" extracts anonymous aggregate statistics from the real data
2. These anonymous aggregates are reviewed and signed off to be publicly released, in keeping with standard procedures for publishing statistics used across the organisation..
3. A "data generator" randomly samples from the anonymous aggregates to generate the synthetic data.

Often the data generator will directly use anonymous aggregate statistics that are already in the public domain.

Exhibit 2 below sets out a simplified, high-level example of how publicly-sourced synthetic data is developed, summarised from the approach used by NHS England.

Exhibit 2: Approach for developing publicly-sourced synthetic data



- Begin with the original anonymised patient dataset
- Within a secure environment, a "metadata scraper" is used - i.e., a tool is used to extract data *about* the data
- Specifically, the scraper extracts aggregate statistical information from the anonymised dataset, for example giving the range and distribution of values for a set of variables (e.g., age, diagnosis)
- Following a disclosure control process to ensure that the mechanism used ensures the data is unidentifiable (i.e., to the same degree as any publicly disclosed health data), the dataset structure and anonymous aggregates are released
- A synthetic data generator then uses the database structure and the anonymous aggregates to randomly create artificial records, creating an entirely new dataset
- For example, each artificial record in a dataset containing records describing genders, ages, ICD 10 diagnoses, appointment dates and times, would be generate by randomly picking a value for each field and then putting them together to form a complete record. Only by coincidence would some synthetic records resemble real records.
- Some additional structural logic may be added to the synthetic dataset (e.g., ward stay start and end dates calculated to fit within hospital admission start and end dates)

Diagram adapted from NHS Artificial Data Pilot

NHS Digital has published a more detailed description of the generation of publicly-sourced synthetic data as well as a codebase to generate publicly-sourced synthetic data (termed "artificial data" by NHS Digital. Open source tools and commercial applications are also available to support the generation of publicly-sourced synthetic datasets (see Annex 1).

4.3 Applications and benefits of publicly-sourced synthetic data

Useful for tasks where the structure and type of data are important but no clinical inference is required
Publicly-sourced synthetic data is useful for tasks where the structure and type of data is important but no clinical or biological inference is required – often in the testing of data processing pipelines, but also in high-level exploration of data resources.[27] Exhibit 3 sets out the key applications and benefits of publicly-sourced synthetic data.

[27] Data Science Campus, *Synthesising the linked 2011 Census and deaths dataset while preserving its confidentiality*, November 20, 2023

Exhibit 3: Key applications and benefits of publicly-sourced synthetic data

Application	Description	Benefits	Case studies
Research (development of models, code and tools)	Researchers can use publicly-sourced synthetic data to begin work on preparatory models, code and tools while access to the real data goes through critical but time-saving safeguards (such as anonymisation) and approvals	• Saves time • Enables collaboration as the same analysis or engineering can be used at multiple sites • Preserves privacy, as the amount of time that researchers need to access sensitive patient information is reduced[28]	*COVID Vaccine Rollout* • In 2020, a team at the Alan Turing Institute was assigned to help determine which groups should be prioritised in the rollout of the vaccine. The provision of COVID outcome data from the UCLH Secure Data Environment (SDE) would take time, as anonymisation and approvals were required. Through the provision of publicly-sourced synthetic data, the Alan Turing Institute could have been able to structure their tools and models–and check their code with international reviewers–before receiving access to the real data. This would have been crucial in expediting the research and ensuring that the vaccine was rolled out to the most vulnerable groups first.
External collaboration	Publicly-sourced synthetic data enables users to collaborate and share code across different environments where real data is not accessible (e.g., allows the development of algorithms which can be deployed across	• Enables code and algorithms to be leveraged across networks (both national and international) without the sharing of real data • Enables broader collaboration on code	*GOSH federated analytics* • Publicly-sourced synthetic data will support the development of inter-hospital analytics via GOSH's federated analytics tooling infrastructure, without the need for data to move outside of hospital Trusts.[29]

[28] Kokosi, T., and Harron, K. "Synthetic data in medical research", The BMJ (2022)

[29] GOSH DRIVE/DRE, *Position regarding Synthetic Health Data use at GOSH* (Dr William Bryant, Prof. Neil Sebire, Dr Natassa Spiridou), v 1.1., 22/06/2023

	federated analytics networks)	with zero risk to patient privacy	
Internal collaboration	Publicly-sourced synthetic data enables internal colleagues access to data at scale, to enable learning and development	● Enables clinicians to share patient data at scale with each other, without concerns about individual patient privacy	
Education and training	Publicly-sourced synthetic datasets can be used in the teaching and training of clinicians, data engineers and scientists	● Enables students to develop skills in handling big datasets, (e.g., auditing and analysis), as well as in data representation, with zero risk to patient privacy	*Digital AMR Hub* ● Publicly-sourced synthetic data will support the sharing of knowledge and skills across researchers, healthcare workers, policymakers, charities, industry and the public to improve surveillance of and action against antimicrobial resistance.[30]
Population-level insights	Publicly-sourced synthetic datasets can be used to explore and compare populations at a high level	● Enables high-level population insights based on aggregates of variables, with no risk to patient privacy	*Turing Synthetic Population Catalyst (SPC)* ● SPC combines aggregate data from a variety of sources–including UK census data and health surveys–to generate synthetic population data that can be easily accessed by researchers. Researchers at the Technical University of Denmark have used synthetic populations generated by the tool to find out why obese people are more susceptible to COVID-19 and other viral diseases.[31]

[30] https://www.ucl.ac.uk/news/2023/jun/new-ps4m-digital-hub-tackle-antimicrobial-resistance

[31] https://www.turing.ac.uk/about-us/impact/powering-population-models-synthetic-data

5. Privately-sourced synthetic data

5.1 What is privately-sourced synthetic data?

Directly learns statistical relationships between variables from real data; very low quantifiable disclosure risk

Privately-sourced synthetic data is generated from a real, private dataset, in order to specifically represent aspects of that dataset. It is generated by computational tools that directly access real datasets, learn their statistical properties, optionally add quantifiable noise to these statistical properties, and use these properties to generate new artificial values. A privately-sourced synthetic dataset is sufficiently similar to the underlying dataset that even complex analysis would result in similar findings across the real and synthetic datasets.

The direct exposure of the generator to real, private patient records means that there is a risk of disclosure. For example, privacy leaks can occur where the synthetic data includes very detailed, rare types of patient profiles, of which there are only a small number in the real patient data; they could also occur during membership inference attacks (whereby an attacker analyses model outputs and makes an educated guess about whether a specific data point was part of the training dataset) and model inversion attacks (whereby the data used to train a model is reconstructed).[32] Nonetheless, the generation of privately-sourced synthetic data can be made quantifiable safer than anonymisation, and privacy controls can therefore be adjusted accordingly, with confidence.

In general, there is a trade-off between utility, i.e., how realistic synthetic data is in preserving the statistical relationships of the underlying real data, and therefore how useful it is to researchers; and privacy, i.e., the risk that information about real patients could be inferred from the synthetic dataset.[33]

5.2 How privately-sourced synthetic data is generated

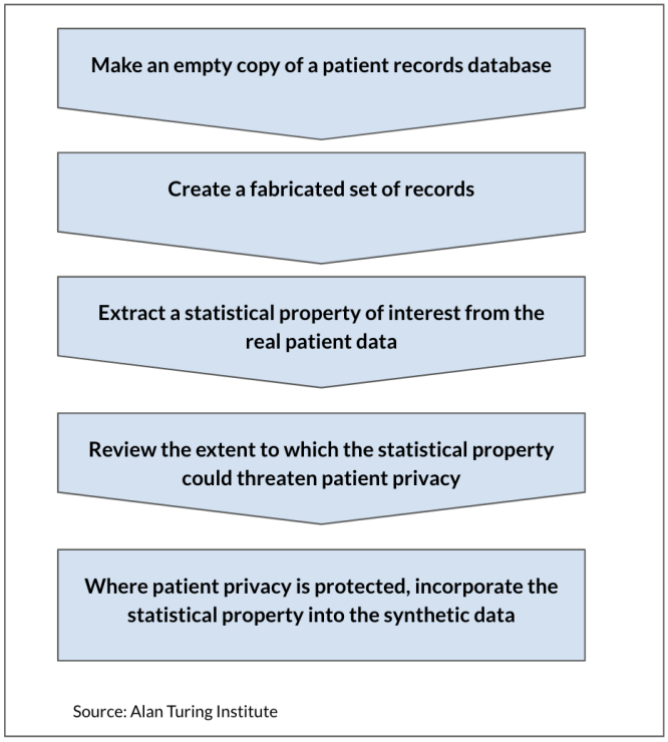
Begin with nonsense data, and introduce statistical relationships until it looks like real data
Privately-sourced synthetic data can be generated by a variety of approaches. One approach is to capture the structure of real-life health data records, populate this structure with fabricated data, and introduce learned statistical relationships into the fabricated data. Whereas anonymised data is collected from real-life events and people, the approach for generating privately-sourced synthetic data is to begin with data that looks nothing like the real data, and as

[32] "Are synthetic health data 'personal data?'", PHG Foundation (2023)

[33] As the ICO notes in its *Draft anonymisation, pseudonymisation and privacy enhancing technologies guidance*, "The degree to which synthetic data is a proxy for the original data depends on the utility of the method and model. The more that the synthetic data mimics real data, the greater the utility it has. At the same time, it may be more likely to reveal individuals' personal data." (Chapter 5)

different statistical relationships and properties are added, it begins to look more and more like the real data.
Exhibit 4 sets out a simplified, high-level approach for developing privately-sourced synthetic data.

Exhibit 4: Approach for developing privately-sourced synthetic data



A range of manufacturing models and tools and no clear consensus on evaluation
Synthetic data is produced by a purpose-built model or algorithm; in this way it contrasts with real data, which is generated not by a model but by real-world events (e.g., patient admissions).[34] The generation of privately-sourced synthetic data is complex, requiring a range of computational tools that directly access real datasets, model their statistical properties, and use the outputs to generate new datasets.

