Synthetic health data Community of Practice (SynD)
Draft Synthetic Health Data Governance Framework, 30 September 2025

Table of Contents

Overview

For the purposes of this Framework, synthetic health data is health data generated by a system or model that mimics and resembles the structure and statistical properties of real health data, and uses real health data as input.

Purpose of this Synthetic Health Data Governance Framework

The Synthetic health data Community of Practice (SynD) was established by the Digital Health Cooperative Research Centre (Digital Health CRC) as a collaborative initiative to advance health data research and related activities in Australia through the use of synthetic health data.

SynD's mission is:

> "To unlock the value of health information through the use of synthetic health data to advance research, education, innovation and service delivery within the health and care sector."[1]

SynD have created this draft Synthetic Health Data Governance Framework (Framework) to support the safe and efficient generation and use of synthetic health data across a range of use cases. Synthetic health data has the potential to unlock the value of health information while protecting the privacy of individuals. This Framework is therefore intended to provide a practical set of guardrails for data custodians, data scientists, researchers and other stakeholders to confidently create, use and share synthetic health data while effectively reducing and managing residual privacy risks.

How to use this Framework

This Framework is intended to apply to use cases that involve either the collection or use of *real* health data for the purposes of generating, using and sharing *synthetic* health data.

[1] From Digital Health CRC, *Synthetic health data Community of Practice: Terms of Reference* (V1.0) available at: https://digitalhealthcrc.com/wp-content/uploads/2024/10/Digital-Health-CRC_Synthetic-Data-Community-of-Practice_Terms-of-Reference_v1.0.pdf

‘Real’ data is information about actual, real-world subjects – as distinct from information that is fictitious or imaginary. Real data can include facts, observations, opinions (whether they are true or not) and can also be used to convey, imply or infer meaning or insights.

This Framework only applies to health data about real people. For the purposes of this Framework, ‘real health data’ is any data or information that is about, or relates to, an actual individual, who can be a living person or a person who has died. It includes information that relates to a person’s health status (physical or mental), health services they received, preferences about future health services, genetic information, as well as *any* personal information collected in the course of receiving a health service. (See the Glossary for a discussion of ‘personal information’ and ‘health information’).

An organisation may hold real health data because it has collected health information directly from the individuals themselves (for example, through providing a health service), or from another source such as another individual (such as a family member), another organisation (such as another health service provider), via publicly available information, or where the organisation has generated or created the information themselves.

This Framework is not intended to apply to *other* types of data that are not about or do not relate to individuals. For example, information about systems, operations or devices that do not relate to individuals will not be in scope of this Framework. ‘Mock data’ or ‘dummy data’ is also not subject to this Framework. ‘Mock data’ and ‘dummy data’ is data that has *not* been derived from, generated from or is otherwise based on information about actual individuals. Mock data and dummy data are entirely and intentionally fabricated. Mock data and dummy data could be generated using a model that applies statistical and relational rules, but *that does not otherwise* use or consume real data in doing so.

The Synthetic Health Data Governance Framework is set out below, and includes five steps that must be worked through for each proposed use case or request for synthetic health data. Each step requires the completion of mandatory assessments designed to assess and manage privacy and other related risks. These assessments are found in the Appendices to the Framework, along with guidance and checklists intended to support organisations with understanding the Framework requirements and documenting the outcomes of their assessments. By following this Framework, organisations can advance use cases with very low privacy risks in a more efficient way that might otherwise be needed for use cases with higher levels of privacy risk.

What is synthetic health data?

‘Synthetic health data’ is not defined in Australian privacy law. A succinct description of synthetic data is offered by the Office of the Privacy Commissioner of Canada:

> “synthetic data is fake data produced by an algorithm whose goal is to retain the same statistical properties as some real data, but with no one-to-one mapping between records in the synthetic data and the real data.

