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:

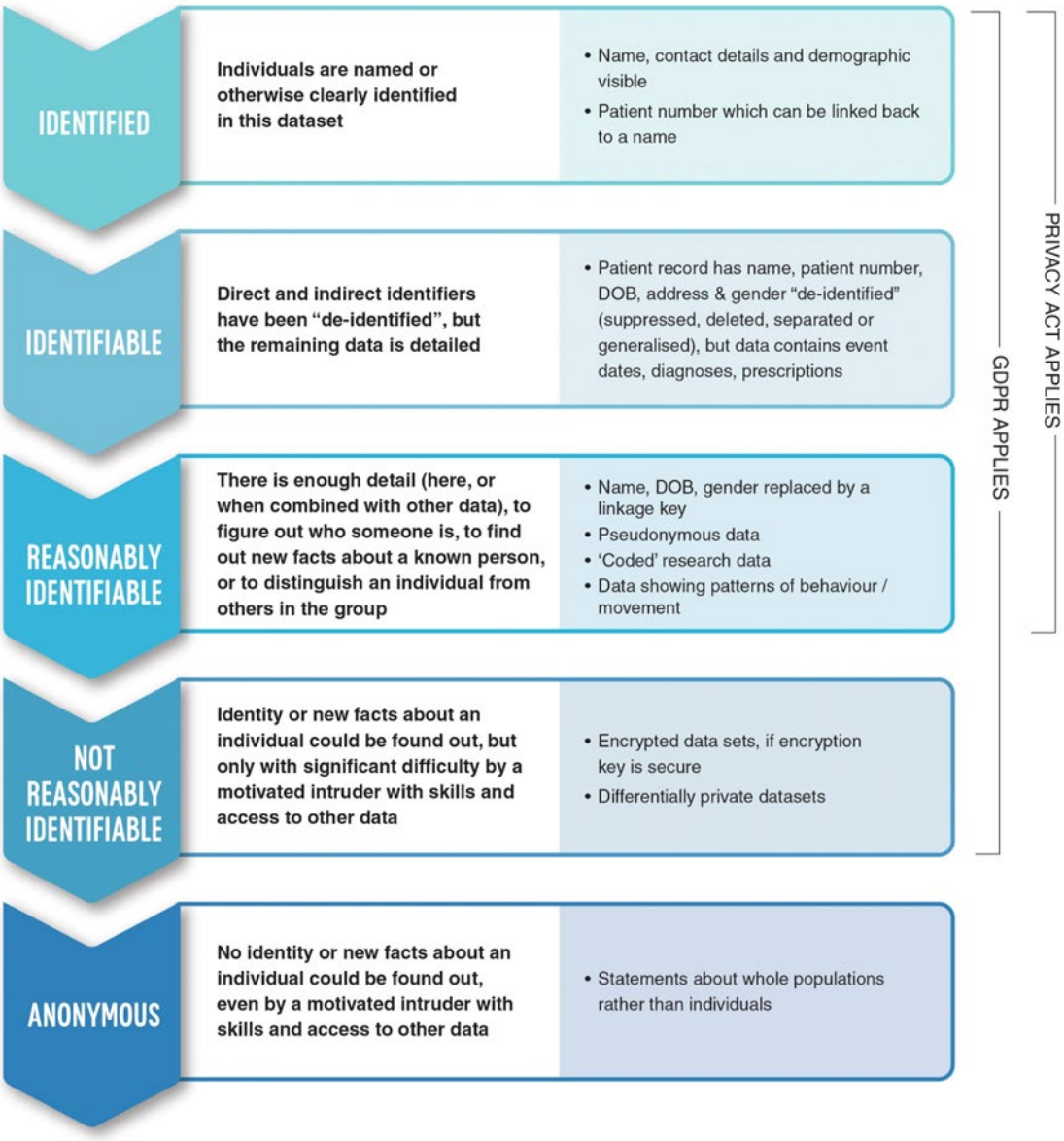
> "Synthetic data is 'artificial' data generated by data synthesis algorithms, which replicate patterns and the statistical properties of real data … when you analyse the synthetic data, the analysis should produce very similar results to analysis carried out on the original real data".[3]

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



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).

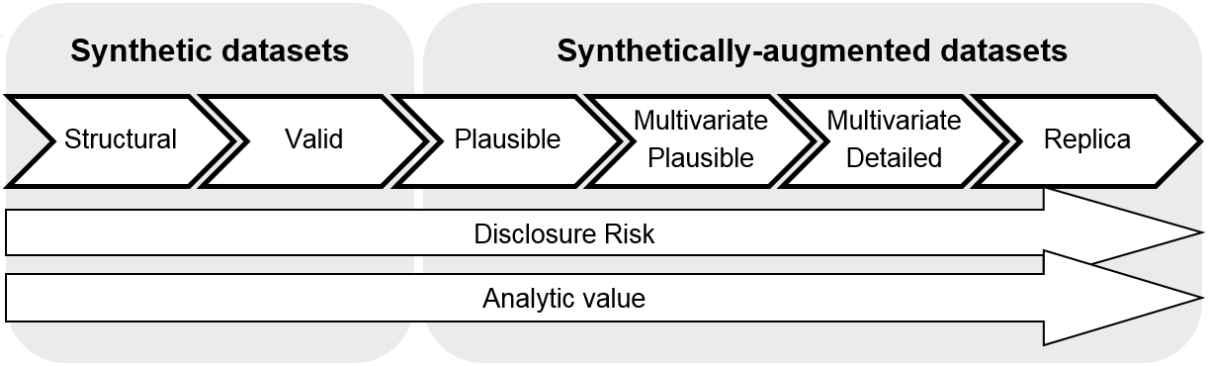
Altering source data will also impact data accuracy. While synthetic health data aims to replicate the structure and patterns of the source data, it does not replicate the exact distribution of the source data. Where a use case requires highly accurate data for analysis and decision-making, synthetic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

About the Synthetic Health Data Governance Framework

Scope

The Framework is designed to support the objectives of data custodians, health organisations, researchers, health system consumers and other stakeholders by providing a structured, practical and risk-based approach for the safe, effective and lawful creation and use of synthetic health data, in relation to both anticipated and future use cases.

Where organisations have their own internal data governance policies and frameworks that apply to health data, this Framework is intended to strengthen these existing policies and frameworks, and not replace them. This Framework is intended only to apply to the creation, use and handling of synthetic health data. It is not intended to cover or provide guidance for a wider range of approaches to data de-identification (such as redaction, data masking, light perturbation).

This Framework explains the steps and assessments organisations need to carry out when seeking to generate and use synthetic health data, so that the benefits of synthetic health data can be realised while ensuring the associated privacy risks are identified and managed.

To ensure that any synthetic health data project will be lawful, appropriate, ethical and safe, **all steps and assessments in this Framework must be completed before access to synthetic health data can be granted.**

Creating, using and sharing synthetic health data can raise questions not only about whether these activities are carried out lawfully, but also about the reliability of the synthetic health data, the ethics of the use case, and how to protect the data when being used or shared. Any creation, use and sharing of synthetic health data must be lawful. The laws and policies that underpin this Framework are explained further in Appendix 3.

This Framework outlines a five-step approach that organisations must complete before access to synthetic health data can be granted:

1. Assess the use case | **2. Assess and prepare source data** | **3. Generate synthetic data** | **4. Assess re-identification risks** | **5. Take steps to manage residual privacy risks**

Steps required under this Framework	Assessments and guidance
Documenting each step *This Framework includes a form to document the outcomes of the assessments required under this Framework and to assist accountable decision-makers to gather the information they need to consider when approving synthetic health data requests*	Request and Assessment Outcomes Form (Appendix 11)
Step 1: Assess the use case *This step uses a risk-based approach for Data Custodians approving the creation of synthetic health data and / or synthetic health data use cases based on legal and ethical considerations, and explains different lawful pathways for progressing high risk and / or complex synthetic health data requests*	Assessments to be completed: ☐ Use Case Assessment (Appendix 4) ☐ Impact Assessment (Appendix 5) Guidance to assist decision-makers: • Decision tree for complex synthetic health data requests (Appendix 8) • Further guidance on different lawful pathways (Appendix 9) • The policy and legal framework underpinning this Framework (Appendix 3)
Step 2: Assess and prepare the source data This step helps *Data Custodians answer the question – 'is this data fit for purpose?'*	Assessments to be completed: ☐ Technical Assessment (Appendix 6)
Step 3: Generate the synthetic health data *A range of synthetic health data generation methods may be suitable under this Framework. Data Custodians, with support from Data Scientists and synthetic health data experts, will need to determine a suitable approach that produces a synthetic health dataset with an appropriate balance of utility and accuracy for each use case.*	
Step 4: Assess and manage re-identification risks *Assessing and managing re-identification risks are critical to ensuring legal privacy compliance. This step is supported by an explanation of different de-identification techniques and ensures organisations effectively manage their privacy risks when*	*Assessments to be completed:* ☐ *[If the SynD orgs wish to adopt a unified re-identification risk assessment, reference here and include in an appendix – see Step 4 for more detail]* ☐ *[If the SynD orgs wish to adopt a unified data utility assessment, reference here*

proceeding with synthetic health data requests.	*and include in an appendix – see Step 4 for more detail]* Guidance to assist decision-makers: • De-identification techniques (Appendix 7) • Decision tree for complex synthetic health data requests (Appendix 8) • Further guidance on different lawful pathways (Appendix 9)
Step 5: Manage residual privacy risks *This step explains the 'how' of using and sharing synthetic health data safely.*	Assessments to be completed: ☐ Safety assessment (Appendix 6)

Only if each of the five steps has been completed successfully, and a safety assessment has been completed, can organisations proceed with sharing synthetic health data.