The growing interest in the medical use of synthetic data has led to the development of open source tools (e.g., Synthea) and commercial applications (e.g., MDClone's Synthetic Data Engine), which take varying approaches to modelling synthetic data (see Annex 1).[35] Several quantitative measures can be used to evaluate the outputs of these methods, in terms of quality (how good the simulacrum is) and safety (how much privacy risk there is; as described below, a basket of approaches is likely required to ensure a broad and thorough evaluation (see Section 9.5). [36]

[34] The Alan Turing Institute and The Royal Society, *Synthetic Data - what, why and how?* (2024)
[35] Giuffrè, M., Shung, D.L. *Harnessing the power of synthetic data in healthcare: innovation, application, and privacy*. *npj Digit. Med.* 6, 186 (2023). https://doi.org/10.1038/s41746-023-00927-3
[36] GOSH DRIVE/DRE, *Position regarding Synthetic Health Data use at GOSH* (Dr William Bryant, Prof. Neil Sebire, Dr Natassa Spiridou), v 1.1., 22/06/2023

5.3 Privacy of privately-sourced synthetic data

Safer than anonymised data: quantifiable privacy loss enables access controls to commensurate
While privately-sourced synthetic data carries disclosure risk, a careful generation process can quantify the privacy loss, and help finely define access policy. By comparison, in anonymised data, combinations of idiosyncratic values can make it possible to identify a patient even when PII is removed.[37] And while anonymised datasets are derived directly from information about individual patients, the process that generates privately-sourced synthetic data at most has access to statistical patterns in the real data, not individual patient records. As NHS England states, "Various anonymisation techniques can be used to turn data into a form that does not directly identify individuals and where re-identification is not likely to take place. However, it is very difficult to entirely remove the chance of re-identification. Wide release of anonymised data will always carry some risks. Synthetic data aims to remove the need for such concerns because there is no "real patient" connected with the data."[38]

It is worth noting that by sheer statistical accident, the privately-sourced synthetic dataset could include a data-point that to exactly mimics a "real patient" across a certain number of data fields (e.g., a randomly created patient that has the same name, date of birth, assigned sex and gender as a real patient). This combination of data has not been learned from the underlying dataset, but is a statistical coincidence produced by one of many millions of random statistical calculations; it is, therefore, not a privacy breach.

Exhibit 5 summarises the key differences between anonymised and privately-sourced synthetic data, in terms of method of generation and privacy risk.

[37] In the recent paper, The urgent need to accelerate synthetic data privacy frameworks for medical research, Arora et al argue that: "As a result of progress in artificial intelligence (AI) research, the concept of anonymised data is beginning to be challenged, as we are increasingly able to extract personal information from data that were previously considered anonymous. Information about age, sex, and ethnicity can be ascertained from electrocardiograms and retinal photographs.3, 4 In medical imaging, it has been established that deep learning models predict race from medical images even after perturbation.5 Furthermore, person-level identification from electroencephalograms has even been possible.6 This possibility for identification has, therefore, created a need for privacy-preserving methods, such as federated learning and synthetic data, to protect against both adversarial and accidental re-identification of patient data"

[38] NHS England Transformation Directorate, *Exploring how to create mock patient data (synthetic data) from real patient data*, March 2022

Exhibit 5 – Key Differences between Anonymised and Privately-Sourced Synthetic Data

	Anonymised data	Privately-sourced synthetic data
Method of generation	● Starts with real data ● Strips out fields and details that can identify a patient (e.g., name, DOB) ● Includes data about real patients	● Starts with artificial data ● Introduces statistical relationships learned from the real data in a controlled and mediated way ● Never includes data about individual patients, only statistical properties derived from the population
Privacy risk	● Difficult to control risk with multidimensional data ○ Combination of dimensions (e.g., demographic, observational, lab data) can make it possible to identify a patient even when PII is removed	● Easier to control risk with multidimensional data ○ Addition of measurable noise to statistical information to quantify privacy loss (and introduce privacy controls accordingly)

Additional privacy-enhancing techniques

Privately-sourced synthetic data therefore generally carries lower privacy risk than anonymised data, provided that there is sufficient and skilled oversight of the privately-sourced synthetic data generation process ("human-in-the-loop"), to assess that only approved statistical properties from sizeable populations are extracted. Standard statistical disclosure controls (e.g., low number suppression), and privacy enhancing technologies such as differential privacy (a mathematical framework for quantifying privacy loss, in order to protect individuals) can be used to further reduce the risk that an individual can be re-identified from a synthetic dataset (although its use in relational datasets is limited).[39] Annex 2 contains further information on differential privacy as a robust and promising approach for protecting individual privacy in the generation of privately-sourced synthetic data.

5.3 Applications and benefits of privately-sourced synthetic data

Useful for tasks where some clinical or biological inference is required

[39] Differential privacy—an algorithm that injects random data into a dataset to protect individual privacy—can protect outlier records from linkage attacks with other data, and has been described as "the gold standard with which to protect information from malicious agents" (Rosenblatt et al., "Differentially Private Synthetic Data: Applied Evaluations and Enhancements" (2020), quoting Dwork et al.,TMAC 2008). It can quantify privacy loss "as well as unique properties such as robustness to auxiliary information, composability enabling modular design, and group privacy" (ICO, *Draft anonymisation, pseudonymisation, and privacy enhancing technologies guidance*, September 2022). See also: Stadler, T., Oprisanu, B., and Troncoso, C., "Synthetic Data - Anonymisation Groundhog Day", 31st USENIX Security Symposium (USENIX Security 22), 2022 https://doi.org/10.48550/arXiv.2011.07018

Because privately-sourced synthetic data carries quantifiable privacy risks (in contrast to anonymised data), it can be shared more readily, with finer-tuned access controls and lighter-touch governance mechanisms. Its use is therefore increasingly seen as a powerful solution to the difficulty of maintaining data privacy whilst exploiting big health datasets. The UK Government Statistical Service has called it "an unprecedented opportunity to innovate with data while safeguarding privacy and fostering public trust"; while the Stanford Technology Law Review has stated that "synthetic data is a viable, next-step solution to the database-privacy problem".[40]

Given that privately-sourced synthetic data is machine-generated, it can be produced at a relatively low cost and for a variety of uses in large volumes—this removes the burden of collection for data controllers and minimises intrusion upon data subjects.[41] It can also be particularly useful in situations where it may be difficult or unethical to collect personal data.[42] It can be used as a placeholder until anonymised data is available, or as a substitute for anonymised or source data.[43] It offers new opportunities for the sharing of datasets between organisations, which has the potential to enhance population-level analysis and operational planning.[44] With care, it can be shared in selective research and educational settings while minimising the risk of compromising "real" patient data and without the same transaction costs incurred in the use of a TRE. It can be shared with patients and the public to demonstrate key public health insights and illustrate improvements in the delivery of patient care.

Exhibit 6 sets out the key applications and benefits of privately-sourced synthetic data.

Exhibit 6: Key applications and benefits of privately-sourced synthetic data

Application	Description	Benefits	Case studies

[40] Quality Centre. 2018. *Government Statistical Service, Privacy and data confidentiality methods: A National Statistician's Quality Review (NSQR)*; Bellovin, Steven M., Preetam K. Dutta, and Nathan Reitinger. 2019. Privacy and synthetic datasets. Stanford Technology Law Review 22 (1): 2–52

[41] ENISA/European Union Agency for Cybersecurity, *Data Protection Engineering: From Theory to Practice* (January 2022); Morley-Fletcher. Note that many methods are not scalable to large volumes

[42] ENISA/European Union Agency for Cybersecurity, *Data Protection Engineering: From Theory to Practice* (January 2022): "For instance, in counterfactual analyses where the goal is to study the causal effects of a specific intervention and implementing this intervention may not be a practical option. Think of situations in which one is interested in the effect of a new treatment on a pathology, or the consequences of an exposure to a risk factor for human health. In all those circumstances it may not be possible to give the new treatment to the entire population, or it may not be ethical to suspend a prior one, and it is not ethical to deliberately expose an individual to a risk factor (e.g. pollution) to check its effects on his health status."

[43] There are some arguments for using synthetic data to correct bias in training data, but this practice is in its very early days and further evidence is required on the accuracy and utility of this method

[44] Arora, A., Wagner, S.k., Carpenter, R., Jena, R., Keane, P.A., *The urgent need to accelerate synthetic data privacy frameworks for medical research*, The Lancet Digital Health, 2024, https://doi.org/10.1016/S2589-7500(24)00196-1

Research (hypothesis development)	Researchers can use privately-sourced synthetic data to identify and test hypotheses or to undertake intensive code testing while waiting for approvals for access to real data	• Saves time • Preserves privacy, as the amount of time that researchers need to access sensitive patient information is reduced[45]	*Data Science Campus (ONS)/ Alan Turing Institute* • Partnership to explore how privately-sourced synthetic national databases can be used provisionally by researchers who wish to test systems or develop methods in initiatives designed for public benefit across a range of activities.[46]
Scientific peer review	Privately-sourced synthetic datasets can be readily shared with other researchers or third parties to facilitate reproducible methods and verify models and analysis strategies.[47]	• Facilitates scientific peer review – particularly vital at a time when approximately 20% of studies are estimated to be fraudulent[48]	*Testing healthcare ML pre-deployment* • Privately-sourced synthetic data is especially useful in addressing concerns around the reproducibility of scientific studies involving machine learning in health data science, where there is a lack of shared code and datasets
Algorithm development	Privately-sourced synthetic datasets can be used to train and validate algorithms, used for example in the development and validation of medical devices	• Reduces costs where the alternative is manual and time-consuming (e.g., in the training of medical-AI image applications) • Jump-starts AI development in areas where data is scarce, expensive, or legally constrained (e.g., biomedical sector)[49]	*MyHealthMyData (MHMD)* • Horizon 2020 EU-funded project where privately-sourced synthetic data has been successfully used to publish clinical data and MRI cardiovascular images to train machine learning tools and to validate clinical decision support applications[50] *Clinical Practice Research Datalink (CPRD)* • Co-sponsored by the MHRA and NIHR, developed synthetic datasets base on patient data from a network of GP practices across the UK, which have been used for improving algorithms and machine learning workflows

[45] Kokosi, T., and Harron, K. "Synthetic data in medical research", The BMJ (2022)

[46] Data Science Campus, Synthesising the linked 2011 Census and deaths dataset while preserving its confidentiality (2023)

[47] Kokosi, T., and Harron, K. "Synthetic data in medical research", The BMJ (2022)

[48] Smith R. – *Time to assume that health research is fraudulent until proven otherwise* – The BMJ (2021);
Grey, A., Bolland, M.J., Avenell, A., Klein, A. A. & Gunsalus, C.K. – *Check for publication integrity before misconduct*, Nature 577, 167-169 (2020)