> In terms of output… synthetic data looks like unmodified identifiable data. Even though it is fake, it retains the same structure and level of granularity as the original”.[2]

Similarly, the UK Information Commissioner’s Office defines synthetic data thus:

> “Synthetic data is ‘artificial’ data generated by data synthesis algorithms, which replicate patterns and the statistical properties of real data … when you analyse the synthetic data, the analysis should produce very similar results to analysis carried out on the original real data”.[3]

For the purposes of this Framework, synthetic health data is health data generated by a system or model that mimics and resembles the structure and statistical properties of real health data, and uses real health data as input.[4]

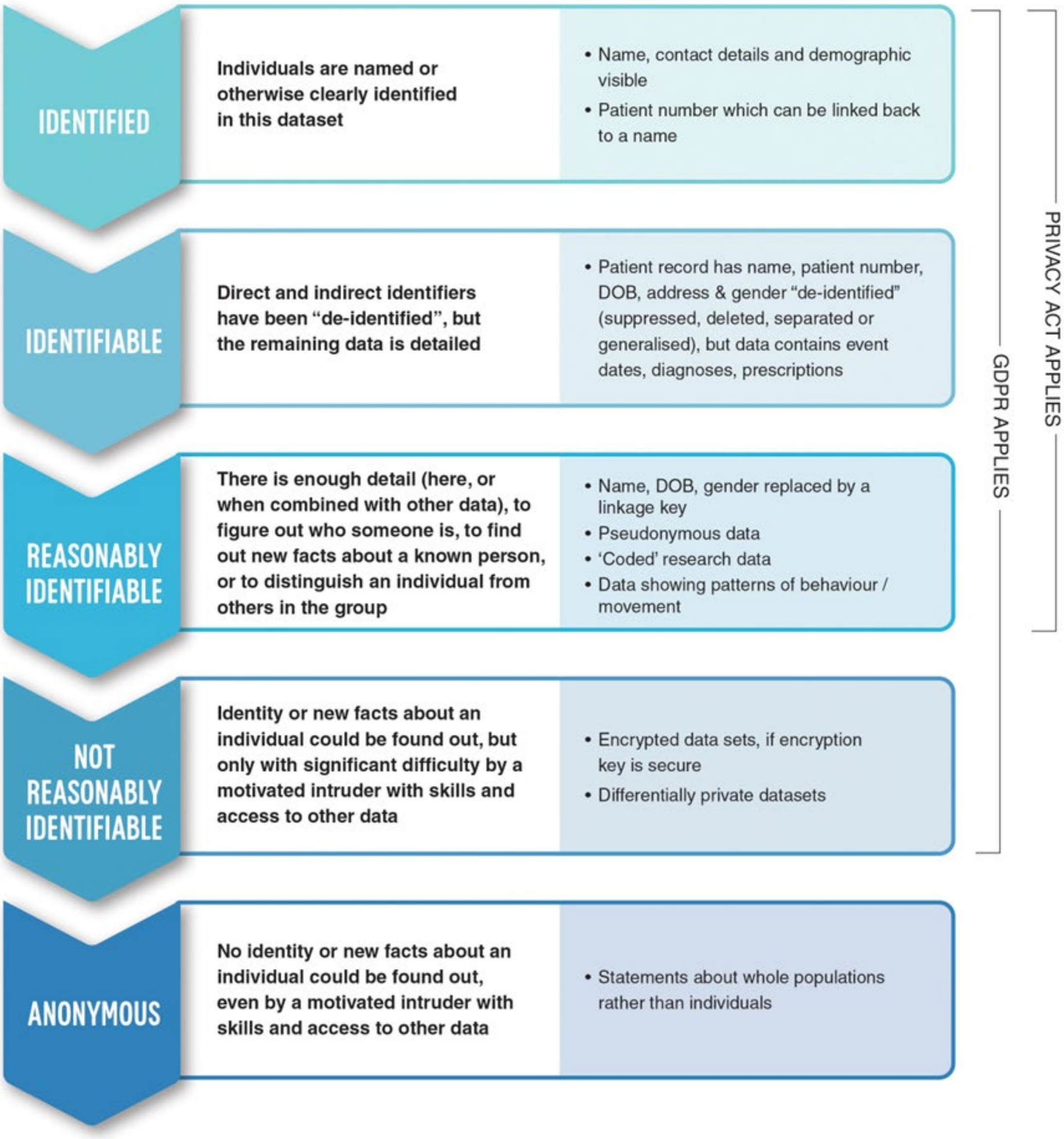
There are different types of synthetic health data as well as different methods that can be used to create it. Not all synthetic health data carries material privacy risk. The level of privacy risk will depend on both the type of source data used to generate synthetic health data (which informs the ‘inherent’ privacy / disclosure risk) and the extent to which the synthetic health data reflects the original source data (the ‘residual’ privacy / disclosure risk).