These steps should be seen as 'cascading', in the sense that they should be completed successively – because if an early step cannot be completed, the synthetic health data request cannot proceed under this Framework.

Audience

This Framework is primarily designed for individuals who have assigned roles for making, or supporting, decisions about data. This includes Data Sponsors, Data Custodians, Data Stewards and Data Scientists or who otherwise fulfil or support the 'accountable decision-maker' role in relation to synthetic health data projects. (See the Glossary at Appendix 2 for the meaning of these terms). It could also include members of data governance committees, research governance committees, ethics committee members and those in Quality and Safety roles. These individuals will have responsibility for decisions about synthetic health data, and/or generating, assessing and protecting synthetic health data.

It is also designed for 'users' of synthetic health data, such as health researchers, students and health data analysts, and any staff involved in handling and protecting synthetic health data.

Roles and responsibilities

Different types of synthetic health data requests will be handled, assessed and ultimately approved (or refused) by people with specific roles in the organisations that are responsible for the data. These include:

Entity	Role
The organisation that holds and controls the source data (= "Data Provider")	**The accountable decision-maker**. Depending on an organisation's own data governance framework, this will usually be the Data Owner or Data Custodian, a role that is typically accountable for, or who 'owns', a particular dataset, and who has the authority to approve certain uses and disclosures of the data. The Data Custodian will typically be supported by Data Stewards, Data Scientists and other relevant stakeholders who can help assess and prepare source data, generate synthetic health data, test for re-identification risk, and ensure data security requirements are met. The Data Custodian may need to consult with the Data Requestor to clarify the purpose of the data request, to discuss whether synthetic health data is appropriate for the particular use case, as well as which method / model will be used to generate the synthetic health data. For synthetic health data requests that are complex or high risk, the Data Custodian may need to engage internal or external expertise with respect to privacy compliance and risk management. The Data Custodian may also need to engage external expertise to test for re-identification risk.

	Where a Data Custodian's organisation does not have the necessary skills or expertise to generate synthetic data, the organisation may need to engage with external suppliers with data science expertise to assist with synthetic data generation. Where this is the case, organisations should ensure robust procurement, contractual and oversight mechanisms are in place to protect the data from any unauthorised use or unauthorised disclosure. Where the re-identification risk cannot be lowered to an acceptable level (based on both the disclosure risk from the data itself, as well as the surrounding controls used to protect the data from re-identification), the Data Custodian may need to seek separate approval to proceed, in accordance with legal and/or risk frameworks. This will also be the case if the Data Custodian intends to share real data with the Data Requestor organisation in order for the Data Requestor to generate synthetic health data. In these cases, steps could involve engaging with privacy experts, ethics committees and internal data governance committees.
The organisation requesting to access or receive synthetic health data generated from the source data (= "Data Requestor")	**The responsible data user**. This will usually be the organisation hosting the research or project lead who either requests synthetic health data from the Data Provider for a specified use case, or in some cases requests real data from the Data Provider with the intention of using it to generate synthetic health data. The Data Requestor will be able to explain the use case for the synthetic health data being requested and will be responsible for ensuring it is being used only for approved purposes. The Data Requestor will consult with the Data Provider to determine if an appropriate synthetic health dataset can be generated, re-used or re-purposed, and what attributes are required. The Data Requestor is responsible for ensuring the synthetic health data is processed and handled in a secure manner, and that the integrity of the dataset is maintained. Where synthetic health data is *transferred* by a Data Provider to the Data Requestor (as opposed to the Data Provider *providing access to* a synthetic health dataset), the Data Requestor who receives the data is responsible for its secure storage and handling. Data security measures for storing and accessing synthetic health data are discussed further in this Framework, and will depend on the level of re-identification risk associated with the synthetic health data and the particular use case. Generally, data security measures should be commensurate with residual privacy and re-identification risks.

End Users	**End Users** are individuals such as data analysts or researchers who will access and use synthetic health data for analysis and insights generation.

Organisations under this Framework may at different times be either the Data Provider or the Data Requestor, depending on the use case.

Broader collaboration for better governance

Requests for synthetic health data will require collaboration between organisations that both hold relevant real data needed to generate synthetic health data, and those who wish to access and use synthetic health data. There may also need to be broader collaboration with other organisations that have synthetic health data expertise but are not otherwise involved in handling data within scope of the synthetic health data request. In practice, collaboration and consultation may be driven by specific teams within organisations, such as project teams or committees.

The organisations should expect there will be discussions around feasibility and availability of data, organisational constraints which may impact generating and sharing synthetic health data, and opportunities to refine what is needed data-wise to support a particular project prior to an organisation receiving a request for synthetic health data. In line with this Framework, organisations should be supportive of synthetic health data generation and use, given the range of benefits to multiple stakeholders and the privacy protective nature of synthetic health data compared with real data. Organisations should also develop supporting synthetic data governance and assurance processes to help streamline requests.

Organisations that are requesting and providing data may also need to liaise with each other as part of their own assessment process. For example, the Data Requestor may need to assist the Data Provider where the Data Requestor has (or is proposing to seek) HREC approval under a research exception pathway because there are material privacy risks in a synthetic health dataset that cannot be further de-identified (that is, the research proposal cannot be considered 'low risk' and so cannot follow 'lower risk research' ethics review pathways[10]).

While the assessments in this Framework must be followed and documented for each synthetic health data request, other assessments (such as Privacy Impact Assessments, Security Assessments, and AI Impact Assessments) beyond this Framework may be required in connection with synthetic health data projects, as required under an organisation's own frameworks and policies.[11] Where this is the case, both the information

[10] 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

[11] In some cases, certain assessments may also be a *legal* requirement. For example, Australian Government agencies are required by law to conduct a Privacy Impact Assessment on any high risk project; see the Australian Government Agencies Code, made under the *Privacy Act 1988* (Cth)

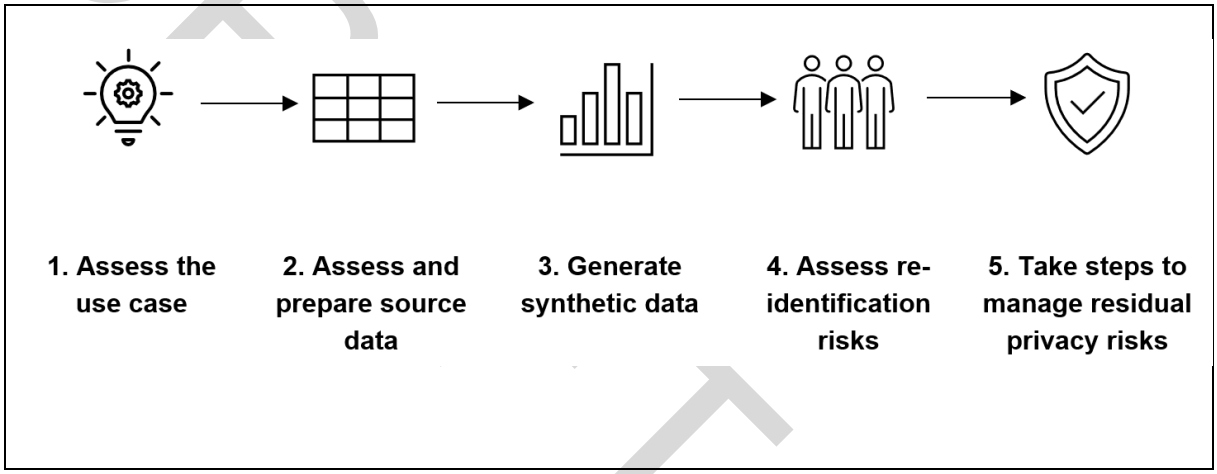
gathered and the assessments completed under this Framework will provide valuable inputs for these activities.

The Framework

Steps for generating and accessing synthetic health data

This Framework sets out the key steps and assessments that must be made in connection with synthetic health data requests.

At a high level, processing synthetic health data requests will involve 5 key steps. Each step must be successfully completed before moving on to the next.[12] Depending on the outcome of each step, some steps may need to be iterated before a request can progress.