[49] Morley-Fletcher

[50] Morley-Fletcher

Clinical trials	Synthetic datasets generated from historical trial data can be used as a control arm for clinical trials, instead of using real patients. They can also be used to define the boundaries of a trial and model and predict patients to include/exclude[51]	• Saves significant time and resources as it reduces the burden of patient recruitment • Mitigates ethical concern around only some trial participants receiving treatment being tested[52]	*Roche* • In 2017, Roche was seeking EU approval of Alecensa (alectinib) as a lung cancer treatment. The EU's conditional approval required more evidence of alectinib's effectiveness relative to the standard of care (certinib). A synthetic control arm of 67 patients was accepted as evidence, speeding up availability of Alecensa in the EU by 18 months *Pfizer Merck Merkel* • The clinical trial for Bavencio (avelumab) from Pfizer and Merck KGaA to treat Merkel cell carcinoma, which was approved in 2018, used data from Electronic Medical Records in a synthetic control arm
Evaluation healthcare policies and pathways	Privately-sourced synthetic data can be used to estimate the benefit of screening and healthcare policies, treatments and interventions	• Improves speed and quality of policy and pathway decision-making with minimal impact on patients and their privacy	*Davis et al. study* • Used micro-simulation to create a synthesised dataset and test policy options in addressing the health service effects of an ageing population.[53]
Enrichment and quality improvement of anonymised datasets	Privately-sourced synthetic data can be used to augment and enrich datasets that are limited or incomplete, or where populations are small or underrepresented	• Improves the quality and clinical usefulness of anonymised data[54] • Mitigates biases in datasets with small populations	*Das et al., COVID-19 modelling* • Synthetic datasets were used by Das et al. in addressing the challenge of data scarcity, by augmenting data volume in imaging studies during the COVID-19 pandemic. Conditional synthetic datasets were created for chest CT scans to classify COVID-19 patients from a population of normal individuals and pneumonia patients. The authors demonstrated that using synthetic datasets could improve the accuracy of the COVID-19

[51] Accenture Life Sciences/Phesi, *Faster and cheaper clinical trials: The benefit of synthetic data*

[52] Accenture Life Sciences/Phesi, *Faster and cheaper clinical trials: The benefit of synthetic data*

[53] Giuffrè, M., Shung, D.L. *Harnessing the power of synthetic data in healthcare: innovation, application, and privacy*. *npj Digit. Med.* 6, 186 (2023). https://doi.org/10.1038/s41746-023-00927-3

[54] Morley-Fletcher, citing Nowok et al. 2017; Brown 2020; El Emam et al. 2020

			detection process compared to the original datasets.[55]
Operational planning	Privately-sourced synthetic datasets can be a means to share datasets between organisations, to enhance population analysis and support operational planning.	• Enhances population analysis and supports operational planning, without sharing personally identifiable data.	*Simulacrum dataset* • Since 2018, a freely-available synthetic cancer dataset that closely resembles the real cancer data (from the National Disease Registration Service) but without any personally identifiable data, called the Simulacrum, has been developed. The Simulacrum allows researchers and clinicians to run preliminary analyses on the synthetic data, explore data structure, and develop analytical code that can be tested on synthetic data before the real data. This can support operational planning for cancer services, e.g., in the estimation of local demand for radiotherapy to enable providers to optimise the supply of services.[56]
Education and training	Synthetic data can be used in the delivery of education and training in healthcare and healthcare-adjacent disciplines (e.g., data science)	• Gives students a more authentic experience of working with data by retaining challenging features such as missing data, with minimal risk to patient privacy	*CRPD* • CRPD has generated a number of synthetic datasets that can be used for training purposes or to improve algorithms or machine learning workflows, including its Cardiovascular disease privately-sourced synthetic dataset and COVID-19 symptoms and risk factors privately-sourced synthetic dataset.[57]

There are some existing privately-sourced synthetic healthcare datasets available–for example via the Health Gym project, the MHRA CRPD datasets, Synthea and Simulacrum–however these are limited to a few adult cohorts and drawn from varying and often inapplicable health contexts (e.g., US focused, primary care focused, cancer focus–see Annex 1 for further details).[58] The

[55] Giuffrè, M., Shung, D.L. *Harnessing the power of synthetic data in healthcare: innovation, application, and privacy*. *npj Digit. Med.* 6, 186 (2023). https://doi.org/10.1038/s41746-023-00927-3

[56] Arora, A., Wagner, S.k., Carpenter, R., Jena, R., Keane, P.A., *The urgent need to accelerate synthetic data privacy frameworks for medical research*, The Lancet Digital Health, 2024, https://doi.org/10.1016/S2589-7500(24)00196-1

[57] https://www.cprd.com/synthetic-data

[58] Kuo, N.IH., Polizzotto, M.N., Finfer, S. *et al.* The Health Gym: synthetic health-related datasets for the development of reinforcement learning algorithms. *Sci Data* 9, 693 (2022). https://doi.org/10.1038/s41597-022-01784-7. CRPD high-fidelity datasets are focused on primary care, one on cardiovascular disease and one on COVID-19 symptoms and risk factors.

opportunity to develop privately-sourced synthetic datasets from rich NHS hospital datasets is significant. In the long-term, synthetic data for research and sharing with external organisations could replace the use of de-identified/anonymised data – reducing all risk of re-identification to a very low level.

6. Regulatory and legal guidance on synthetic data

An emerging regulatory and legal area

Synthetic data is an emerging regulatory and legal area. As the PHG Foundation explains in its report on the status of synthetic health data in UK protection law, "[t]he pace of technical progress is outstripping regulatory guidance and it is unclear whether, or under what conditions, synthetic health data will be considered 'personal data' governed by data protection law (the UK GDPR and EU GDPR)". Although "data authorities across the EU and UK are cautiously positive about the potential of synthetic data", "regulators and the courts are yet to grapple fully with synthetic data generation".[59]

A risk-based approach is recommended

The ICO has broadly defined synthetic data as non-personal, and therefore out of scope of the UK GDPR or the common law duty of confidentiality: "[t]o the extent that synthetic data cannot be related to identified or identifiable living individuals, it is not personal data and therefore data protection obligations do not apply when you process it."[60] However, recent draft regulatory guidance reflects the ICO's increasing sense of nuance in its appraisal of synthetic data. acknowledging that this is an active research area and that there are associated privacy risks—specifically a trade-off between utility and risks to revealing individual data, and the potential for "model inversion attacks". The ICO's proviso, "*to the extent that synthetic data cannot be related to identified or identifiable living individuals*" is key here: the extent to which synthetic data should be treated as personal data depends on the level of risk that an individual can be identified from a dataset (both at the time of processing and as technology develops over time). In its draft guidance on privacy enhancing technologies, the ICO clarifies:

> You should consider whether the synthetic data you generate is personal data. You should focus on the extent to which individuals are identified or identifiable in the synthetic data, and what information about them would be revealed if identification is successful.[61]

At the heart of this guidance is the acknowledgment that not all synthetic data is the same in terms of privacy risk, and therefore not all synthetic data should be treated the same. The risk

[59] "*Are synthetic health data "personal data"?*, PHG Foundation (2023)

[60] ICO. PHG report: "the latest approach of the CJEU and ICO provide room for determining that synthetic data could be considered anonymous or non-personal data."

[61] ICO, *Draft anonymisation, pseudonymisation and privacy enhancing technologies guidance*. Chapter 5: Privacy-enhancing technologies (PETs). September 2022

that an individual can be identified from a synthetic data set depends upon the nature of the source data and the relationship the synthetic data has with it.

7. Proposed high-level two-tier approach

A simple, risk-based approach

The ICO's draft guidance posits a risk-based approach to the processing of synthetic data. This is echoed by the PHG Foundation, which advocates a "proportionate" approach to regulating synthetic data, whereby data controllers undertake a risk assessment of the identifiability of different categories of synthetic data to classify them as non-personal or otherwise.[62] Within this regulatory framework, publicly-sourced synthetic data can be classified as "non-personal" under the ICO's original guidance, and therefore not governed by UK GDPR; whilst privately-sourced synthetic data must be treated with care.

This paper therefore proposes the following simple, risk-based policy for the generation and release of synthetic data (Exhibit 7).

Exhibit 7: High-level summary of proposed approach to synthetic data

1. **Publicly-sourced synthetic data can always be released, provided that its generation mechanism has been approved and the anonymous aggregates have undergone a standard disclosure control process**
   - The generation mechanism for publicly-sourced synthetic data must be approved following the completion of a Data Protection Impact Assessment (DPIA) and sign-off by Information Governance; this includes review and sign-off of:
     - The algorithm
     - The anonymised aggregate statistics used in the generation of the synthetic data, in accordance with standard procedures for publishing statistics used across the organisation
     - Any noise/disclosure control methods used
     - Any rule-based logic applied
   - In approving the generation mechanism, IG would be looking at its safety in data protection terms and its appropriate use cases; these use cases must be defined and documented (e.g., cannot be used for direct clinical care purposes)
   - Once the generation mechanism has been approved, there is no requirement for further approval for the ongoing manufacture, use and release of publicly-sourced synthetic data– provided that the generation mechanism remains unchanged[63]
   - Publicly-sourced synthetic datasets must be labelled as such and must be accompanied by a clear description of the method of generation

[62] "Are synthetic health data 'personal data'?", PHG Foundation (2023)

[63] This approach is used by GOSH

- The scope of internal use cases/applications must be defined and documented (e.g., cannot be used for direct clinical care purposes)
- This approach is similar to that in use at GOSH for "dummy data" which has similar privacy risk profile

2. **Privately-sourced synthetic data can be released with moderate privacy controls, in keeping with the attendant privacy risk**
   - Privately-sourced synthetic data is safeguarded via user access controls and a clear digital chain of custody
   - These controls are markedly less stringent than those required within TREs and are consistent with the controls used for datasets that carry a small privacy risk
   - The generation and evaluation of privately-sourced synthetic data should be overseen by a dedicated and qualified Data Audit Team
   - Once evaluated, privately-sourced synthetic data sets are approved for release by Information Governance by use case
   - Privately-sourced synthetic datasets must be released with appropriate documentation regarding their generation and metrics of similarity to the original dataset
   - Given that this is an emerging area of data processing, a pilot programme should be deployed before broader rollout

Approach is proportionate to privacy risk

This high-level, risk-based approach is proportionate and simple; it facilitates access to synthetic data whilst putting protections in place where there is any risk, however small, to patient privacy; it allows for clarity and transparency, to support audit and monitoring. The approach is a form of "Data Protection Engineering", which "can be perceived as part of data protection by Design and Default. It aims to support the selection, deployment and configuration of appropriate technical and organisational measures in order to satisfy specific data protection principles".[64]

Sections 8 and 9 below set out the policy approaches to publicly-sourced and privately-sourced synthetic data in greater detail.

