

DRAFT**Synthetic Data Community of Practice (SynD)***Synthetic Health Data Governance Framework***Table of Contents**

Overview	2
Purpose of this synthetic health data governance framework	2
How to use this framework	2
What are the benefits synthetic health data?	3
Protecting privacy with synthetic health data	4
Limitations of synthetic data	5
What is synthetic data?	7
About this Framework	10
Scope	10
Audience	13
Responsibility for decision-making about synthetic data	13
Responsibilities under this framework – who to consult	15
The Framework	16
Guiding principles for handling synthetic data requests	16
Steps for generating and accessing Synthetic Data	17
Step 1: Assess the use case	17
Step 2: Assess and prepare the source data	19
Step 3: Generate the synthetic data	20
Step 4: Assess and manage re-identification risks	21
Step 5: Manage residual privacy risks	25
Final steps	25
APPENDIX 1: About synthetic data	28
A source dataset	28
A generative model	29
The synthetic data	29
APPENDIX 2: Glossary	31
APPENDIX 3: The policy and legal framework underpinning this framework	37

APPENDIX 4: Use case assessment.....	42
APPENDIX 5: Impact assessment.....	44
APPENDIX 6: Technical assessment	49
APPENDIX 7: De-identification techniques.....	52
APPENDIX 8: Decision tree for complex synthetic data scenarios	56
APPENDIX 9: The lawful pathways explained	58
APPENDIX 10: Safety Assessment	66
APPENDIX 11: Synthetic data request assessment outcomes form	75
APPENDIX 12: Privacy obligations regulating the use & disclosure of health information.....	83

Overview

Purpose of this synthetic health data governance framework

The Synthetic 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 in Australia through the use of synthetic data.

SynD's mission is:

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

SynD have created this draft synthetic data health governance framework to support the safe and efficient generation and use of synthetic 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 intended to provide a practical set of guardrails for data custodians, data scientists, researchers and other stakeholders to confidently create, use and share synthetic 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 and using synthetic health data. ‘Real’ health data is any data or information that is about or relates to an individual, who can be a living person or a person who has died. Health data is 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 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 is not about or does not relate to individuals. For example, information about systems, operations or devices that do

¹ From Digital Health CRC, *Synthetic 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

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.


The framework is set out below under "This framework". The framework includes five steps that must be worked through for each proposed use case or request for synthetic 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. . 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.

What are the benefits 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. [As an alternative, I can incorporate the above information in a table format, e.g. populate the below table. We could either separate by stakeholder and / or use case. If that is the desired approach, I suggest holding a small workshop to help me articulate the range of benefits against each use case.]



Synthetic Health Data potential use cases:


Use Case	Key Stakeholder Benefit (add columns for additional stakeholders)			
	Health Consumer / Community Stakeholders	Data Custodian	Researchers /Research Organisation	Ethics / Data Governance Committee
Proof of Concept analysis and feasibility testing	Can help support Consumers engagement in Research Design including input into Research Questions	Can help improve data requests and efficiencies with data approvals.	Could inform whether to proceed with a research funding request / research protocol saving time, effort, cost etc.	Could provide an early step in a Research Protocol to provide additional quality assurance, governance and ethics support.
All use cases	Offers strong privacy	Helps reduce privacy risks	Helps reduce privacy risks	Reduces complexity of

	protection and reduces risk of privacy related-harms	associated with handling real data 	associated with handling real data	privacy considerations that must be balanced against public interest considerations when using real 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 data is often promoted as ‘privacy enhancing technology’ that can assist organisations to protect the confidentiality and privacy of the data they hold. Synthetic 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 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 data will ‘look real enough’ to be appropriate for analysis, while materially lowering privacy risk. 

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

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

Reducing the amount of real data and transforming it into synthetic data reduces the likelihood that an individual can be identified or ‘singled out’ from the dataset,³ or that inferences about them can be made from the data. In this context, synthetic 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 data as a data security measure *even where* a researcher is or could be authorised to access and use the real data for

³ 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.

analysis, but where synthetic data is sufficient for their use case.

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

Synthetic 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 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 environment in which the data is stored and accessed. Organisations seeking to use synthetic 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. such that the privacy law no longer applies.

Limitations of synthetic data

Synthetic data is not risk-free and is not suitable for all use cases. The challenge when creating synthetic 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 datasets that need to be managed.