Step 1: Assess the use case

Synthetic health data will not be suitable for all use cases. For example, synthetic health data will not be suitable for analysis that leads directly to clinical decisions impacting individuals, or for use cases where a high degree of individual data accuracy is required.

The creation of synthetic health data will often involve the use of a 'real' dataset as its starting point. If that source dataset contains personal information – i.e. potentially identifiable information about individual humans – any use of that dataset for a particular purpose, including using it to create a synthetic health dataset, will need to comply with the 'Use' principle/s in the applicable privacy law. This is discussed further in Appendix 9.

[12] This five-step approach is based on the Personal Data Protection Commission Singapore's *Privacy Enhancing Technology (PET): Proposed Guide On Synthetic health data Generation*, July 2024, available at: https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/other-guides/proposed-guide-on-synthetic-data-generation.pdf

In cases where synthetic health data will be generated from real health data, it is important that, when assessing a proposed use case, organisations are guided by the original purpose for which the individuals represented by the data originally provided their information.

Individuals interacting with the health system (i.e. health consumers) typically do so to access health care services relevant to their needs. This could include, for example, accessing care related to illnesses, injuries, disabilities or health conditions, or for care related to maintaining or improving their health. Organisations that provide health care services to individuals collect their health information for the *primary purpose* of delivering this care. Under privacy law, these organisations are only permitted to use and disclose health information for the primary purpose of collection, or for limited secondary purposes in certain circumstances. (See Appendices 3 and 9 for a further discussion of privacy law).

Using health data to generate synthetic health data is a secondary use. Unless an exception applies, the use case for the synthetic health data request must be one that is *directly related* to the primary purpose of collection, and individuals should *reasonably expect* that their health information will be used for this purpose.

In order to meet this test, use cases will be suitable for synthetic health data requests under this Framework if all three tests below can be met:

1. The use case is for a **clear 'public benefit' purpose** related to providing health services, and where the expected benefits from the use case are related to consumer health or health system outcomes. This requirement will help constrain use cases to those that are aligned with the primary purpose of the original collection.
2. The stated aim for creating and managing the synthetic health dataset is to achieve a **'de-identified' dataset for the use case, which significantly minimises the risk to individuals** compared to if the source dataset were used for that use case. Organisations agree that a synthetic health dataset with only a very low risk of re-identification will be suitable for use cases to proceed under this Framework.
3. The organisation that collected and holds the source data has set expectations with health consumers about how their health information will be used. This means there should be transparency and **public communication about using synthetic health data** for public benefit projects before organisations facilitate synthetic health data requests under this Framework. Organisations are not required to obtain consent from individuals or to provide them with individual notices. However, these communications should be co-designed with health consumers and communities, and created with a range of health consumers as the intended audience as a way to build and maintain a social licence for synthetic health data generation and use. Organisations could also conduct social research with health consumers to better understand current community expectations when it comes to using real health data to generate synthetic health data for a range of different use cases, and to determine the level and type of communication that

may be required to provide transparency and inform expectations about this use of health data.

Communication strategies could include a communication on the organisation's website, posters in areas where consumers are likely to see them, information included (or linked to) in an organisation's published privacy policy, uplifting communications that already speak to health research initiatives, or via another channel that is deemed suitable by the organisation to help educate, inform, provide transparency and maintain trust.

Under this Framework, a use case will not be suitable to proceed if it does not meet these tests and the request should be considered 'complex'. The Data Provider (i.e. the entity that holds and controls the source data) ultimately bears the legal risk of using source data to generate synthetic health data for sharing. The Data Custodian at the Data Provider should be comfortable that *using real data* to create the synthetic health dataset is ultimately for a purpose that is 'directly related' to the primary purpose of collection and would be 'reasonably expected' by the individuals who are represented in the source data. If the Data Custodian is not satisfied, the synthetic health data request can proceed on this basis, an alternative lawful privacy pathway must be determined.

Use cases may also have mixed benefits or purposes that will need to be assessed under this test. For example, a Data Requestor may request synthetic health data for the purposes of developing and training an AI-based clinical diagnostic support tool to be sold and distributed by a medical technology company. While there may be a clear 'public benefit' purpose to the use case (e.g. faster and more accurate medical diagnoses for health consumers), there is also potentially a material commercial benefit to a private company. Given the privacy and potential social licence impacts associated with this type of 'mixed benefit' use case, the use case should be considered 'complex' and be subject to review by an HREC before proceeding to carefully consider the ethical implications and to ultimately approve the use of synthetic health data (*even* if the synthetic health data would otherwise have a very low re-identification risk).

See *Appendices 8 and 9* for guidance on dealing with complex synthetic health data requests.

The Data Provider should use the Use Case Assessment Checklist in Appendix 4 to determine whether a proposed use case is acceptable.

After determining whether a proposed synthetic health data use case can proceed to the next steps under this Framework, the Data Provider should then ask, *should we proceed with the request and share this data?* Each request to generate and share synthetic health data needs to be considered in terms of whether the other risks involved in providing the data can be adequately managed and whether it is ethical to proceed with the request.

The Data Provider should use the Impact Assessment Checklist in Appendix 5 to answer the question: *should* we generate and share the synthetic health data?

Step 2: Assess and prepare the source data

Once a use case is considered suitable to proceed under this Framework, the next step is to determine whether there are any conditions or restrictions on the source data's use, and whether the source data available is fit for purpose.

Key questions to be considered include:

- What insights need to be generated from the synthetic health data?
- What data needs to be included to satisfy the use case at hand?

Limits or restrictions on using source data
The Data Custodian must assess if there are any conditions on the source dataset that limit its use (and whether the creation of synthetic health data is not already included within those conditions). As an example, use of the NSW Lumos Data Asset is constrained by the terms of an HREC approval, which reflects promises made to the original data owners (e.g. General Practices). There may also be limits or restrictions that apply to certain datasets as a result of contractual agreements with original data owners. In these circumstances, the Data Custodian will need to examine whether a new use case (in this case, creating a synthetic health dataset) is permitted under those limitations or restrictions, or if further approvals need to be obtained in order to proceed.

Data linkage
If preparing the source data involves data linkage (i.e. combining data from different sources) prior to generating the synthetic health dataset, the Data Custodian must consider if there are limitations or restrictions on all data sources that are intended to be linked.

If data required for linkage prior to generating the synthetic health dataset needs to be disclosed by one organisation (e.g. a health department in one state) to another organisation (e.g. a health department in another state), both organisations must ensure that both the disclosure and the subsequent collection and use of the source data are lawful. Given the heightened privacy risks associated with data linkage activities involving data sharing between multiple organisations, a Privacy Impact Assessment (PIA) should first be completed to assess whether the data flows required in the circumstances will be lawful.

Is the source data fit for purpose?
Before an accountable decision-maker commits to using source data to generate synthetic health data, they should first confirm that the appropriate data is available and is of sufficient quality. The Data Custodian will also need to consider what is the minimum amount of data needed to generate a synthetic health dataset that will be suitable for the use case at hand. For example, a use case may relate to a particular disease or focus on a particular time period, which does not require the full source dataset. If appropriate, the Data Custodian should prepare a subset of the source data that includes only those aspects, attributes and / or fields needed for the use case. The Data Custodian and the Data Requestor will also need

to consider whether a synthetic health dataset is appropriate for each use case in the circumstances.

Data and fields containing **directly identifying information** (such as names, addresses, phone numbers, date of birth, date of death, unique identifiers such as patient numbers, Medicare numbers or drivers licence numbers) must also be removed or otherwise treated with appropriate, effective and tested de-identification techniques to reduce the risk they will be 'leaked' via the synthetic health dataset (if they haven't already been removed or treated).

Assessing the source data
To assist with assessing the fitness for purpose of the source data, the accountable decision-maker at the Data Provider **must complete the Technical Assessment Checklist** in this Framework at Appendix 6 and be satisfied that the source data being requested is fit for purpose before progressing the synthetic health data request to the next steps.

Step 3: Generate the synthetic health data

There are various methods for generating synthetic health data. The elements of each generative model should be considered when determining which one is most appropriate for a particular use case.

Organisations generating synthetic health data will need individuals with the necessary expertise to carry out the synthesis, such as data scientists. If third-party expertise is required to generate the synthetic health data (which may include the source data being transferred off-premises), organisations will need to put in place appropriate due diligence / vetting processes, security controls, contractual protections and oversight (organisations should rely on their organisational policies and procedures to assist with these tasks, which could include the need to complete a Privacy Impact Assessment).

When creating a synthetic health dataset, the Data Provider with the Data Requestor will need to consider the desired level of analytical value and preservation of relationships between variables that need to be retained in the dataset. The dataset will need to be representative enough for the use case, while also keeping statistical disclosure risk to a minimum. The Data Provider and the Data Requestor should define the key statistics that must be preserved for the use case. These statistics will need to be taken into account when generating the synthetic health data, while also aiming to keep statistical disclosure risk to a minimum.