8. Publicly-sourced synthetic data: Proposed principles, policies and procedures

8.1 Overview

This proposed approach for publicly-sourced synthetic data has been developed and rolled out at GOSH for "dummy data".[65]

[64] ENISA/European Union Agency for Cybersecurity, *Data Protection Engineering: From Theory to Practice* (January 2022)

[65] GOSH DRIVE/DRE, *Position regarding Synthetic Health Data use at GOSH* (Dr William Bryant, Prof. Neil Sebire, Dr Natassa Spiridou), v 1.1., 22/06/2023; GOSH DRIVE/DRE: *Use of 'structural synthetic data' at GOSH: Proposed algorithm for approval.* GOSH defines dummy data as data that resembles not only the structure of the real dataset but also applies rules from the aggregated, anonymised real data to ensure that the range and distribution of each data column are plausible (e.g., to ensure that the value range for patient age is within the bounds of reality). The generation of this synthetic data therefore does not require direct access to the real dataset; it merely requires access to the heading, data type, and aggregate-level, independent rules for each data column. These

8.2 Principles

Publicly-sourced synthetic data can always be publicly released, provided the mechanism for its generation has been approved. Publicly-sourced synthetic data shall therefore be treated as open data, in line with NHS Digital's support for open data, and to improve transparency in health and care.

8.3 Policies and procedures

Prior to the release of publicly-sourced synthetic data, the mechanism for its generation must be approved by the relevant Information Governance authority, for example via a Data Protection Impact Assessment (DPIA). This approval shall include the review and sign-off of:

- The algorithm used in the generation of the synthetic data.
- The anonymised aggregate statistics used in the generation of the synthetic data; where these are not already publicly available, they must be approved in accordance with standard procedures for publishing statistics used across the organisation.
- Any noise or disclosure control methods used (e.g., small number suppression, use of ranges).
- Any rule-based logic applied (e.g., making sure that end dates are the same as or later than start dates for individual events).
- A definition of appropriate internal use cases and applications (e.g., publicly-sourced synthetic data may not be used for direct clinical care purposes).

Once a mechanism for publicly-sourced synthetic data generation has been approved, there is no requirement for further governance approval for the generation, use and release of publicly-sourced synthetic data, provided that the generation mechanism remains unchanged, in terms of (a) algorithm used/methodology used; (b) anonymised aggregate statistics used. Where there is a request to include a new or additional anonymised aggregate statistic into the publicly-sourced synthetic data, Information Governance approval must be sought for the release of that statistic, in accordance with the organisation's standard procedures for publishing statistics. Where such guidelines are not in place, the recommendation is to use the ONS's method for preventing disclosure of personal information in the publication of health statistics.[66]

The publicly-sourced synthetic dataset must be clearly labelled as synthetic data at all times, and must be accompanied by documentation describing the generation methods/algorithms used to create the data.

"rules" could be already available in the public domain (e.g., distribution of patient ages) or could be derived from a real local dataset and approved for release by the relevant information governance authority.

[66] This broadly covers the following steps: (1) Determining user requirements; (2) Understanding the key characteristics of the data and outputs; (3) Assessing disclosure risk (low, medium, high - defined in terms of likelihood of an attempt to identify an individual and impact of identification; each risk category comes with recommendations on level of protection); (4) Legal and policy considerations; (5) Disclosure control methods (e.g., suppression, rounding); (6) Implementation.

9. Privately-sourced synthetic data: Proposed principles, policies and procedures

9.1 Overview

Proposed approach based on best practice and expert recommendations
Where the proposed policy approach to publicly-sourced synthetic data is simple, the practical aspects of protecting data privacy in privately-sourced synthetic datasets are more complex.

In this section, we propose a set of principles, policies and procedures for the generation and release of privately-sourced synthetic healthcare data, with a view to ensuring the utility of the synthetic data whilst maintaining the privacy of individuals within the underlying datasets. As stated above, the governance of synthetic data is an emerging regulatory area; in drafting the recommendations below we have therefore not only drawn on the latest regulatory guidance, but also on recent relevant academic papers,[67] recommendations by the Royal Society, The Alan Turing Institute and the PHG Foundation ,[68] expert interviews,[69] current best practice for the processing of private data,[70] international case studies,[71] and emerging guidelines on the application of AI in healthcare.[72] It should be noted that even in applying these principles, policies and procedures, there will always be a residual risk of re-identification; however with improvements to synthetic data generation techniques and a consistent application of standard privacy evaluation frameworks, this risk could be minimised above the acceptable threshold.[73]

9.2 Scope

Includes both complex and non-complex clinical data types
The proposed principles, policies and procedures are intended for both non-complex clinical data types–i.e., tabular, time-series, and text-based–and complex data types based on high-dimensional modalities such as images, free-text records, genomics and sensors. The oversight of the generation of complex modes of data is likely to require a distinct set of specialist skills, to ensure patient privacy is preserved.

9.3 Principles

Key principles for the generation and release of privately-sourced synthetic data

---

[67] Relevant papers include: Giuffrè and Shung (2023); Goncalves et al. (2020); Kokosi et al. (2022); See Biography.

[68] Royal Society/Alan Turing Institute (2024); PHG Foundation (2023); Royal Society (2023); ENISA/EU Agency for Cybersecurity (2022). See Biography.

[69] E.g., Dr May Young, Professor Neil Sebire. See Interviews.

[70] Hetherington et al, *Design choices for productive, secure, data-intensive research at scale in the cloud* (2019); NHS England Transformation Directorate, *Exploring how to create mock patient data (synthetic data) from real patient data*, March 2022

[71] The proposed user access and management controls for privately-sourced synthetic datasets have been informed by those used for publicly available anonymised datasets from international health records - i.e., MIMIC-IV (USA); AmsterdamUMCdb and eICU Collaborative Research Database (eICU-CRD) v2.0 (USA).

[72] https://transform.england.nhs.uk/information-governance/guidance/artificial-intelligence/

[73] GOSH DRIVE/DRE, *Position regarding Synthetic Health Data use at GOSH* (Dr William Bryant, Prof. Neil Sebire, Dr Natassa Spiridou), v 1.1., 22/06/2023

The proposed principles for governing the generation and release of privately-sourced synthetic data are intended to strike the appropriate balance between enabling access to data whilst protecting individual patient privacy. Exhibit 11 below sets out the proposed principles.

Exhibit 11: Proposed principles for the generation and release of privately-sourced synthetic data

Human-in-the-loop: The generation of synthetic data shall be overseen by a technically skilled and clinically experienced team, to ensure that clinical expertise and situational understanding is incorporated into the synthetic data, thereby improving privacy and utility (e.g., ensuring plausible scenarios, mitigating biases, ensuring any statistical mimicry is strictly necessary). In practice, this may require the establishment of a dedicated expert team (see Synthetic Data Audit Team, Section 9.4 below).

Privacy as Priority: In developing privately-sourced synthetic datasets, patient privacy shall be the utmost priority, both as a value to uphold and a right to be protected. All efforts shall be made to ensure that appropriate techniques are used to ensure patient privacy (including standard statistical disclosure controls and privacy enhancing techniques such as differential privacy);[74] and a Data Protection Impact Assessment must be completed and Information Governance approval given prior to the release of the dataset.

Specific Use Cases*:* The creation and application of a privately-sourced synthetic dataset shall be tailored to and approved for a specific use case, both to reduce risks to individual privacy (by limiting the extent to which the synthetic data mimics the statistical relationships within the underlying data) and to ensure the utility of the data (by ensuring that the privately-sourced synthetic dataset is fit-for-purpose). By focusing on specific use cases, the privately-sourced synthetic data will resemble some (useful) aspects of the real data, but not others, thereby reducing disclosure risk.

Evaluation and Audit*:* The quality, utility, privacy and bias of privately-sourced synthetic data shall be evaluated prior to release, using the best techniques available.[75] The privacy of the synthetic dataset shall be audited on an ongoing basis to ensure it remains robust in the face of data changes and evolving security threats.

Proportionate Security*:* Proportionate security controls shall be put in place in the release of privately-sourced synthetic datasets; specifically, controlled user access and a clear digital chain of custody (i.e., the recording and documentation of the synthetic data through its lifecycle, encompassing data generation, storage, sharing and destruction).[76] These controls are in line with best practice for the

[74] Wording taken from Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nat Mach Intell.* 2019;1:389–399. doi: 10.1038/s42256-019-0088-2.

[75] Utility - the synthetic data must be fit for its defined use; Quality - it must be a sufficient representation of the real data; Privacy - it mustn't 'leak' or expose any sensitive information from the real data. Source: NHS England Transformation Directorate, *Exploring how to create mock patient data (synthetic data) from real patient data*, March 2022

[76] The concept of a "digital chain of custody" is posited in Giuffrè, M., Shung, D.L. *Harnessing the power of synthetic data in healthcare: innovation, application, and privacy*. *npj Digit. Med.* 6, 186 (2023). https://doi.org/10.1038/s41746-023-00927-3. The digital chain of custody provides transparency, traceability and accountability at each stage of the data lifecycle, and therefore safeguards the integrity, security, and privacy of the privately-sourced synthetic dataset.

release of datasets where there is low, but not zero, risk to patient privacy, but are significantly less stringent than those required for anonymised datasets.

Transparency: The process and mechanism of generating a privately-sourced synthetic dataset shall be clearly documented, alongside potential limitations, data biases, and evaluation results (including metrics of similarity to the original dataset). The dataset shall be clearly labelled as "synthetic" at all times.

Informed and Engaged Patients: Where an organisation is generating privately-sourced synthetic data from anonymised patient datasets, it shall provide relevant transparency information and privacy notices to patients (e.g., on the website, in waiting areas). This information should explain the purpose of using synthetic data, the way it is generated, and the implications for individual privacy.

Ongoing Review: Policies and procedures on the generation and release of privately-sourced synthetic data shall be regularly reviewed and updated to ensure that they are fit for purpose given developments in technology, regulation, and privacy threats.