The statistical value of synthetic data will vary depending on the extent to which the source data has been altered to create the synthetic 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 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 increases (i.e. the risk that an individual can be identified from the synthetic data).

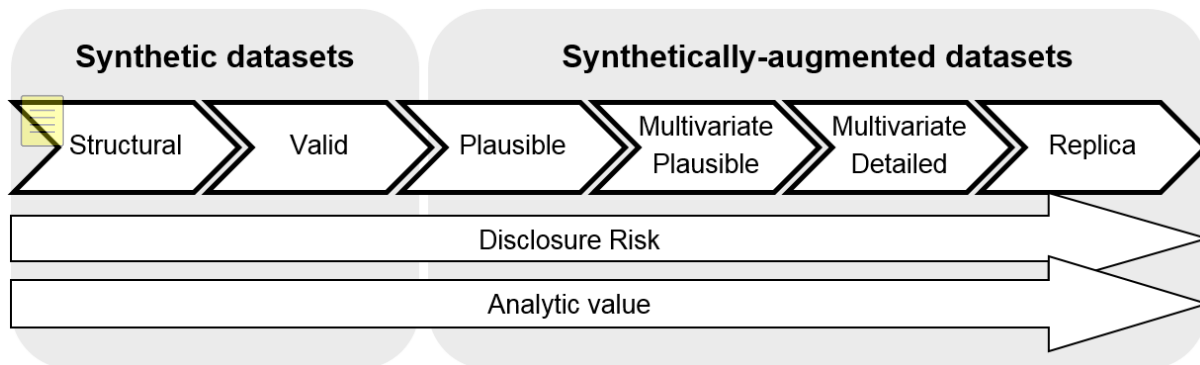
Altering source data will also impact data accuracy. While synthetic data aims to replicate the structure and patterns of the source data, it does not replicate the exact distribution of the source data. Where a use case requires highly accurate data for analysis and decision-making, synthetic data will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

Generating synthetic 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 data. Biases can also be carried across to the synthetic data, where they exist

in the source data or can be inadvertently created by the model used to generate the synthetic data.

Whether these limitations will materially impact the suitability of synthetic 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 data will be suitable for the use cases described above ('What are the benefits of synthetic health data?'). Where a very high degree of accuracy is required in a synthetic dataset, organisations should plan to validate that the dataset is accurate *enough* for the use case before proceeding with analysis.

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



Source: UK Office for National Statistics (ONS)⁴

When will synthetic data not be suitable for a particular use case?

Noting the above limitations, synthetic data will not be an appropriate substitute for real data for all use cases. While synthetic 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.

⁴ ONS methodology working paper series number 16 - Synthetic data pilot, 15 January 2019. Available at: <https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaper/series/onsmethodologyworkingpaperseriesnumber16syntheticdatapilot>

For example, synthetic data should not be used for use cases involving clinical decision-making (such as medical diagnoses), where research projects require real data for certain analyses, or when dealing with legal requests.

What is synthetic data?

'Synthetic 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".⁵

For the purposes of this framework, synthetic data is data generated by a system or model that mimics and resembles the structure and statistical properties of real data.⁷