Once generated, the Data Provider should check the synthetic health dataset and validate that it meets the expected parameters and the model has worked correctly. Organisations may wish to create multiple versions of the synthetic health dataset and average the conclusion based on the results from the different versions.

The organisation should ensure it has documented the model that was trained and used to generate the synthetic health data. The model must be stored securely and separately from the data or otherwise destroyed if it is no longer needed. If a model is to be reused or modified for other use cases, it should only be accessed by authorised personnel. Access to the model must be controlled, monitored and logged to reduce the risk of model leakage.

As a general rule, the model should not be provided to the Data Requestor or an End User who either has access to or will receive the synthetic health dataset. If a Data Requestor or an End User wishes to access the model, it must be for a purpose that is acceptable to the Data Provider, and steps should be taken to reduce the risk that the Data Requestor or End User could use the model to potentially rebuild the original source dataset (or aspects of the dataset).

Users who have access to the synthetic health dataset for analysis must not be able to access the source dataset or a related source dataset unless they are doing so for an approved purpose.

Step 4: Assess and manage re-identification risks

The UK Information Commissioner's Office has noted that there is no standard available as to how synthetic health data should be generated, and warns:

> "Synthetic health data may not represent outliers present in the original personal data. You will need to assess whether the personal data on which the synthetic health data was trained can be reconstructed. Further additional measures (e.g. Differential Privacy) may be required to protect against singling out".[13]

('Singling out' is a phrase in UK/European data protection law to mean that an individual may be distinguished from the group, and thus 'identifiable' for the purposes of the definition of 'personal data', or 'personal information' as it is known for the purposes of privacy law in Australia.)

Thus again, unless synthetic health data is created completely from scratch or in a manner which does not involve real data about individuals, the way in which it is created could lead to some re-identification risks being carried over from the source dataset.

> "If a synthetic health dataset preserves the characteristics of the original data with high accuracy, and hence retains data utility for the use cases it is advertised for, it

[13] UK Information Commissioner's Office, "Chapter 5: Privacy-enhancing technologies (PETs) – Draft", September 2022, p.38; available from https://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/ico-call-for-views-anonymisation-pseudonymisation-and-privacy-enhancing-technologies-guidance/

simultaneously enables adversaries to extract sensitive information about individuals."[14]

Risks that personal information may be ascertained or disclosed from a dataset could include:

Identity disclosure
Identity disclosure occurs when data is re-identified and a person's identity can be assigned to a record. Identity disclosure can arise by one of two ways: by either matching a person to data (such as taking an individual, and finding data that matches them), or matching data to a person (such as starting with the data and finding the individual to whom that data relates).

Attribute disclosure
Attribute disclosure occurs when new facts can be learned or inferred about an individual from the dataset.

Membership disclosure
Membership disclosure occurs if it can be determined if an individual's data was in the source dataset that was used to generate the synthetic health dataset.

Because 'personal information' (as defined in privacy law) includes information or opinion regardless of *whether it is true or not*, even disclosures that are inaccurate or incorrect will risk breaching an organisation's privacy obligations.

If a re-identification attack were successful, the re-identification of consumers and resultant risk of unauthorised disclosure of personal information from the synthetic health dataset would pose a legal compliance and reputational risk.

Once a synthetic health dataset has been created, there will be additional legal compliance issues if the data in the synthetic health dataset could contain 'personal information'. Following the creation of the synthetic health dataset, the Data Provider should therefore take additional steps to reduce the risk of re-identification or disclosure of personal information. This will likely involve post-generation review and modification activities carried out by data scientists or those with similar expertise. The dataset may need to be further modified in order to meet certain criteria designed to reduce re-identification risk. **Common techniques to reduce re-identification risks are described in Appendix 7. These techniques can help support synthetic health data use and ensure a higher level of privacy protection.**

After applying additional de-identification techniques, the overall re-identification risk level of the synthetic health dataset must be considered and tested via a robust Re-identification Risk Assessment.

[14] See Theresa Stadler, Bristena Oprisenu and Carmela Troncoso, "Synthetic health data – Anonymisation Groundhog Day," v6, 24 January 2022; available at https://arxiv.org/abs/2011.07018 Note however that the word 'utility' can also have a more specific meaning.

Only if the results of the Re-identification Risk Assessment indicate that the re-identification risk is 'very low' can the synthetic health data request proceed to Step 5.

Re-identification Risk Assessment Methodologies

Privacy protection is the first principle guiding synthetic data generation, even though synthetic data ultimately represents a trade-off between fidelity, utility, and privacy. Privacy evaluation is therefore a critical yet often misunderstood component of synthetic data governance. While synthetic data aims to protect privacy by generating artificial records without direct links to real individuals, residual privacy risks persist. Some generative models incorporate privacy-preserving mechanisms by design (for example, DPGAN and PATEGAN implement differential privacy, while ADSGAN targets re-identification risks [1-3]), yet issues such as model overfitting and the inadvertent retention of sensitive patterns can compromise privacy. Consequently, robust privacy assessment remains essential, as risks such as membership and attribute inference may arise when synthetic data preserves statistical patterns from the original dataset. **Appendix 7** provides guidelines for best practice for assessing re-identification risk for synthetic data.

Steps following the Re-identification Risk Assessment

The results of the Re-identification Risk Assessment will determine next steps for the use case.

A 'very low' risk of re-identification means that even though it may be technically possible to re-identify an individual from the information, doing so is so impractical that there is almost no likelihood of it occurring.[15] The OAIC advises:

> "As part of assessing the likelihood of identification, entities should also consider whether an entity (or a particular person) may be especially motivated to attempt to identify someone".[16]

If the results of the Re-identification Risk Assessment indicate there is a *more than a very low risk of re-identification*, there are two options for next steps:

- In consultation with the Data Requestor, apply additional de-identification techniques until the re-identification risk has been lowered to a very low level and the synthetic health data request can proceed to Step 5. (This outcome should be supported by completing another Re-identification Risk Assessment.)
- If the re-identification risk cannot be lowered to a very low level, the synthetic health dataset must be considered 'personal information', and privacy law

[15] This is the standard of de-identification used by the OAIC for information to no longer be regarded as 'personal information' for the purposes of the Privacy Act. See: Office of the Australian Information Commissioner, *What is personal information?*, May 2017, Available at https://www.oaic.gov.au/agencies-and-organisations/guides/what-is-personal-information
[16] Office of the Australian Information Commissioner, *What is personal information?*, May 2017, Available at https://www.oaic.gov.au/agencies-and-organisations/guides/what-is-personal-information

obligations will continue to apply to the way it is handled. This means the Data Custodian cannot use or share the synthetic health dataset further until a lawful pathway has been determined. This may involve needing to seek a waiver of consent from an HREC.

See the decision tree and how it relates to different scenarios involving health data in Appendix 8. See Appendix 9 for an explanation of the lawful privacy pathways for secondary uses and disclosures of health data.

Summary of next steps following completion of a Re-identification Risk Assessment

Re-identification Risk Assessment Results	**Privacy risk**	**Next steps**
'Very low' risk of re-identification	Low	Privacy risks associated with the data have been materially reduced. Proceed to Step 5. If the use case is for research, consider if the research is otherwise eligible to follow 'lower risk research' ethics review pathways.[17]
More than a very low risk of re-identification, but there are options to further reduce the risk	Medium	Data Provider to consult with Data Requestor and apply further de-identification techniques and privacy controls. Re-do the Re-identification Risk Assessment to determine results and associated privacy risk.
More than a very low risk of re-identification, risks cannot be further reduced	High	The dataset is 'personal information' and privacy law applies. Treat use case as a 'complex request' and seek alternative lawful pathway to support the use case; most likely seek a waiver of consent from an HREC in order to proceed.

Organisations should also be aware that re-identification risk can change over time. Factors impacting re-identification risk should be monitored, as Re-identification Risk Assessments may need to be refreshed over time to ensure organisations can identify any changes in the risk level and can manage their obligations accordingly. The OAIC has advised:

> "The feasibility of a particular method of identifying an individual can change with new developments in technology and security, or changes to the public availability of certain records. If an entity has decided that the information it holds does not allow the identification of individuals, that decision should be reviewed regularly in light of any such developments."[18]

[17] 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

[18] OAIC guidance, What is personal information?, available at: https://www.oaic.gov.au/privacy/privacy-guidance-for-organisations-and-government-agencies/handling-personal-information/what-is-personal-information

Factors which may impact re-identification risk could include:

- The time that has passed since the Re-identification Risk Assessment was completed
- How any outputs from the use case will be handled (e.g. whether outputs will be published or shared outside of the Data Requestor, and in what form)
- Any changes to the way the synthetic health dataset will be shared with the Data Requestor (e.g. if the full synthetic health dataset will be transferred to the Data Requestor, as opposed to the Data Provider granting access to approved End Users)
- If there have been any data breaches or security incidents involving the source data
- If there have been any data breaches or security incidents involving data *related* to either the source data or the synthetic health dataset (even if the data was leaked from a different organisation), this could increase the risk of re-identification. For example, if health data held by another organisation (but which may still reasonably relate to individuals represented in the synthetic health dataset) has been exposed on the dark web.
- If there have been any privacy incidents or breaches involving the Data Requestor that may impact their data security posture
- Whether there have been any technological or security developments that may impact re-identification risk

Data Utility Assessment

In addition to assessing the synthetic health dataset for re-identification risk, the dataset should also be assessed to measure data utility and data fidelity to ensure it is suitable for the use case at hand.