[2] Office of the Privacy Commissioner of Canada, “When what is old is new again – The reality of synthetic health data”, OPC blog post, 12 October 2022; available at https://priv.gc.ca/en/blog/20221012

[3] UK Information Commissioner’s Office, “Chapter 5: Privacy-enhancing technologies (PETs) – Draft”, September 2022, p.35; available from https://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/ico-call-for-views-anonymisation-pseudonymisation-and-privacy-enhancing-technologies-guidance/

[4] This definition builds on the definition used by the IAPP: https://iapp.org/resources/article/key-terms-for-ai-governance/

Inherent privacy risk of a source dataset based on degree of identifiability



IDENTIFIED	Individuals are named or otherwise clearly identified in this dataset	• Name, contact details and demographic visible • Patient number which can be linked back to a name
IDENTIFIABLE	Direct and indirect identifiers have been "de-identified", but the remaining data is detailed	• Patient record has name, patient number, DOB, address & gender "de-identified" (suppressed, deleted, separated or generalised), but data contains event dates, diagnoses, prescriptions
REASONABLY IDENTIFIABLE	There is enough detail (here, or when combined with other data), to figure out who someone is, to find out new facts about a known person, or to distinguish an individual from others in the group	• Name, DOB, gender replaced by a linkage key • Pseudonymous data • 'Coded' research data • Data showing patterns of behaviour / movement
NOT REASONABLY IDENTIFIABLE	Identity or new facts about an individual could be found out, but only with significant difficulty by a motivated intruder with skills and access to other data	• Encrypted data sets, if encryption key is secure • Differentially private datasets
ANONYMOUS	No identity or new facts about an individual could be found out, even by a motivated intruder with skills and access to other data	• Statements about whole populations rather than individuals

Source: Helios Salinger [5]

Where a synthetic health data use case starts with a low inherent privacy risk due to the type of source data, this Framework does not need to be applied.

However, where the source data is about *individuals* (whether they are alive or have died), the synthetic health data project will have a 'high' inherent privacy risk and this Framework must be applied. Individuals do not need to be named in a dataset for the data to be about them or for them to be 'identifiable'.

[5] Helios Salinger, *Demystifying De-Identification*, August 2025. Available at: https://www.heliossalinger.com.au/downloads/demystifying-deid/

Input data type / source	Does this Framework apply?
Data about subjects that do not relate to individuals. E.g. data about inventories, systems	No. Inputs have very low inherent privacy risk.
Mock data about fictitious individuals, i.e. random data that is not derived from a real dataset or based on real people. May be generated using statistical modelling. E.g. random fictitious numbers, names or email addresses representing a fictitious or virtual population	No. Inputs have low inherent privacy risk.
Real data about individuals collected from publicly available sources. E.g. names and details about individuals collected from websites	Yes. Inputs have high inherent privacy risk.
Real data about individuals already held by organisations, originally collected for another purpose. E.g. patient records in clinical systems.	Yes. Inputs have high inherent privacy risk.

The challenge when creating synthetic health data is how to make it appear real, or 'real enough', to support the relevant use case, while protecting real patients' privacy. Whilst balancing the trade-off between fidelity, utility and privacy, the privacy protection should be the first principle for synthetic health data generation and management.

By applying this Framework, data custodians can effectively manage and assess the privacy risks associated with both the source data being used to generate the synthetic health data and the synthetic health data itself, and reduce the level of privacy risk to an acceptable level to support the use case at hand.

Further information about synthetic health data and how it is generated is set out in Appendix 1 below.

Benefits of synthetic health data

Health data is critical for health research and for generating insights into, and responding to, the needs of Australian health consumers and the Australian health system. Restrictions on accessing health data, or lengthy delays in accessing health data, can have an impact on providing timely and beneficial outcomes for health consumers and on improvements and innovations for stronger healthcare systems.

Synthetic health data has the potential to provide opportunities for organisations to access the *statistical value* of real health data while reducing the risk of *statistical disclosure,* which could lead to the identification of individuals whose information is included in the original or source dataset.

Synthetic health data can therefore be seen as an ‘enabler’ for many use cases by effectively reducing many of the risks and complexities associated with using real health data, while also providing a range of benefits for multiple stakeholder groups.