9.4 Roles and responsibilities

Clear roles and responsibilities

Synthetic Data Audit Team: A dedicated Synthetic Data Audit team shall be put in place to oversee the generation, evaluation and audit of privately-sourced synthetic datasets produced and released by the data controller (e.g., NHS Trust). This team shall include skilled and qualified data scientists, and shall rotate in relevant healthcare professionals (dependent on the dataset and use case) to ensure that the synthetic dataset being developed protects patient privacy (i.e., only includes necessary similarities to real data) and is relevant and represents real-world medical scenarios and distributions ("human-in-the-loop").[77] Specific responsibilities shall include:

- Agreeing the generation of a privately-sourced synthetic dataset for a specific use case;
- Determining the technology and mechanism used to generate the privately-sourced synthetic dataset, based on privacy, quality and utility (e.g., this may include agreeing the use of a specific third party synthetic data supplier);
- Overseeing the generation of privately-sourced synthetic datasets, to ensure privacy, quality and utility, including ensuring the inclusion of privacy-enhancing technologies;
- Auditing and evaluating the privacy, quality and utility of privately-sourced synthetic datasets;
- Ensuring that all privately-sourced synthetic datasets are clearly labelled as such and accompanied by clear information on their generation and evaluation;
- Supporting the Information Governance Team to ensure that policies and procedures relating to the generation and release of privately-sourced synthetic data remain robust and up-to-date.

[77] Giuffrè, M., Shung, D.L. *Harnessing the power of synthetic data in healthcare: innovation, application, and privacy*. *npj Digit. Med.* 6, 186 (2023). https://doi.org/10.1038/s41746-023-00927-3

Information Governance Team: The data controller's (NHS Trust's) Information Governance Team shall be responsible for:

- Ensuring that policies and procedures related to privately-sourced synthetic data generation and release reflect the latest NHS guidelines, regulatory guidance and requirements, best practices, and patient views;
- Reviewing the effectiveness and ensuring the implementation of the policies and procedures on at least an annual basis, making changes to policy and procedure as required;
- Approving the release of any privately-sourced synthetic dataset for a specific use case;
- Approving user access control procedures and overseeing their implementation (e.g., approving individual user access).

Patients and Public Involvement Team: The relevant NHS Trust/Biomedical Research Centre's Patients and Public Involvement Team is responsible for:

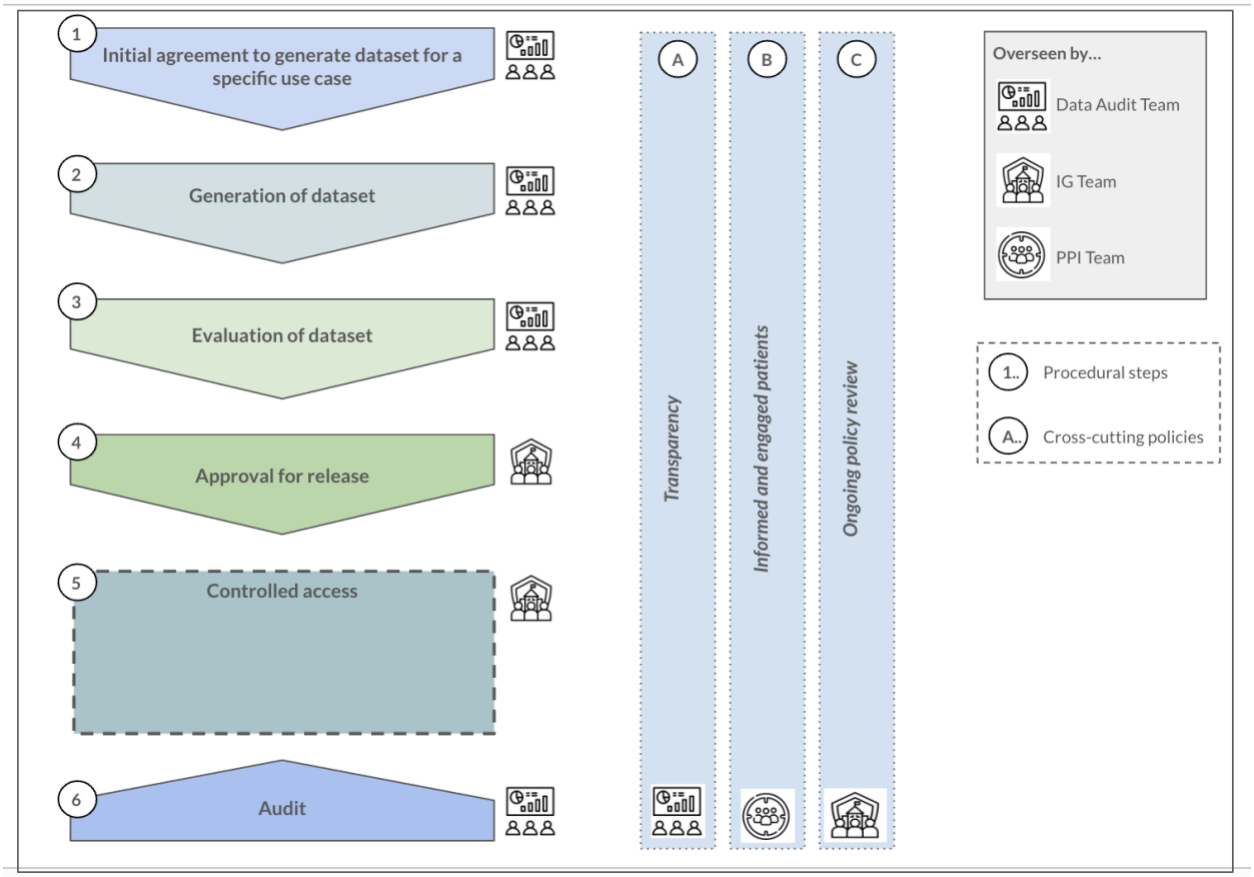
- Drafting clear and transparent communications to explain the purpose of using privately-sourced synthetic data, the way it is generated, and the implications for individual privacy;
- Conducting PPI events to canvas patient views on the generation and use of privately-sourced synthetic data, and feeding these back into the Information Governance team.

9.5 Policies and procedures

Procedure for generation and release of privately-sourced synthetic data

An overview of the proposed procedure for the generation and release of privately-sourced synthetic data is shown in Exhibit 12 below. The exhibit sets out the procedural steps (marked 1-6) as well as the policies governing all activities (marked A-C).

Exhibit 12: Overview of procedure for generation and release of privately-sourced synthetic data



The proposed step-by-step procedure is as follows:

1. Initial agreement to generate a privately-sourced synthetic dataset for a specific use case
Any proposed privately-sourced synthetic dataset shall be required to address a specific use case. This is both to reduce risks to individual privacy and ensure the utility of the dataset.

The Synthetic Data Audit Team shall use its combined data science and clinical research expertise to agree the generation of a privately-sourced synthetic dataset for a specific use case, based on feasibility and impact.

2. Generation of the dataset
In overseeing the generation of a privately-sourced synthetic dataset for a specific use case, the Synthetic Data Audit Team shall:

- Determine the technological mechanism by which the privately-sourced synthetic dataset shall be produced (this can include approach, such as Bayesian trees/GANs, and/or external suppliers);
- Ensure that only necessary statistical information from the real underlying dataset is included to support the approved use case;
- Ensure cost-benefit analysis to consider privacy/utility trade-off before the introduction of any statistical property into a synthetic dataset;

- Ensure the quality and utility of the synthetic dataset through the preservation of biological relationships (e.g., female-specific diseases should not include male patients; well-established clinical symptom-diagnosis pairs are preserved) and alignment of data to "ground truths" for continuous and categorical variables (e.g., age, gender, ethnic distributions of disease);
- Ensure the mitigation of biases in the privately-sourced synthetic dataset;
- Ensure the utilisation of standard statistical disclosure controls as required in generating the synthetic data–e.g., low number suppression, differencing, attribute disclosure, sparsity, identity disclosure, variable minimisation;
- Ensure the implementation of well-evidenced privacy-enhancing techniques such as differential privacy, as appropriate (see Annex 2).

Technical generation of synthetic data may be conducted in-house by specialist data scientists, via a third party, such as MD Clone or Roche, or using open source tools, such as Simulacrum or Synthea. Annex 1 sets out a comparison of commercial and open source synthetic data offerings.

3. Evaluation of the synthetic dataset for privacy, quality and utility

Prior to the release of a privately-sourced synthetic dataset, the Synthetic Data Audit Team shall ensure that a set of checks is conducted to evaluate the privacy, quality, and utility of the dataset. Given that there is currently no industry standard for this type of evaluation, a range of evaluation approaches should be used to provide the broadest possible assessment of the data.[78] Detailed guidelines on a set of approaches that can support a balanced and thorough evaluation of a privately-sourced synthetic dataset are set out in Annex 3. This evaluation methodology is in line with that used by NHS AI Lab Skunkworks in its development and testing of synthetic data.[79] Evaluation may be outsourced to a third party supplier but the Synthetic Data Audit Team must be satisfied that is reliable and robust,

Where the Synthetic Data Audit Team is satisfied that the combined evaluation measures demonstrate that the dataset has sufficient privacy, quality and utility, it shall complete a Data Protection Impact Assessment (DPIA) and request that the relevant Information Governance Team approve the release of the synthetic dataset.

4. Approval for release

Where the Information Governance Team is satisfied that the release of the privately-sourced synthetic dataset presents no significant risk to patient privacy, based on the evaluations of the Synthetic Data Audit Team and the completion of a DPIA, it shall approve the release of the dataset.

5. Controlled Access (Proportionate Security)

Where TREs apply security measures both at the point of access (user access and management) and within the data environment (technical controls such as restrictions on access nodes, package delayed), the proposed security measures for privately-sourced synthetic data are focused on access control (i.e., ensuring users are legitimate and agree to the terms of use) and a clear digital chain of custody to enable traceability and accountability (i.e., ensuring that where users breach terms of use they are held accountable). Where a TRE is described as a "Walled Garden", the security controls required for the handling of privately-sourced synthetic data can be characterised as a "Picket Fence"; this is

[78] This approach was used by the NHS Transformation Directorate's experiment in creating and evaluating synthetic patient data

[79] https://nhsx.github.io/skunkworks/synthetic-data-pipeline

proportionate, as while there is a small residual privacy risk in handling privately-sourced synthetic data, it is lower than that involved in handling anonymised data.