[Placeholder for SynD members: adding a data utility / fidelity assessment was suggested earlier as feedback. Similar to the re-identification risk assessment, the SynD community may wish to describe acceptable data utility assessment methodologies, or alternatively may wish to wish to settle on an agreed approach to assessing data utility / fidelity and link to it here in the Framework.]

Depending on the outcome of the re-identification risk assessment and the data utility / fidelity assessment (for example, if desired levels have not been achieved in the synthetic health dataset) data custodians may need to iterate steps 3 and 4 until requirements are met.

If a 'very low' level of re-identification risk cannot be achieved in order to maintain the necessary level of data utility required for the use case at hand, the request should be considered 'complex' and the synthetic health dataset must be considered 'personal information'. Privacy law obligations will continue to apply to the way it is handled (see above for options where this is the case).

Step 5: Manage residual privacy risks

Once the accountable decision-maker is satisfied that the synthetic health dataset is sufficiently de-identified to be shared with the Data Requestor and End Users, they must answer the final question: *How* do we share this data - *safely*?

Each synthetic health data request now needs to be considered in terms of ensuring a safe sharing and storage environment.

The accountable decision-maker must only approve sharing the synthetic health dataset once satisfied that it is safe to do so.

See Appendix 10 for more information about safe sharing, including the Five Safes Framework, a mandatory Safety Assessment Checklist, information about Data Sharing and Data Use Agreements, and links to further resources.

Final steps

Responsibility for approving the creation, sharing and use of a synthetic health dataset ultimately sits with the accountable decision-maker at the Data Provider. **It is the responsibility of the accountable decision-maker or their delegate to ensure that the steps and assessments set out under this Framework have been completed, prior to issuing their approval for a synthetic health dataset to be created and shared with the Data Requestor.** The Data Requestor must be willing to assist the Data Provider with information or action needed to facilitate the assessment and decision-making process.

If a request to create and share a synthetic health dataset is not approved, the Data Provider must provide reasons and further context where appropriate.

All synthetic health data requests and their outcomes must be documented. The Request and Assessment Outcomes form (attached at Appendix 11) should be used to document synthetic health data requests, assessment results and approvals.

Both the Data Requestor and the Data Provider will have responsibility for maintaining synthetic health data decision artefacts.

Relevant material would usually include:

- a copy of the request / data specification
- any consultation / meeting notes
- methodology notes, including documentation about the model trained and used to generate the synthetic health data and its parameters
- statement(s) of data quality

- any conditions the requester has been asked to meet
- for complex synthetic health data requests, documentation of the lawful privacy pathway to create and share the synthetic health data (see Appendix 8 for further guidance on complex requests)
- documentation of the accountable decision-maker's approval to create and share the synthetic health data
- any agreed modifications
- metadata describing the synthetic health data provided
- any Privacy Impact Assessment completed in connection with the synthetic health data request
- the Re-Identification Risk Assessment and Data Utility Assessment completed in connection with the synthetic health dataset
- any supporting Data Sharing Agreement and Data Use Agreement(s), as well as any other relevant agreements (such as any Memorandums of Understanding, Schedules, contract or confidentiality undertaking)
- where synthetic health data is not created or provided, the reason for that decision

Re-using, re-purposing or re-synthesising synthetic health datasets

Data Providers and Data Requestors may propose a use case where a synthetic health dataset already exists that would be suitable in the circumstances.

Data Requestors may also wish to use a synthetic health dataset already provided for a different or expanded use case, or for re-synthesis.

In these circumstances, the steps in this Framework should still be followed:

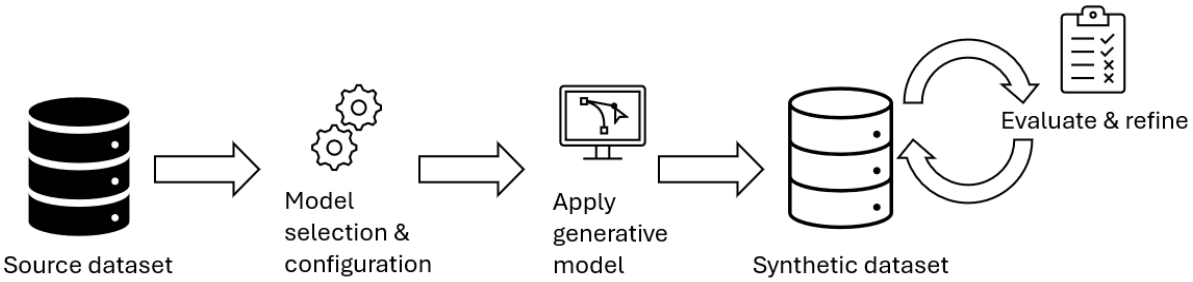
- The new use case must be assessed to determine if it is acceptable under this Framework (Step 1)
- The Data Provider must consider whether the data associated with the synthetic health dataset is fit for purpose in light of the use case (Step 2)
- Consideration should be given to whether the synthetic health dataset has the desired levels of data utility and fidelity for the use case (Step 3)
- The Data Provider must consider whether there are any internal or external factors that could impact the re-identification risk associated with the synthetic health dataset, which means a new Re-identification Risk Assessment must be completed (Step 4).
- The Data Provider and the Data Requestor must manage residual privacy risks and ensure the synthetic health data is protected (Step 5)

APPENDIX 1: About synthetic health data

To develop synthetic health data at scale, but without any one-to-one mapping, requires:

- A 'source dataset', containing 'real' records
- A generative model, and
- A new table or file, in which to house the synthetic records created.

These three elements are illustrated in the following diagram.[19]



These three elements can be further described as follows.

A source dataset

This is the original dataset containing records about a group of 'real' individuals, such as customers, patients or students. The synthetic health data to be generated will be expected to emulate the statistical properties of this source data.

Some source datasets will include data fields that clearly and directly identify individuals, from direct identifiers such as names, date of birth and unique numbers. Others might include indirect or 'quasi' identifiers such as age, gender and postcode, which in combination can render some individuals unique in the dataset (i.e. distinguishable from the rest of the group) and thus 'identifiable' in law.

Others again might have already robustly controlled for both direct and indirect identifiers via de-identification techniques, but the 'attribute' data is itself rich enough that some individuals will be unique in the dataset, and thus 'identifiable' in law. For example, even without any direct or indirect identifiers about a patient, the parts of a patient record that record event dates (such as doctor visit, hospital admission, surgery) and clinical information (such as conditions or treatment) can themselves render a patient unique in the dataset.

[19] Diagram adapted from UTS, "Synthetic Lumos - Technical Transfer", presentation slides, 20 December 2021

In all such cases, the source dataset contains 'personal information' as defined in privacy law.

To demonstrate using an example, in the source system of hospital admissions dataset, you might have 50% male and 50% female patients. You might also know from the source dataset that 5% of patients have had a caesarean section. If you break down that figure by gender, it represents 10% of female patients and 0% of male patients.

A generative model

This is the statistical model used to generate the synthetic health data. It is derived from the source dataset. Statistical tables are created from the source dataset to generate a probabilistic model.

> "A generative model is able to 'learn' the statistical properties of the source data without making strong assumptions about the underlying distributions of variables and correlations among them". [20]

The statistical properties of the source data might include, for example, the distribution of patients across gender, age ranges, geographic regions, as well as reason for hospital admission, but *without* necessarily correlating one data field (such as gender) to another (such as reason for hospital admission).

A generative model applied to the hospital admissions dataset which doesn't consider correlations will generate 50% of patients as male and 50% as female. Separately, it will generate 5% of all patients as admitted to hospital for a caesarean section.

The synthetic health data

This is the data generated from the generative model: many thousands of individual records of 'fake' individuals.

Because the data about individual synthetic 'patients' is generated by the generative model, unless correlations were designed into the generative model, attributes will be distributed across the synthetic patient records according to the statistical properties of the source dataset.

In the synthetic version of our hypothetical hospital admissions dataset, 5% of male patients will show a caesarean section as the reason for hospital admission, as will 5% of female patients.