Potential use cases for synthetic health data

- **Health research and related activities**

  Currently, it can take considerable time for researchers to access sufficient health data for research projects. These delays in accessing health data contribute to inefficiencies in current health data research workflows, where researchers must wait until they receive data before progressing their funded research projects.

  Synthetic health data has the potential to improve the scale and quality of health research and expand data capacity (and health system outcomes overall) in the face of these delays by:

  - Allowing researchers to be more productive in the time between receiving project funding and receiving access to real health data by carrying out preliminary research tasks (noting that real health data will be used to finalise projects)
  - Allowing researchers to conduct Proof of Concept (PoC) analysis and feasibility testing to inform whether they proceed with research funding requests
  - Allowing research and funding bodies and health data custodians to use PoC results to vet research proposals when evaluating and prioritising which projects to fund.

- **Healthcare management activities**

  Synthetic health data can be used to carry out preliminary steps ahead of using real health data for healthcare system testing, management, planning, funding and evaluation activities.

- **Education and training activities (including hackathons)**

  Synthetic health data can be used for educational and training purposes where the outcomes of these activities are not used for other purposes (such as research or clinical purposes, for example). Analysing health data requires specialised analytical skills; synthetic health data can be used to train the next generation of health data scientists by providing them with access to a broad range of data that might not otherwise be available or accessible.

- **Health technology development and improvement activities**

Synthetic health data can be used as an alternative for real health data for health technology development and improvement activities such as designing medical devices/clinical analytics tools to the PoC stage, testing new clinical systems, or building and/or improving AI systems.

- **Testing and validation (non-representative) activities**

Synthetic health data offers a secure and efficient alternative to production test data, mitigating risks in testing environments. It also allows for the creation of diverse and complex scenarios. Synthetic health data can also be used to improve model performance in populations under-represented in real health datasets.

Benefits for key stakeholders

Synthetic health data has the potential to provide a range of different benefits to key stakeholders, in addition to the use cases described above. For example:

- **Healthcare system consumers and community stakeholders**

Using health data for secondary purposes (such as research or management of healthcare activities) can sometimes create a tension between an individual's right to privacy and accessing the value of their personal information for these purposes. Synthetic health data can be a powerful safeguard to protect and respect patient privacy and dignity while also supporting these activities.[6]

Synthetic health data offers stronger privacy protection than real health data, and reduces the risk and severity of privacy related-harms that could arise from misuse, interference, unauthorised access or disclosure, or loss of real health data.

Health system consumers and communities can also receive the benefits of more *efficient* research and health management activities that use synthetic health data (as well as consumers and communities providing input into the design of these activities), where they ultimately result in faster improvements being made to the delivery of healthcare services to communities (including those represented by the data).

- **Health Data Custodians**

Synthetic health data can help improve data requests and efficiencies with data approvals. It can also be used for modelling and creating more representative datasets that may not already exist, or to create customised datasets to ensure they are fit-for-purpose, including addressing risks around bias, under-representation, data availability and / or gaps or errors in real data. Synthetic health data can also be used

[6] See *FZP v Sydney Children's Hospitals Network* [2025] NSWCATAD 144 at [17] for a statement which illustrates the tension between research and privacy. Available at: https://www.caselaw.nsw.gov.au/decision/1977c6de0803ecdd2bd5e685

to increase the scalability, volume and quality of datasets needed for analysis. It can reduce the need to collect data to create new datasets (and the time and expense it takes to do this), or to annotate, correct or otherwise 'clean' real datasets before they are suitable to be used for analysis. Synthetic health data also helps reduce the privacy risks associated with handling and protecting real health data.

- **Health Organisations**

Health organisations, such as state and territory health departments, Primary Health Networks, and private health organisations, can use synthetic health data for more efficient healthcare system testing, management, planning, funding and evaluation activities. Using synthetic health data for these activities also helps reduce the privacy risks associated with handling and protecting real health data.

- **Researchers / Research Organisations**

Using synthetic health data for research related activities (including PoC and analysis and feasibility testing) can help inform whether to proceed with a research funding request / research protocol, saving time, effort, cost and resources. Using synthetic health data for these activities also helps reduce the privacy risks associated with handling and protecting real health data.