This approach set out below is consistent with best practice for datasets where there is low, but not zero, risk to patient privacy. Specifically, it is aligned with the UK Data Service's approach for "safeguarded" datasets–i.e., datasets that do not require access via a TRE ("controlled datasets"), but nor are they publicly available ("open datasets"; see Annex 4 for further details).[80] It is also consistent with the approaches used by publicly available anonymised datasets sourced from international health records–i.e., MIMIC-IV (USA); AmsterdamUMCdb and eICU Collaborative Research Database (eICU-CRD) v2.0 (USA).[81] These databases, although anonymised, still contain detailed information regarding the clinical care of patients, so have privacy safeguards applied to users to ensure that data is handled with the appropriate sensitivity; however these safeguards are considerably less stringent than those used in a TRE.

Individuals wishing to access privately-sourced synthetic data ("applicants") must comply with the following provisions:

Be an approved researcher with a recognised reference	• Affiliated with a recognised academic or clinical institution • Reference is a clinical, academic research or education lead, easily identifiable as such through an online directory, institutional webpage or equivalent source[82]
Complete a certified, approved training course on the handling of sensitive clinical data	• E.g., completion of NHS Digital's Data Security Awareness (NHSD) course[83] • Proof of completion must be supplied

Complete, sign and submit an Access Request Form and End User Licence Agreement (both applicant and reference)	• The agreement includes:
	  ○ Affiliation and institutional email
	  ○ Intended use (lawful, non-commercial, scientific research or education purposes only)
	  ○ Agreement to Terms of Use
	  ○ Agreement to sanctions if Terms of Use are breached (e.g., removal of user permissions)

[80] UK Data Service, Access Levels and Conditions

[81] MIMIC-IV is a publicly available database sourced from the electronic health record of the Beth Israel Deaconess Medical Center (2008-2019); AmsterdamUMCdb is the first freely accessible intensive care database from within the European Union containing de-identified health data related to tens of thousands of European intensive care unit admissions, including demographics, vital signs, laboratory tests and medication; eICU-CRD comprises 200,859 stays at ICSs and step-down units across 208 hospitals in the continental United States.

[82] Cf., for AmsterdamUMCdb, the reference must be a practising intensivist, easily identifiable as such through an online directory, institutional webpage or equivalent source.

[83] Other potential courses are CITI's Data or Specimens Only Research (DSOR) course (free of charge and currently required to gain access to AmsterdamUMCdb, MIMIC-IV, and eICU) or an equivalent course (e.g., BROK from NFU).

	○ Agreement to publish a note at the end of the project to share what the data has been used for. ● The agreement can be adapted from existing access request forms/end user licence agreements used for datasets which have limited privacy risks, e.g.: ○ The UK Data Service's End User Licence Request Form for Safeguarded Data (adapted version currently used for CHIMERA) ○ AmsterdamUMCdb's Access Request Form and End User Licence Agreement

The Information Governance Team shall agree or refuse applicants based on their compliance with the terms set out above. Upon approval, the privately-sourced synthetic dataset will be available for the user to access.

6. Audit
The Synthetic Data Audit Team shall audit privately-sourced synthetic datasets on a regular basis to ensure that they remain safe in the light of evolving security threats and data changes, as well as to ensure that they do not perpetuate or exacerbate existing biases in healthcare.

Cross-cutting policies
In addition to these procedural steps, we propose the following policies, which cut across all activities:

A. Transparency
When released, all privately-sourced synthetic datasets must be clearly labelled as synthetic and accompanied by a clear account of the methodology of generation as well as the results of evaluation.

B. Informed and Engaged Patients
Where a NHS Trust or Research Centre is generating privately-sourced synthetic data from patient datasets, it shall make concerted efforts to provide transparency information and privacy notices to patients, via a number of channels, for example on websites and in waiting areas. This information should explain the purpose of using privately-sourced synthetic data, the way it is generated, and the implications for individual privacy. Trusts should also use PPI events to gather patient and public views on the generation and use of synthetic data, in order to shape ongoing policy. Annex 5 includes an example script from UCLH for use in an animated patient information video about synthetic data.

C. Ongoing policy review

The Information Governance Team shall keep their policies and procedures relating to the generation and release of privately-sourced synthetic data under continual review, to ensure that they remain effective in an evolving technological and regulatory landscape. NHS Trusts may consider engaging a third party organisation to annually review regulatory compliance, conformance with best practice, and accuracy and safety in relation to the generation and release of privately-sourced synthetic data.

9.6 Phased implementation

A phased rollout is recommended

We propose a phased rollout of the proposed policies and procedures for privately-sourced synthetic data, to safeguard patient privacy and allow for process and governance improvements before broader rollout. Specifically, we propose the selection of 2-3 pilot cases within UCLP. These can be independently evaluated post-hoc, after which process and/or governance changes can be agreed before further rollout. Annex 6 sets out potential use cases for the first phase of rollout.

10. Conclusion

The explosion in digital healthcare data and the rapid advancement of computational science present an enormous opportunity for major advances in clinical research and improvements to patient care and public health. They also present a considerable risk to patient privacy. Important safeguards are therefore put in place to restrict and control access to healthcare data, which in turn have a chilling effect on clinical research and innovation. Synthetic data presents a significant opportunity to improve access to digital healthcare data whilst safeguarding patient privacy.

This paper sets out a simple, proportionate framework to govern the generation and release of different types of synthetic healthcare datasets, based upon their risk to patient privacy. Specifically, it proposes that publicly-sourced synthetic data, as non-personal data, can always be publicly released; whereas privately-sourced synthetic data, which carries a low but variable privacy risk, requires moderate privacy controls, which are significantly less stringent than those in place for real healthcare datasets. The proposed policies, procedures and organisational requirements underpinning this framework have been developed in line with regulatory guidance, expert interviews, academic research, and emerging best practice. A phased implementation approach and ongoing policy review is advised to ensure that patient privacy continues to be safeguarded as the technological and regulatory landscape evolves.

11. Annexes

Annex 1: Survey of synthetic data providers

Overview of commercial synthetic healthcare data providers

Tool	Roche	MD Clone	Veil AI	Replica Analytics
Overview	Pharmaceutical & diagnostics company with synthetic data tools	Israeli-backed company specialising in synthetic healthcare data for research and analytics	Spin-off from University of Helsinki	Technology company that specialises in synthetic data generation for the healthcare industry Acquired by Aetion in 2022
Primary use cases	Clinical trials (often for rare diseases), training machine learning applications	Healthcare analytics (clinical and operational), research and innovation	Machine learning model training	Healthcare, clinical trials and AI models
Generation techniques	Likely includes advanced AI and ML models, statistical modelling Partnerships with in AI-focused startups suggest use of deep learning methods such as GANs	Uses a platform called MDClone ADAMS, uses statistical modelling, likely Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs)	Combination of ML and advanced AI techniques, GANs	Uses a platform called Replica Synthesis, uses Machine Learning and deep learning techniques
Privacy	States that tools are HIPAA/ GDPR compliant and all synthetic models are irreversible	States that data is de-identified and HIPAA- and GDPR-compliant	Differential privacy; privacy risk assessments to ensure HIPAA/ GDPR compliance	Differential privacy; states that synthetic data is “non-identifiable”
Platform/ Interface	Largely focused on Roche use cases and collaborations	Self-service platform, allowing users to query and generate synthetic data on demand	Flexible configuration, API integration	Flexible and customisable
Key customers	Internal use (drug discovery) Partnerships with biopharma Collaborations with hospitals in oncology and rare disease research	Healthcare providers (e.g., University of Chicago Medical Center, Ottawa Hospital), to improve care delivery Academic institutions (e.g., Washington University,), for academic research Government Health Departments (e.g., US Dept of Veterans Affairs), to support public health research/policy	Healthcare technology companies AI startups	Pharmaceutical companies (e.g., Merck Canada) Clinical research organisations Healthcare providers (e.g., Children's Hospital of Eastern Ontario) Regulatory agencies Academic institutions (University of Alberta)

Overview of open source synthetic healthcare data providers

Tool	Synthea	Simulacrum	Synthpop
Overview	Developed by US non-profit (MITRE). Synthetic patient generator that models realistic patient records and clinical histories for research and analysis	Developed by Health Data Insight CiC (HDI), with support from AstraZeneca (AZ) and IQVIA Imitates cancer patient records held securely by the National Disease Registration Service (NDRS) in England	Developed by University of Edinburgh An R Package designed for generating synthetic versions of synthetic microdata for statistical disclosure control
Primary use cases	Creation of realistic EPRs Healthcare research and policy Testing healthcare applications	Cancer research.	Statistical modelling Regression modelling Simulation studies Machine learning development Educational training
Generation techniques	Monte Carlo simulation, Generic Module Frameworks, Python coding (e.g., Medication Diversification Tool) Difficult to specifically generate large numbers of specific subsets (eg. asian women over 45 with diagnosed with x, y and z)	Statistical modelling, probability distributions, clustering algorithms; Bayesian networks; detailed methodology here	Sequential regression modelling, classification and regression trees (CART), and parametric methods; more detail here
Data source	Uses PADARSER, the Publicly Available Data Approach to the Realistic Synthetic EHR, i.e., health incidence statistics (CDC), population statistics (US census), clinical practice guidelines and protocols	Based on datasets held on the National Disease Registration Service (NDRS) Cancer Analysis System (CAS) at NHS England	Does not come with its own predefined data source. Instead, it is an R package designed to generate synthetic versions of any dataset provided by the user.
Privacy	Publicly available data is used	Based on anonymised dataset	Statistical disclosure functions
Platform/ interface	Command-line interface	Command-line interface	R programming
Key users	Academic researchers, IT vendors, public health organisations, healthcare providers, pharmaceutical companies	Healthcare researchers, public health officials, pharmaceutical companies, health IT developers, and educational institutions.	Research (Scottish Longitudinal Study), Teaching (Scottish Centre for Administrative Data Research)

Annex 2: Differential privacy

Differential privacy is a concept first proposed in 2006 by Dwork et al.[84] It is "a precise mathematical constraint that ensures the privacy of individual pieces of information in a database while answering queries about the aggregate. The concept of DP is based on the notion of adding noise to the data to protect the privacy of individuals".[85] To paraphrase, the theory states that if you add enough noise to the synthetic data then it's hard to tell whether any one patient was in the real data. That is why it is called "differential" privacy: it looks to minimise the difference that any particular person makes on the output.

Differential privacy introduces a parameter, ε (epsilon), which determines how much of an effect on the output any individual can have. If every individual has a very small effect on the output (small ε) then their privacy is not infringed when the data is shared. The tradeoff is that there is reduced accuracy in the data once noise has been added. High ε gives higher accuracy but reduced privacy. Low ε gives reduced accuracy but retains privacy.