This hypothetical scenario is intended to illustrate the 'privacy protection vs. data utility' trade-off that can occur when creating synthetic health data. Privacy protection is strengthened by not mapping correlations, however more simplistic representations of

[20] 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

populations in synthetic health datasets may then limit the quality or accuracy of results. In this scenario, if the correlation between “gender” and “reason for hospital admission” is not maintained, then the synthetic health dataset will indicate a percentage of male patients who have had caesarean sections. Whether this will impact a project’s outcomes will depend on the use case and the purpose for the analysis. For example, if the use case is to create dataset to train nursing students to use a patient management system, the synthetic health dataset may be suitable for this purpose. However, the synthetic health dataset would not be suitable for a use case that involves testing or researching a clinical hypothesis. The more that complex correlations are maintained from the source dataset (i.e. in order to make the synthetic health dataset ‘more real’), then the greater the likelihood of replicating unique patients from the source dataset, which leads to increased re-identification and privacy risk.

APPENDIX 2: Glossary

Accountable decision-maker	In the context of data under this Framework, this is usually the Data Sponsor (Executive Director level) or their delegate, or the Data Custodian. For complex data use and sharing proposals, it could be a Chief Executive or a Deputy Secretary.
Aggregated data	Aggregated data (as distinct from unit record data) is produced by grouping information into categories, typically with a combined count (i.e. numerical value) within each category.
API	Application Programming Interface
APPs	Australian Privacy Principles, found in the Privacy Act
Attribute disclosure	When new facts can be learned or inferred about an individual from a dataset.
Collection	A 'collection' of information occurs when the information comes into the possession or control of an organisation.
Confidentiality Undertaking	A Confidentiality Undertaking is a document containing a number of undertakings made by a data recipient pertaining to the handling of shared data. A Confidentiality Undertaking may be required to be executed by a Data Provider prior to data sharing (for example, if required by organisational data government frameworks and policies).
Data	Any facts, statistics, instructions, concepts or other information in a form that is capable of being communicated, analysed or processed (whether by an individual or by a computer or other automated means). For the purposes of this document, 'information' and 'data' are used interchangeably. Data may or may not include 'personal information' or 'health information, or other 'special category' information.
Data asset or dataset	A data asset or a dataset is a body of information or data, managed as a single unit, which is recognised as having value to the organisation and enables it to perform its business functions.
Data breach	If personal information has been lost, or accessed or disclosed without authority. A data breach will be 'notifiable' if the breach is

	likely to result in serious harm to one or more affected individuals.
Data Custodian	Makes decisions about the management of, access to and release of a data asset, including the definition of quality, and ensuring the asset is registered or catalogued.
Data fidelity	A measure of how accurate, complete, reliable and consistent data is in terms of representing the actual, real-world subject.
Data masking	The process of modifying, obscuring or replacing original data for security or confidentiality purposes.
Data Owner	The person or organisation responsible for the creation of the data, and who exercises authority over the data. The Data Owner may delegate or transfer certain aspects of its authority and its responsibilities to a Data Custodian, including via an agreement. For example, a general practice that collects patient data may (as the Data Owner) provide this data to another organisation, such as a state health department (as the Data Custodian) for specific purposes under an agreement.
Data Provider	The organisation which holds and controls the source health data that is the subject of a synthetic health data request. is disclosing data to one or more of the other organisations
Data Requestor	The organisation that is requesting the generation of synthetic health data from source data held by one of more of the other organisations
Data Sponsor	Undertakes data ownership on behalf of an organisation and ensures appropriate data governance policies are in place. The Data Sponsor may have the authority to approve data sharing.
Data Steward	Has day to day management of a data asset on behalf of the Data Sponsor, including ensuring that data quality and other standards are met. Provides support to Data Custodians and Data Sponsors.
Data utility	A measure of the value or 'usefulness' of data to achieve a goal or objective within a particular context.
De-identified data	'De-identified' data means that a person's identity is no longer apparent, or cannot be reasonably ascertained following the

	successful application of one or more de-identification techniques to 'personal information'[21]
Disclosure	The provision of personal information to another party outside an organisation
DSA	Data Sharing Agreement
DUA	Data Use Agreement
Dummy data	Sometimes described as a 'placeholder' or 'substitute' for real data. Dummy data will typically be fabricated to mimic the structure of real data for software or algorithmic testing purposes, but is non-meaningful and is not suitable for analysis.
Fake data	An umbrella term that means artificially generated data. Dummy data, mock data and synthetic health data can all be described as 'fake data'
Five Safes	A framework for considering how to control two types of privacy risks, when sharing data within a controlled setting
Health consumer	Individuals who use (or will use) health services, including their family and carers
Health information	Personal information that is about a person's: • physical or mental health • disability • current, past or future health services provided to them • wishes about future health services • actual or intended donation of body parts, organs or body substances • genetic information predictive of health • healthcare identifiers, and • all other information collected in the course of providing a health service.
HREC	Human Research Ethics Committee

[21] See the NSW Information & Privacy Commission Fact Sheet 'De-identification of personal information', 2020, https://www.ipc.nsw.gov.au/fact-sheet-de-identification-personal-information and the Office of the Australian Information Commissioner's guide 'De-identification and the Privacy Act', 2018, https://www.oaic.gov.au/privacy/privacy-guidance-for-organisations-and-government-agencies/handling-personal-information/de-identification-and-the-privacy-act

Identity disclosure	When data is re-identified and a person's identity can be assigned to a record. Identity disclosure can arise by one of two ways: by either matching a person to data (such as taking an individual, and finding data that matches them), or matching data to a person (such as starting with the data and finding the individual to whom that data relates).
Information	See 'data'
Insights	Information or details derived from data once it has been processed or analysed; the message conveyed by (or in reference to) data.
Membership disclosure	Membership disclosure occurs if it can be determined if an individual's data was in the source dataset that was used to generate a synthetic health dataset.
Mock data	Simulated or fictious data that is not derived or created from real data. It may be designed to replicate the structure and format of real data but does not contain or relate to real data records.
NHMRC	National Health and Medical Research Council
OAIC	Office of the Australian Information Commissioner
Output	The outcomes resulting from data use. For example, data analysis, results, insights, reports, or other information generated from the data.
Personal information	In the context of this Framework, personal information means any information about a person or that relates to a person who is at least reasonably identifiable. 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. Information does not have to be 'private' to be included in this definition. It can be true or false, an opinion or fact, recorded in a material form, or not recorded at all. Personal information can be about any living person, or a person who has died.[22] Personal information includes 'health information' and other special category data.
Perturbation	The process of modifying data for security or confidentiality purposes by making small changes intended to obscure original

[22] This meaning encompasses the elements of 'personal information' as defined in the range of privacy laws described in Appendix 3.

	values without impacting the overall statistical properties of the dataset (e.g. by adding 'noise' to the data).
PIA	Privacy Impact Assessment
Privacy Act	Privacy Act 1988 (Cth)
Real data	'Real world' data that relates to actual people, places, events, etc.
Redaction	The process of permanently removing or concealing data for security or confidentiality purposes.
Secondary purpose / secondary use	Using personal information for a purpose *other than* the primary purpose for which the information was originally collected
Sensitive information	Personal information relating to an individual's ethnic or racial origin, political opinions, religious or philosophical beliefs, trade union membership, sexual orientation activities, criminal record, health or genetic information, and some aspects of biometric information. Sensitive information is subject to additional legal privacy protections
Sharing	Data sharing involves data being provided from one organisation (the Data Provider) to another party at another organisation (known as the Data Requestor or End User)
Source data	The original data collected and held by the Data Provider from which a synthetic health dataset will be generated
Statistical disclosure risk	The risk that the identity of individuals, or new information about known individuals, within a dataset can be revealed. Includes both attribute disclosure and identity disclosure.
Statistical properties	Characteristics of a dataset that can be measured, analysed or interpreted
Synthetic health data	Data generated by a system or model that can mimic and resemble the structure and statistical properties of real health data, and uses real health data as input.[23]
Synthetic health data request	A synthetic health data request could include: One organisation requests another organisation to generate a synthetic health dataset for a specific project

[23] From the IAPP: https://iapp.org/resources/article/key-terms-for-ai-governance/

	An organisation wishes to establish a synthetic health dataset for multiple potential projects / purposes An organisation (or End User) requests to access or use a synthetic health dataset that was created for a different purpose or use case
Unit record data	Also called 'micro' data, this is data at the level of a single observation, for example data items relating to a unique individual, or a particular entity (such as a general practice)
Use	The use of personal information by a person inside an organisation
'Very low risk'	In the context of this Framework, 'very low risk' of re-identification means that even though it may be technically possible to identify an individual from information, doing so is so impractical that there is almost no likelihood of it occurring.[24]

[24] This is the standard of de-identification used by the OAIC for information to no longer be regarded as 'personal information' for the purposes of the Privacy Act. See: https://www.oaic.gov.au/privacy/privacy-guidance-for-organisations-and-government-agencies/handling-personal-information/what-is-personal-information

APPENDIX 3: The policy and legal framework underpinning this Framework

The system of privacy regulation in Australia is best described as 'patchwork'. Given the range of current and potential future SynD members, a range of privacy laws may apply to a single synthetic health data project.