- **Ethics Committees and Data Governance Committees**

Synthetic health data reduces the complexity of privacy considerations that must be balanced against public interest considerations when using real health data. It could also allow Ethics Committees to provide an early step in a Research Protocol to provide additional quality assurance, governance and ethics support.

Where synthetic health data has a very low re-identification risk (such that it is effectively 'de-identified' and no longer 'personal information' or 'health information' under privacy law) statutory privacy Guidelines (such as approved under s 95 and s 95A of the *Privacy Act 1988* (Cth), as well as those issued under the *Health Records and Information Privacy Act 2002* (NSW)) will not apply.

Under the *National Statement on Ethical Conduct in Human Research*, research projects that involve the collection, use and disclosure of 'de-identified' data may also be eligible for 'lower risk research' ethics review pathways (where the research also otherwise carries a lower risk to participants or the community).[7] While researchers will still need to address other ethical considerations and follow Institutional ethics processes in line with the *National Statement*, synthetic health data with very low re-identification risk does not require a 'waiver of consent' granted by a Human Research Ethics Committee (HREC) as this data is not 'personal information' as defined by privacy law. Where ethics review and HREC approval is required, seeking preapproved types of research using a synthetic health dataset from an HREC, or

[7] See National Health and Medical Research Council, *National Statement on Ethical Conduct in Human Research* (2025) at 5.1.15 – 5.1.18. Available at: https://www.nhmrc.gov.au/about-us/publications/national-statement-ethical-conduct-human-research-2025

approval of a ‘data governance and use case approval’ approach under a research protocol, may also be options to streamline use of the data.

Protecting privacy with synthetic health data

The secondary use of health data can bring enormous benefits. However, any use of health data elevates the risk of harm to patients in the event of any misuse, interference, unauthorised access or disclosure, or loss of the data.

Privacy laws create guardrails for the use of health data in order to strike the right balance between those potential risks and benefits.

Synthetic health data is often promoted as ‘privacy enhancing technology’ that can assist organisations to protect the confidentiality and privacy of the data they hold. Synthetic health data can provide a technical alternative to using real data for analysis and insight generation, thereby reducing the need for researchers to access real health data about individuals. Synthetic health data can be generated in such a way that it has many of the same statistical properties as the source dataset, and for many use cases, synthetic health data will ‘look real enough’ to be appropriate for analysis, while materially lowering privacy risk.

Synthetic health data can help organisations manage privacy risks in two key ways:

1. Synthetic health data as a data security measure (privacy risk mitigation)

   Reducing the amount of real data and transforming it into synthetic health data reduces the likelihood that an individual can be identified or ‘singled out’ from the dataset,[8] or that inferences about them can be made from the data. In this context, synthetic health data can be a powerful privacy risk mitigation tool to protect individuals represented in the real dataset from harm, in the event of any misuse, interference, unauthorised access or disclosure, or loss of the data. Organisations may therefore decide to use synthetic health data as a data security measure *even where* a researcher is or could be authorised to access and use the real data for analysis, but where synthetic health data is sufficient for their use case.

2. Synthetic health data as a strategy to enable uses not otherwise possible (legal strategy)

   Synthetic health data can be used as an alternative to real health data, where the use of real health data would *otherwise not be permitted* under privacy law. However, whether this strategy will be effective will depend on a) what statistical properties are required to be maintained in the synthetic health dataset to ensure the data still has the necessary level of analytical value and utility for the use case at hand, and b) the types of controls used to manage the residual re-identification risks in relation to the

[8] A person may be ‘identifiable’ if they can be ‘distinguished’ from all other members of a group. This may not necessarily involve identifying the person by name.

environment in which the data is stored and accessed. Data custodians seeking to use or share synthetic health data as a legal strategy to support particular use cases that would *otherwise not be permitted* under privacy law will need to robustly test for re-identification risks before they can confidently consider the data to be *not reasonably identifiable of any real individual* such that the privacy law no longer applies.

Limitations of synthetic health data

Synthetic health data is not risk-free and is not suitable for all use cases. The challenge when creating synthetic health data is how to make it appear real, or ‘real enough’, to support the relevant use case. Organisations should also expect that there will be residual privacy risks associated with synthetic health datasets that need to be managed, as well as other risks and impacts (such as ethical considerations) that may need to be managed.