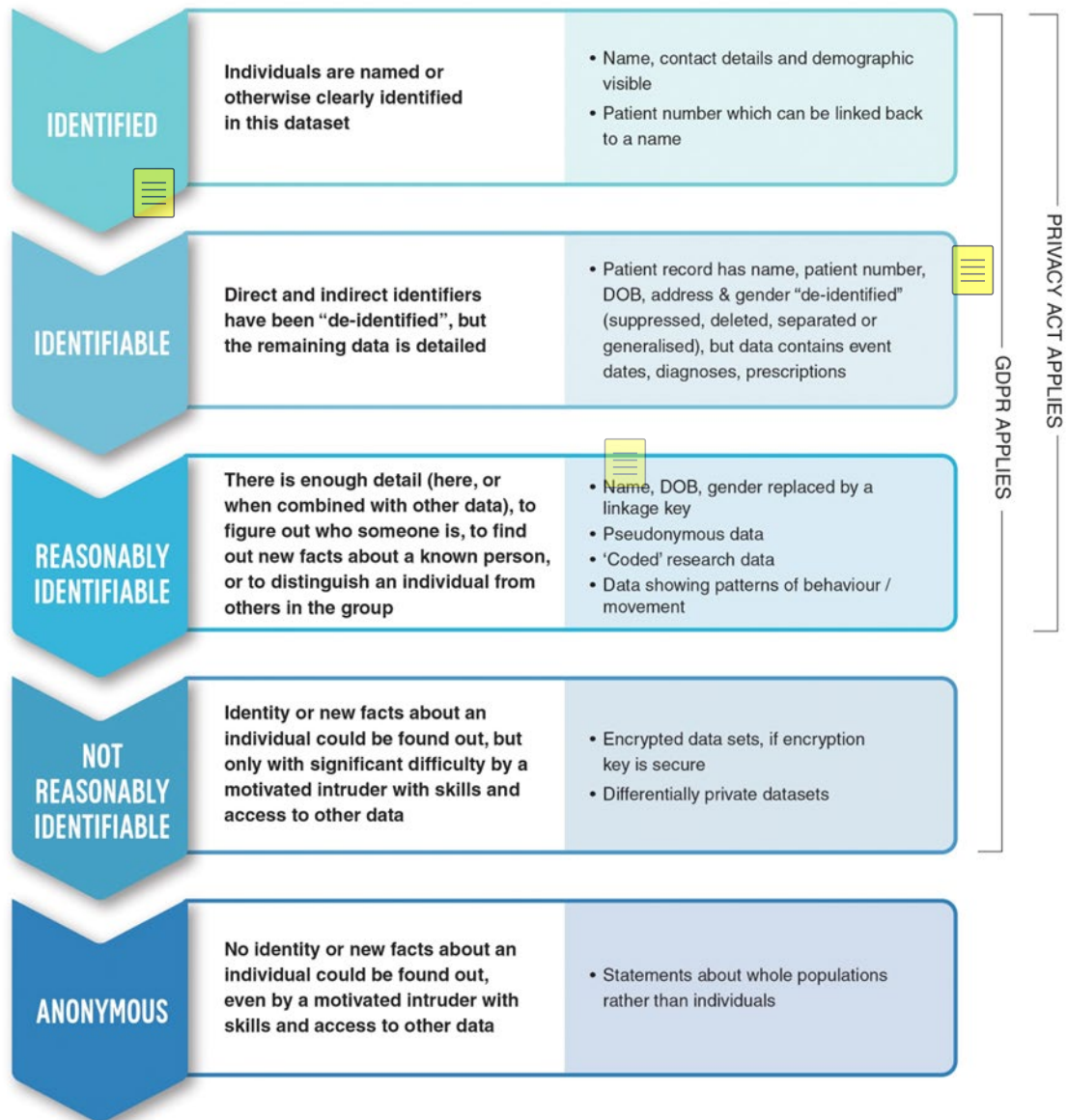
There are different types of synthetic data as well as different methods that can be used to create it. This framework is focused on "Tabular" synthetic health data, as opposed to other forms of health data such as synthetic images for computer vision, synthetic text for Natural Language Processing (NLP) or synthetic time-series data. In terms of generation methods, this framework does not consider all methods. It is not intended to cover rule-based generators, simulators, or domain-expert-driven approaches.

Not all synthetic data carries material privacy risk. The level of privacy risk will depend on both the type of source data used to generate synthetic data (which informs the 'inherent' privacy / disclosure risk) and the extent to which the synthetic data reflects the original source data (the 'residual' privacy / disclosure risk).

⁵ Office of the Privacy Commissioner of Canada, "When what is old is new again – The reality of synthetic data", OPC blog post, 12 October 2022; available at <https://priv.gc.ca/en/blog/20221012>

⁷ From the IAPP: <https://iapp.org/resources/article/key-terms-for-ai-governance/>

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



Source: Helios Salinger⁸

Where a synthetic 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 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'.

⁸ 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 modeling. 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	Yes. Inputs have high inherent privacy risk.

The challenge when creating synthetic data is how to make it appear real, or 'real enough', to support the relevant use case meanwhile to protect real patients' privacy. Whilst balancing the trade-off between fidelity, utility and privacy, the privacy protection should be the first principle for synthetic 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 data and the synthetic data itself, and reduce the level of privacy risk to an acceptable level to support the use case at hand.

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