Federal government agencies and private sector organisations are generally regulated by the Australian Privacy Principles (APPs), which are found in the federal *Privacy Act 1988* (Cth).[25]

State/Territory and local government agencies, including public universities, are generally regulated by their own State/Territory privacy laws. Some States/Territories have multiple privacy laws and sets of privacy principles. For example, NSW has the Information Protection Principles (IPPs) set out in the *Privacy and Personal Information Protection Act 1998* (NSW) (PPIP Act) and the Health Privacy Principles (HPPs) set out in the *Health Records & Information Privacy Act 2002* (NSW) (HRIP Act). However, some States – such as South Australia – do not have any specific privacy laws regulating either State or local government agencies.

Some private sector organisations will be regulated by *both* federal and State/Territory privacy laws. For example, the HRIP Act applies the NSW HPPs to private sector health service providers, in addition to the federal APPs.

Each privacy law contains a set of privacy principles, each of which covers similar ground, regulating the collection, use and disclosure of 'personal information' (or a sub-set of personal information, such as 'health information' or 'sensitive information'), as well as data security, data quality, access and correction rights. As a very general statement, privacy laws typically prohibit the disclosure of personal information, unless an exception applies.

The privacy laws that apply to SynD organisations (both current and future) could include:

Commonwealth privacy laws

- the *Privacy Act 1988* (Cth) (Privacy Act)[26]

The Privacy Act applies to Commonwealth government agencies (including Commonwealth public universities) and private sector organisations. This includes all private sector and NGO health service providers across Australia.

[25] There is an exemption for private sector organisations with a turnover of less than $3M pa, but this exemption does not apply to 'health service providers'. This means that even very small NGO and community organisations which offer health services will therefore be regulated by the APPs.
[26] See: https://www.legislation.gov.au/C2004A03712/latest/text

Australian Capital Territory

- the *Information Privacy Act 2014* (ACT) (IP Act),[27] and
- the *Health Records (Privacy and Access) Act 1997* (ACT) (HRPA Act)[28]

The IP Act applies to ACT public sector agencies (including ACT public universities) and sets out 13 Territory Privacy Principles (TPPs).

The IP Act does not apply to the handling of 'health information' about individuals, which is instead regulated by the HRPA Act.

The HRPA Act applies to both public sector and private sector health service providers and contains 12 Privacy Principles (PPs).

New South Wales

- the *Privacy and Personal Information Protection Act 1998* (NSW) (PPIP Act),[29] and
- the *Health Records and Information Privacy Act 2002* (NSW) (HRIP Act)[30]

NSW public sector agencies (including NSW public universities) must comply with the PPIP Act which has 12 Information Protection Principles (IPPs) and the HRIP Act which has 15 Health Privacy Principles (NSW HPPs). The HRIP Act also applies to all private sector health service providers.

Northern Territory

- the *Information Act 2002* (NT) (Information Act)[31]

Northern Territory public sector agencies (including NT public universities) must comply with the IP Act, which contains 10 Information Privacy Principles (NT IPPs).

Queensland

- the *Information Privacy Act 2009* (QLD) (IP Act)[32]

[27] See: https://www.legislation.act.gov.au/View/a/2014-24/current/html/2014-24.html
[28] See: https://www.legislation.act.gov.au/View/a/1997-125/current/html/1997-125.html
[29] See: https://legislation.nsw.gov.au/view/html/inforce/current/act-1998-133
[30] See: https://legislation.nsw.gov.au/view/html/inforce/current/act-2002-071
[31] See: https://legislation.nt.gov.au/Legislation/INFORMATION-ACT-2002
[32] See: https://www.legislation.qld.gov.au/view/whole/html/speciallabel/bill-2022-041/act-2009-014 (this version of IP Act indicates the amendments that will be made by the IPOLA Act)

The IP Act contains 13 Queensland Privacy Principles (QPPs). Queensland government agencies (including Queensland public universities) must comply with the QPPs.

South Australia

There are no specific privacy laws in South Australia. However, the Department of Premier and Cabinet has issued a privacy Instruction for South Australian public sector agencies (Premier and Cabinet Circular 12 - Information Privacy Principles Instruction)[33]. The Instruction contains 10 Information Privacy Principles (SA IPPs) which apply to the handling of personal information. The Instruction was last re-issued in May 2020.

While the Instruction creates a binding policy for public sector agencies, it is not law and cannot be enforced by a court. The Instruction creates the Privacy Committee of South Australia, which handles privacy complaints relating to the SA IPPs.

Tasmania

- the *Personal Information Protection Act 2004* (Tas) (PIP Act)[34]

Tasmanian public sector agencies (including Tasmanian public universities) are required to comply with the PIP Act, which has 10 Personal Information Protection Principles (PIPPs).

Victoria

- the *Privacy and Data Protection Act 2014* (Vic) (PDP Act),[35] and
- the *Health Records Act 2011* (Vic) (HR Act)[36]

Victorian public sector agencies (including Victorian public universities) must comply with the PDP Act, which has 10 Information Privacy Principles (IPPs) and with the HR Act which has 11 Health Privacy Principles (Vic HPPs).

Private sector health service providers are also required to comply with the HR Act.

[33] See: https://www.dpc.sa.gov.au/resources-and-publications/premier-and-cabinet-circulars/DPC-Circular-Information-Privacy-Principles-IPPS-Instruction.pdf
[34] See: https://www.legislation.tas.gov.au/view/whole/html/asmade/act-2004-046
[35] See: https://www.legislation.vic.gov.au/in-force/acts/privacy-and-data-protection-act-2014/031
[36] See: https://www.legislation.vic.gov.au/in-force/acts/health-records-act-2001/049

Western Australia

- the *Privacy and Responsible Information Sharing Act 2024* (WA) (PRIS Act)[37]

Until very recently, Western Australia did not have comprehensive privacy legislation to regulate how the WA public sector handles personal information. At this stage, it is anticipated that the privacy provisions in the PRIS Act will commence in 2026.[38]

The PRIS Act creates 11 Information Privacy Principles (WA IPPs) that apply to the handling of personal information (including health information) by WA public sector organisations (including WA public universities).

The PRIS Act also creates a responsible information sharing legislative framework that governs the handling of WA government data. Under this Framework, information sharing agreements can be formed under the terms of the PRIS Act, giving legal authority for the collection, use and disclosure of government information.[39] These agreements may be made between WA public entities and other parties – either other WA public entities or external entities.

Other relevant laws

Organisations may also be subject to other laws that impact data handling. These laws typically govern specific functions or services of an organisation, and can guide the collection, use, disclosure and management of personal information. An example of a law governing a specific function or service would include the *Public Health Act 2010* (NSW).

Other relevant policies

SynD members may also have their own policies that apply to the creation and handling of synthetic health data. These policies will need to be applied by these organisations alongside this Framework where a synthetic health data request is within scope of this Framework.

[37] See: https://www.legislation.wa.gov.au/legislation/statutes.nsf/law_a147470.html
[38] News story: Interim advice for all agencies about the protection of personal information, 9 December 2024, accessible at https://www.wa.gov.au/government/announcements/interim-privacy-position-0
[39] See Part 3, Division 5, PRIS Act.

Summary of applicable Australian privacy laws that regulate the handling of health information**

Where privacy laws establish specific obligations for the handling of 'health information' or 'sensitive information' (as a distinct category of 'personal information') – either through separate legislation or a specific set of privacy principles – these are the relevant privacy obligations that will apply to the handling of health data. Other non-privacy laws may also impact the handling of health information.

Organisation	**Privacy Act/APPs**	**ACT HRPA Act/PPs***	**NSW HRIP Act/HPPs***	**NT IP Act/IPPs***	**QLD IP Act/QPPs***	**SA Privacy Instruction/ IPPs***	**TAS PIP Act/PIPPs***	**VIC HR Act/HPPs***	**WA PRIS Act/IPPs***
Private sector health service providers (e.g. GPs, private hospitals)	Y	Y	Y	N	N	N	N	Y	N
Health NGOs (e.g. PHNs)	Y	Y	Y	N	N	N	N	Y	N
State/Territory public sector agency (e.g. public hospitals, state public universities)	N	Y	Y	Y	Y	Y	Y	Y	Y
Commonwealth government agencies (e.g. Department of Health, Disability and Ageing)	Y	N	N	N	N	N	N	N	N

* If operating in the State or Territory

** Where Entity A engages Entity B as a 'contracted service provider', the privacy laws that apply to Entity A may then also apply to Entity B.