The statistical value of synthetic health data will vary depending on the extent to which the source data has been altered to create the synthetic health dataset, as well as how ‘real’ and accurate the outputs are required to be for a particular use case. Altering values or statistical properties associated with real data when generating synthetic health data can risk concealing or ‘losing’ potentially useful properties from the source data (i.e. it may not be ‘real enough’ to support a particular use case). On the other hand, if anomalies or outliers from the source data are not removed, the risk of statistical disclosure may increase (i.e. the risk that an individual can be identified from the synthetic health data).

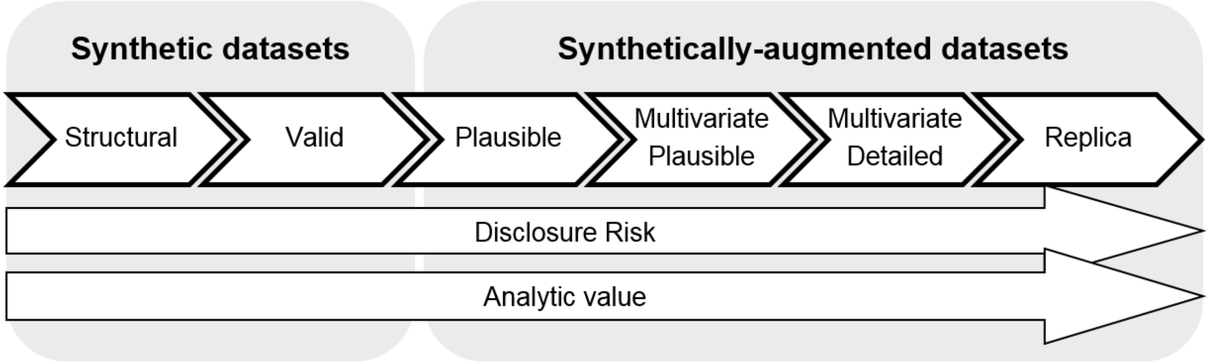
Altering source data will also impact data accuracy. While synthetic health data aims to replicate the structure and patterns of the source data, it does not replicate the exact distribution of the source data. Where a use case requires highly accurate data for analysis and decision-making, synthetic health data will not be the best solution.

Generating synthetic health data also relies on the quality of the source data. Data issues present in the source data (such as inaccurate, incomplete or outdated data) can impact the quality of the synthetic health data. Biases can also be carried across to the synthetic health data, where they exist in the source data or can be inadvertently created by the model used to generate the synthetic health data.

Whether these limitations will materially impact the suitability of synthetic health data for a particular use case very much depends on the use case at hand and the level of data accuracy that is required in the circumstances. In most circumstances, synthetic health data will be suitable for the use cases described above (‘Benefits of synthetic health data’). Where a very high degree of accuracy is required in a synthetic health dataset, organisations should plan to validate that the dataset is accurate *enough* for the use case before proceeding with analysis.

It is also worth noting that a single source dataset may give rise to multiple synthetic versions, with the specific characteristics of each version depending in part on its intended use case. Because not all information can be preserved simultaneously, data scientists and researchers generating synthetic data can exercise discretion in determining which correlations and structural features should be reproduced with greater fidelity.

Synthetic health dataset spectrum: a high-level scale to evaluate synthetic health data based on how close the synthetic health data resembles the original data, the purpose of the synthetic health data and the disclosure risk



Source: UK Office for National Statistics (ONS)[9]

When will synthetic health data <u>not</u> be suitable for a particular use case?

Noting the above limitations, synthetic health data will not be an appropriate substitute for real data for all use cases. While synthetic health data is statistically similar to real data, it is not an exact replica. This means that for use cases where data accuracy is critical, real data should be used. This could include use cases where there is a risk of an *adverse consequence* to an individual (or group of individuals) in circumstances where:

- the data being used has *any* inaccuracies
- the data excludes outliers from the real source data, or
- where the decision being made requires a high level of credibility and assurance.

For example, while synthetic health data may be suitable for developing and testing clinical support tools, synthetic health data itself should not be used as a substitute for real health data about a patient when making a medical diagnosis. Synthetic health data will also be unsuitable where research projects require real health data for certain analyses (e.g. where highly accurate data is required), or when dealing with legal requests.

[9] ONS methodology working paper series number 16 - Synthetic health data pilot, 15 January 2019. Available at: https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaperseriesnumber16syntheticdatapilot