Differential privacy is "gaining broad acceptance as a solid, practical, and trustworthy privacy framework and its application has been also explored with synthetic data," although it "has not yet seen widespread adoption in the health sector",[86] and its use in relational datasets is limited. New approaches are emerging for application in synthetic healthcare data, for example; for example, Jordan et al. have developed a novel approach incorporating a modification to the Private Aggregation of Teacher Ensembles (PATE) methodology, incorporating it into GANs for the creation of privacy-preserving synthetic data, and have further suggested a new metric, "synthetic similarity". Further information can be found here; Giuffrè and Shung deem the approach "a significant stride forward in the realms of machine learning and privacy [...] enabling the production of high-quality synthetic data that balances privacy protection with utility across datasets and applications".[87]

[84] Dwork, C., *Differential privacy*, In: Bugliesi, M., Preneel, B., Sassone, V., Wegener, I. (eds) Automata, Languages and Programming. ICALP 2006. Lecture Notes in Computer Science, vol 4052. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11787006_1 (2006)

[85] Giuffrè, M., Shung, D.L. *Harnessing the power of synthetic data in healthcare: innovation, application, and privacy*. *npj Digit. Med.* 6, 186 (2023). https://doi.org/10.1038/s41746-023-00927-3

[86] Giuffrè & Shung

[87] Giuffrè & Shung

Annex 3: Evaluation methods for privately-sourced synthetic datasets

This approach replicates that used by NHS AI Lab Skunkworks in their synthetic patient data trial. Given the current absence of an industry standard for the evaluation of synthetic datasets, a basket of evaluation methods as used by Skunkworks enables a robust and broad-based approach for assessing the privacy, quality and utility of a privately-sourced synthetic dataset. The evaluation methods are as follows:

- **An evaluation capability from Synthetic Data Vault (SDV)**: uses metrics to check whether the synthetic data is a good substitute for real data without causing a change in utility.
- **Collision analysis**: checks that no two records are exactly the same in the input and synthetic datasets.
- **Correlation analysis**: compares the relationship between the two datasets to see if patterns have been accurately preserved in the synthetic dataset.
- **Evaluating the Gower distance**: looks at the closeness of similarity between the input and synthetic datasets to make sure they are not too similar.
- **Comparing each dataset using Principal Component Analysis**: reduces the size of the dataset to its principal components whilst keeping as much information as possible to see how similar the input and synthetic datasets are, and therefore to determine whether the synthetic dataset is useful.
- **Propensity testing**: checks whether a model can differentiate between the real and synthetic data. E.g., use a logistic regression model that has been trained on input data; combine the real and synthetic data, then fit the logistic regression model to the data set. Using the fitted model, it is possible to see how well it differentiates between the real and synthetic data by looking at its ability to predict how likely each row was real or synthetic.
- **Comparison of the Voas-Williamson statistic**: a global goodness of fit metric that compares the variation over degrees of freedom in the synthetic and ground truth data.
- **Comparison of statistical distributions of the features:** enables a high level view of the similarity of the two datasets, the categorical and numerical columns were compared visually. For a more in depth overview of both the real and synthetic datasets, pandas-profiling can be used to generate reports for each. Pandas profiling is a way of quickly exploring data using just a few lines of code instead of trying to understand every variable.

The Alan Turing Institute and The Royal Society's joint paper, *Synthetic Data - what, why and how?* (2024), also proposes a set of evaluation methods for synthetic data generation methods and outputs.

Annex 4: UK Data Service: Access Levels and Conditions

Each dataset in the UK Data Service collection has an access level designated by the data provider depending upon the detail, confidentiality and sensitivity of the data: open, safeguarded, or controlled. Differences in how a specific dataset can be accessed vary according to these levels. The table below sets out the data types and the corresponding access levels and conditions. We have added our proposed classification of publicly-sourced synthetic data, privately-sourced synthetic data and anonymised data alongside these levels (green column). We have also assigned monikers to characterise the access controls for each level (purple column).

UK Data Service Access Levels and Conditions

Access Level	Data type	Access Conditions	Proposed classification of synthetic data	Characterisation of access controls
Open data	Non-personal data, no privacy risk	Open access, no user registration.*	Publicly-sourced synthetic data	"Open"
Safeguarded data	Data with some privacy risk, although this risk is very low.	Users must register and accept End User Licence before downloading data.**	Privately-sourced synthetic data	"Picket Fence"
Controlled data	Detailed, sensitive or confidential data	Data can only be accessed via a Trusted Research Environment	Anonymised data	"Walled Garden"

Source: UK Data Service Access Levels and Conditions

* Some may be subject to an Open Government Licence (OGL) or a Creative Commons Licence (CCL)

** Some safeguarded data may have additional conditions attached

Essentially, we propose that:

- Publicly-sourced synthetic data should be released publicly (Tier 0 – "Open").
- Privately-sourced synthetic data should have moderate user access applied (Tier 2 – "Picket Fence") – but these are significantly less strict and cumbersome than those applied in a Trusted Research Environment (Tier 3 – "Walled Garden").

Annex 5: Example UCLH script for patient information video about synthetic data

Health data is an incredible resource that can help us understand health and improve patient care. Analysing health data can be difficult but using something called synthetic data can help us do it more quickly whilst enhancing privacy for patients.

It takes a lot of work and time to prepare data and get to the point where researchers and hospital staff can work together to analyse data about real people's health and treatment.

The problem is the preparation needed. Researchers need to plan their approach and this is difficult without knowing what real heath data looks like. That data is rightly protected - so researchers can't take a peek before all of their approvals are in place. They can only guess at what kinds of things the data will show.

Synthetic data gives researchers something closer to reality, rather than pure guesswork.

Synthetic data gives researchers an idea of what the real data is likely to look like, the trends they are likely to see and the way the data is structured. It isn't the real data. It's just a preview—like a trailer of a movie.

Importantly, synthetic data isn't data about individual real people. You won't be able to identify your neighbour and his gallstones, but you will know that your neighbourhood is likely to include people with gallstones.

Synthetic data is not taken directly from individual patient records. Instead it is created using:

- *the columns that are present in a dataset, and*
- *trends calculated from anonymised, aggregated datasets (for example, the top 10 reasons for hospital admissions, or the age range of people with a certain disease who developed complications)*

We use this available information to weight a set of virtual dice which are then rolled to manufacture data that looks like real patients. Because the dice are weighted, instead of rolling a 15 year old with a heart attack, and a 50 year old with a rugby injury (both rare but possible), the dice are more likely to roll a 15 year old with a rugby injury and a 50 year old with a heart attack.

By learning from real information we can weight the dice to produce high quality synthetic data that is as realistic as possible, without risking patient privacy. And for the researcher this is fantastic.

The researcher can start exploring the kind of trends that are likely to emerge or the kind of code needed to interrogate the real data. So when the researcher actually gets to work on the real data, a lot of time has been saved, the researcher is more likely to get to important information, and the real data doesn't need to be accessed for so long.

And because it is not data about real individual people, anyone can explore it, whether they are students studying computational research, or health workers with a hunch about something to investigate.

Not only that, but we can be more transparent about research using health data by showing people what researchers are doing when they look at health data and how they are doing it.

Annex 6: Potential use cases for phase 1 rollout of privately-sourced synthetic data

Organisation	Description	Contact
DATA-CAN	UK-wide partnership unlocking the power of health data to improve cancer care.	
Digital Health Hub for AMR	Interdisciplinary team focused on harnessing emerging digital technologies to transform antimicrobial stewardship and one-health surveillance across humans, animals and the environment	
UCL Industry Exchange Network (UCL IXN)	Gives UCL computer science students customised industry experience throughout their course, and gives companies access to fast-track innovation.	
UCL EPSRC CDT	Postgraduate study and research in Engineering Solutions for Antimicrobial Resistance.	
UCL Centre for Doctoral Training in Data Intensive Science	CDT programme to develop advanced skills and techniques in Data Intensive Science, including the collection, storage and analysis of large datasets, as well as the use of complex models and algorithms	
tech4health	Collaboration between UCL and Ulster University for Doctoral Training in Digital Health Technologies	
UCL Academic Careers Office (ACO)	Promotes, supports and develops all aspects of academic and clinical academic careers in health related disciplines across UCL. Specific potential use cases via Translational Innovation Networks (collaborations with Industry)	Felipe Fouto
CHIMERA	UCL centre launched in Autumn 2020 uses tools such as machine learning to analyse currently unused intensive care data to find clues that will improve the care of critically ill adults and babies	Prof Rebecca Shipley/ Dr Steve Harris
UCL Computer Science	MSc programmes in computer science	Professor Ivana Drobnjak
International collaborations?		

12. Glossary

Anonymisation	The process by which personal data is irreversibly altered in such a way that a data subject can no longer be identified (by "all the means reasonably likely to be used"), directly or indirectly, including by a motivated bad actor.[88] The practice of anonymisation of healthcare data involves the removal of all directly identifiable traits of patients, alongside additional de-identification procedures (e.g., data perturbation, generalisation, swapping) to reduce disclosure risk (for example, when processing DOB, DOD, rare conditions/ procedures). No reference back to the original patients is retained, so they cannot be directly re-identified by linkage to another dataset.[89]
Confidentiality	Confidentiality in the NHS is the practice of keeping patient information private and respecting the rights of individuals to have their information protected. This includes information about a patient's health, care needs, lifestyle, and family.[90]
Data Protection Impact Assessment (DPIA)	A process designed to identify risks arising out of the processing of personal data and to minimise these risks as far and as early as possible. DPIAs are important tools for negating risk, and for demonstrating compliance with the GDPR.[91]
De-identification	The process by which personal data is altered so that explicit identifiers are removed or hidden, but they could still be linked back to an individual with additional information.
Differential privacy	Security definition which means that, when a statistic is released, it should not give much more information about a particular individual than if that individual had not been included in the dataset[92]. Differential privacy introduces a parameter, epsilon, which determines how much effect any individual can have on the output. High epsilon gives higher accuracy but reduced privacy; low epsilon gives reduced accuracy but retains privacy.
Fidelity	The accuracy, precision, or realism of synthetic data. Also known as "quality".

[88] Chevrier R, Foufi V, Gaudet-Blavignac C, Robert A, Lovis C. Use and Understanding of Anonymization and De-Identification in the Biomedical Literature: Scoping Review. J Med Internet Res. 2019 May 31;21(5):e13484. doi: 10.2196/13484. PMID: 31152528; PMCID: PMC6658290

[89] GOSH DRIVE/DRE, *Position regarding Synthetic Health Data use at GOSH* (Dr William Bryant, Prof. Neil Sebire, Dr Natassa Spiridou), v 1.1., 22/06/2023

[90] NHS Digital, *A guide to confidentiality in health and social care* (March 2022)

[91] Data Protection Commission

[92] The Royal Society, *From privacy to partnership: The role of privacy enhancing technologies in data governance and collaborative analysis* (January 2023)

"Five Safes" Framework	Framework for protecting the confidentiality of personal data. "Safe People": only trained and specifically accredited researchers can access the data; "Safe Projects": data are only used for ethical, approved research with the potential for clear public benefit; " Safe Settings": access to data is only possible using secure technology systems - the data never leaves the Trusted Research Environment; "Safe Data": researchers only use data that have been de-identified to protect privacy; "Safe Outputs": all research outputs are checked to ensure that they cannot be used to identify subjects.[93]
Membership inversion attack	A privacy attack where the attacker analyses model outputs and makes an educated guess about whether a specific datapoint was part of the training dataset (e.g., by querying the model with the datapoint and analysing its response).
Model inversion attack	A privacy attack where the attacker reverse engineers the synthetic model to reconstruct the original data used to train it.
Personal data / personally identifiable information (PII)	Information relating to an identified or identifiable living (or deceased) individual; identifiers include names and other information that would directly or indirectly link those data to that individual.
Privacy	Defined by the UK GDPR as the right and ability of individuals to control the access to and use of their personal data
Privacy Enhancing Technologies (PETs)	A suite of tools that can help maximise the use of data by reducing risks inherent to data use. Some PETs provide new techniques for anonymisation, while others enable collaborative analysis on privately-held datasets, allowing data to be used without disclosing copies of data.[94]
Privately-sourced synthetic data	Synthetic data that has been generated from a real, private dataset in order to specifically represent aspects of the underlying dataset. A synthetic data generator accesses and learns the statistical properties of the real dataset and uses these to produce artificial values.
Publicly-sourced synthetic data	Synthetic data that takes the structure of a real dataset and applies publicly available statistical information from aggregated, anonymised real data. This statistical information is either already in the public domain or has been approved for public release using a standard disclosure control procedure for publishing health statistics.

[93] HDR UK

[94] The Royal Society, *From privacy to partnership: The role of privacy enhancing technologies in data governance and collaborative analysis* (January 2023)

Structural synthetic data	Synthetic data that uses the structure of a real dataset, without retaining and of its statistical information.
Synthetic data	Artificially generated data designed to mimic real data for a particular purpose, but not containing any data directly collected from real patients. Synthetic data is a risk minimisation technique to protect privacy whilst maximising use of data.[95] This paper defines three broad types of synthetic data: structural synthetic data, publicly-sourced synthetic data, and privately-sourced synthetic data
Trusted Research Environment (TRE)	A highly secure computing environment that provides remote access to de-identified health data for researchers to use in research that can save and improve lives. Also known as a "Data Safe Haven".[96]
Utility	The quality and usefulness of data for analysis and decision-making processes. The utility of synthetic data depends upon how well synthetic data mimics the statistical relationships within the source data. There is a relationship between utility and privacy risks.

[95] Great Ormond Street Hospital DRIVE/DRE, "Position regarding Synthetic Health Data use at GOSH" (September 2024)
[96] HDR

13. References, Interviews, PPIs

References

Accenture Life Sciences and Phesi, *Faster and cheaper clinical trials: The benefit of synthetic data*
Arora, A., Wagner, S.k., Carpenter, R., Jena, R., Keane, P.A., *The urgent need to accelerate synthetic data privacy frameworks for medical research*, The Lancet Digital Health, 2024, https://doi.org/10.1016/S2589-7500(24)00196-1
Bellovin, Steven M., Preetam K. Dutta, and Nathan Reitinger, 2019, "Privacy and synthetic datasets", Stanford Technology Law Review 22 (1): 2–52
Chevrier R, Foufi V, Gaudet-Blavignac C, Robert A, Lovis C. Use and Understanding of Anonymization and De-Identification in the Biomedical Literature: Scoping Review. J Med Internet Res. 2019 May 31;21(5):e13484. doi: 10.2196/13484. PMID: 31152528; PMCID: PMC6658290
Darzi, Lord, *Independent Investigation of the National Health Service in England*, September 2024
Data Science Campus, *Synthesising the linked 2011 Census and deaths dataset while preserving its confidentiality*, November 20, 2023
Dwork, C., *Differential privacy*, In: Bugliesi, M., Preneel, B., Sassone, V., Wegener, I. (eds) Automata, Languages and Programming. ICALP 2006. Lecture Notes in Computer Science, vol 4052. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11787006_1 (2006)
El Emam, K., Mosquera, L., Jonker, E., and Sood, E., "Evaluating the utility of synthetic COVID-19 case data", JAMIA Open, 2021, Vol. 00, No. 0; doi: 10.1093/jamiaopen/ooab012
El Emam, K., *Seven Ways to Evaluate the Utility of Synthetic Data*, IEEE 2020
European Parliament. 2016. Regulation (EU) 2016/679 of the European Parliament and the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). https://eur-lex.europa.eu/eli/reg/2016/679/oj
ENISA/European Union Agency for Cybersecurity, *Data Protection Engineering: From Theory to Practice* (January 2022)
Giuffrè, M., Shung, D.L. *Harnessing the power of synthetic data in healthcare: innovation, application, and privacy*. *npj Digit. Med.* 6, 186 (2023). https://doi.org/10.1038/s41746-023-00927-3
Goncalves, A., Ray, P., Soper, B. *et al.* Generation and evaluation of synthetic patient data. *BMC Med Res Methodol* 20, 108 (2020). https://doi.org/10.1186/s12874-020-00977-1
GOSH DRIVE/DRE, *Position regarding Synthetic Health Data use at GOSH* (Dr William Bryant, Prof. Neil Sebire, Dr Natassa Spiridou), v 1.1., 22/06/2023
GOSH DRIVE/DRE: *Use of 'structural synthetic data' at GOSH: Proposed algorithm for approval*
Harris, S. et al., "Critical Care Health Informatics Collaborative (CCHIC): data, tools and methods for reproducible research: a multi-centre UK intensive care database", *International Journal of Medical Informatics* (2018) https://doi.org/10.1016/j.ijmedinf.2018.01.006
Hetherington et al, *Design choices for productive, secure, data-intensive research at scale in the cloud* (2019)
Human Perspectives in Health Sciences and Technology Series, *Personalized Medicine in the Making: Philosophical Perspectives from Biology to Healthcare*, Ed. Chiara Beneduce & Marta Bertolaso: "New Solutions to Biomedical Data Sharing: Secure Computation and Synthetic Data", Edwin Morley-Fletcher
Jobin A, Ienca M, Vayena E., *The global landscape of AI ethics guidelines*, *Nat Mach Intell*. 2019;1:389–399. doi: 10.1038/s42256-019-0088-2
Kokosi, K., De Stavola, B, Mitra, R., Harron, K., at el, "Using synthetic administrative data for research", September 2021

Kokosi, T., and Harron, K. "Synthetic data in medical research", The BMJ (2022)
Kuo, N.IH., Polizzotto, M.N., Finfer, S. *et al.* The Health Gym: synthetic health-related datasets for the development of reinforcement learning algorithms. *Sci Data* 9, 693 (2022). https://doi.org/10.1038/s41597-022-01784-7
Lopez, C. and Elbi, A., "On the legal nature of synthetic data", (2022) NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research
McPherson, R., R. Shokri, and V. Shmatikov. 2016. Defeating image obfuscation with deep learning. *Journal of Petrology* 43 (9). https://doi.org/10.1093/petrology/43.9.1707
Myles et al, "High-fidelity synthetic patient data applications and privacy considerations", Journal of Data Protection & Privacy Vol. 6, 4, 334–354
NHS Digital, *A guide to confidentiality in health and social care* (March 2022)
NHS England Transformation Directorate, *Exploring how to create mock patient data (synthetic data) from real patient data*, March 2022
Page, Hector, Charlie Cabot, and Kobbi Nissim. 2018. *Differential privacy: An introduction for statistical agencies*. National Statistician's Quality Review. https://gss.civilservice.gov.uk/ wp-content/uploads/2018/12/12-12-18_FINAL_Privitar_Kobbi_Nissim_article.pdf
PHG Foundation, *Are synthetic health data 'personal data'? A PHG Foundation report independently commissioned by the MHRA to assess the status of synthetic health data in UK data protection law* (May 2023)
Quality Centre. 2018. *Government Statistical Service, Privacy and data confidentiality methods: A National Statistician's Quality Review (NSQR)*, https://gss.civilservice.gov.uk/policy- store/privacy-and-data-confidentiality-methods-a-national-statisticians-quality-review-nsqr/
Rosenblatt, L. et al, "Differentially private synthetic data: applied evaluations and enhancements" (2020), https://arxiv.org/abs/2011.05537
Stadler, T., Oprisanu, B., and Troncoso, C., "Synthetic Data - Anonymisation Groundhog Day", 31st USENIX Security Symposium (USENIX Security 22), 2022 https://doi.org/10.48550/arXiv.2011.07018
Taichman, D.B. et al. *Data sharing statements for clinical trials: a requirement of the International Committee of Medical Journal Editors*. The Lancet 378, e12-e14 (2017)
The Alan Turing Institute and The Royal Society, *Synthetic Data - what, why and how?* (2024)
The Royal Society, *From privacy to partnership: The role of privacy enhancing technologies in data governance and collaborative analysis* (January 2023)
Yue et al,., "Synthetic Text Generation with Differential Privacy: A Simple and Practical Recipe", July 2023, arXiv:2210.1438
Understanding Patient Data

Interviews

Name and title	Organisation	Date

Dr May Yong Senior Research Software Engineer	The Alan Turing Institute	15/04/24
Martin Donnelly Head of Data Governance	Advanced Research Computing, UCL	21/03/24
Professor Neil Sebire Professor of Paediatric and Developmental Pathology	UCL GOS Institute of Child Health	12/09/24

Public and Patient Involvement Workshops (PPIs) - BRC

Subject	Date	Attendees
Use of Synthetic Data in Research and Education	01/24	Public contributors, patients, Masters students
Use of Data in Research	03/21	UCLH patients, local residents and members of public, information governance specialists, clinicians, UCLH non-executive director