APPENDIX 4: Use Case Assessment

Use Case Assessment Checklist

The accountable decision-maker must assess the use case for the synthetic health data request against the following three tests. Each of these three tests must be met in order for the synthetic health data request to proceed to the next steps:

Test	What must be met
1.	The use case is for a **clear 'public benefit' purpose** related to providing health services, and where the expected benefits from the use case are related to consumer health or health system outcomes. *For the purposes of this Framework, the types of use cases that may meet this test include:* • *Research-related activities, where the research is for a public benefit purpose. These activities could include conducting PoC analysis and feasibility testing, carrying out preliminary research tasks; or using PoC results to vet research proposals* • *Tasks related to the management, planning and funding of healthcare system activities, where the task does not require real data* • *Education and training activities that are related to analysing health data for beneficial purposes (such as hackathons that test models for generating insights into population health)* • *Healthcare technology development and improvement activities that are for a clear public benefit purpose*
2.	The stated **aim for creating and managing the synthetic health dataset is to achieve a 'de-identified' dataset for the use case, that significantly minimises the risk to individuals** compared to if the source dataset (i.e. real data) was used for the use case.
3.	The organisation that collected and holds the source data has **set expectations with health consumers** about how their health information will be used through suitable public messaging.

If the Data Custodian is not satisfied that the synthetic health data request meets these tests, an alternative lawful privacy pathway must be determined (see Appendix 8 for guidance on dealing with complex synthetic health data requests).

Further Resources

Relevant organisational policies that may need to form part of an organisation's use case assessment can be set out / linked to here

- WA Department of Health, Synthetic health data: Governance and Technical Guidelines for the Generation and Use of Synthetic health data
- …

APPENDIX 5: Impact Assessment

Impact Assessment Checklist

The accountable decision-maker must assess the synthetic health data request in consideration of the following:

Considerations:	
Public interest	Will the synthetic health data use case be in the public interest? This question should be considered in relation to: • what the objective is when creating a synthetic health dataset, as well as • each request for access and use of the synthetic health dataset
Resourcing	What will be the resourcing impacts on the organisation, both in expenses and time? Will this project create a high opportunity cost - i.e. will resources be diverted from other projects more closely aligned to the organisation's policy or strategic objectives?
Beneficiaries	Will the synthetic health data use case be for the benefit of many or a few? Can we maximise the utility of the data to be shared, to benefit multiple parties or derive useful insights to assist health care delivery and systems?
Community expectations and trust	Will the synthetic health data use case meet community expectations about how their data will be handled? Will the data sharing take place in a manner that maintains or builds trust with stakeholders?
Privacy impacts	A Privacy Impact Assessment (PIA) will be needed for high-risk projects involving data sharing for linkage prior to creating a synthetic health dataset, and where re-identification risk cannot be lowered to a very low level, in order to manage legal compliance risks.
Data ethics	Does the synthetic health data use case raise concerns with regard to ethical considerations, in terms of the potential for harmful impacts on individuals, particular cohorts or sections of the community? See further discussion below.
Indigenous data sovereignty	Does the data include data about indigenous status? See further discussion below.

Data ethics

Examples of data types that could raise ethical concerns include data about:

- Health information (including physical and mental health)
- Disability

- Address or housing status
- Alcohol and drug use
- Sexuality
- Children
- Family violence / AVO's
- Aboriginality
- Ethnicity
- Language background other than English
- Religion
- Trade union membership
- Text-based fields / unstructured data
- Staff information
- Practice performance
- Services to be commissioned, where one of the organisations may be eligible to bid to provide the service during a procurement process
- Identifiable cohorts, even where individuals who belong to the cohort are not themselves reasonably identifiable from the data

Examples of scenarios that could raise ethical and legal concerns include:

- There is data linkage across multiple data sources and/or platforms
- There is legislation (other than the privacy laws) protecting this specific type of data
- High risk technologies will be applied to the data. For example, generative AI technologies to conduct analysis and generate new information and insights, where the technology has not been assessed as safe
- The data or results will be published or shared with an external party
- The risk of NOT proceeding with the use case, where meaningful benefits might otherwise flow to health consumers as a result of the use case

The ethics of synthetic health data

Even where synthetic health data has been robustly re-identified, organisations should be aware that there may still be ethical concerns that can arise from the use of this data and which should be considered. These include:

Incorrect recognition of patients

Even if they are completely incorrect, if a person with access to a synthetic health dataset believes that they recognise an individual in a synthetic health dataset, and as a result they 'learn' new facts about that person such as a medical condition (whether that fact turns out to be correct or not), privacy harm can be done to a real person.

Incorrect inferences
Some methods of controlling for re-identification risk include suppressing some data fields (e.g. country of birth), and/or removing small cell counts (e.g. people with unique combinations of variables). However, this may further entrench disadvantage, if some groups are then excluded from analysis.

Similarly, any inherent biases in the reference data will be carried through to the synthetic health data.

If the synthetic health dataset is not truly representative of the real population, there is a risk that a person with access to the synthetic health dataset could draw inferences or statistical conclusions that are invalid. Whether those inferences are about individuals, populations, medical conditions, medical interventions or the quality of health service providers, invalid conclusions could lead to poor quality policy or operational decisions, or interventions in the health system.

Therefore if the synthetic health data is to be used to make policy or operational decisions that could have consequences for individuals, it will be important to detect and correct bias in the generation of synthetic health data, and ensure that the synthetic health data is representative.[40]

In addition, the revelation that inferences drawn from the synthetic health dataset were incorrect could in turn undermine confidence in the integrity of the *source* dataset, and the program of data collection that underpins it.

Correct inferences
Even if *correct*, inferences drawn from synthetic health data could lead to policy or operational decisions that negatively impact vulnerable populations. Disregard of the risk of such outcomes would be contrary to the principles of ethical design and beneficence.

For example, while not creating a privacy risk for any individual patients, inferences drawn and publicised about the health outcomes for patients of a particular rural hospital or GP service could lead to a loss of confidence in those local health services. Local communities might face further disadvantage if the inferences were used to cut funding, and/or even poorer health outcomes might result if patients avoided using their local services at all.

Indigenous data sovereignty

Synthetic health data use cases, which involve data about Aboriginal and Torres Strait Islander peoples, are subject to the same requirements of this Framework as any other synthetic health data use case. However some additional considerations will also apply.

[40] 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/

SynD organisations are committed to Indigenous data governance, which is the right of Indigenous peoples to autonomously decide what, how and why Indigenous data and information are collected, accessed and used. It ensures that data on or about Indigenous peoples reflects their priorities, values, cultures, worldviews and diversity.

Proponents of Indigenous Data Sovereignty argue that in the past, research has too often been done 'on' rather than 'for' or 'by', Aboriginal and Torres Strait Islander peoples and communities. Thus, Indigenous Data Sovereignty is achieved when projects to support policy development or research:

- are led by Aboriginal or Torres Strait Islander experts
- have teams that include Aboriginal and Torres Strait Islander researchers or policy-makers, and
- collect and use data that reflects Aboriginal and Torres Strait Islander values and frameworks, such as cultural determinants of health.

Proposals to generate synthetic health data which include data about Aboriginal and Torres Strait Islander peoples may need to be approved by an appropriate accountable decision-maker at an organisation.

Collaborating and co-designing with Aboriginal and Torres Strait Islander groups (such as Aboriginal Community Controlled Health Organisations (ACCHOs)) on synthetic health data projects may provide opportunities to empower and support Indigenous Data Sovereignty, and to support research and data analytics designed to benefit Aboriginal and Torres Strait Islander communities. Synthetic health data can provide a pathway for these benefits to be realised in a timely and cost effective manner, and can support research investment and advocacy efforts for Aboriginal and Torres Strait Islander groups.

Where a synthetic health dataset is still considered 'health data' and / or ethics review is required, using or disclosing this data for research about Aboriginal and Torres Strait Islander peoples will also require the approval of a registered Aboriginal HREC, such as the Aboriginal Health & Medical Research Council (AH&MRC), or the Research Ethics Committee of the Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS).

Further resources

Relevant organisational policies that may need to form part of an organisation's impact assessment can be set out / linked to here

E.g. Privacy Impact Assessment guides or frameworks, AI assessment frameworks, data ethics guides, etc.

    - Office of the Information Commissioner's (OAIC's) Guide to undertaking Privacy Impact Assessments
- The Lowitja Institute has produced an Information Sheet, 'Indigenous Data Governance and Sovereignty', for organisations involved in research.[41]
- The National Health and Medical Research Council (NHMRC)'s *National Statement on Ethical Conduct in Human Research* outlines key matters for ethical consideration, including considerations specific to participants in research.[42]
- NHMRC 2003, *Values and Ethics Guidelines for Ethical Conduct in Aboriginal and Torres Strait Islander Peoples about health research ethics*.[43]
- NHMRC 2018, *Keeping Research on Track II: A companion document to Ethical conduct in research with Aboriginal and Torres Strait Islander Peoples and communities: Guidelines for researchers and stakeholders*[44]
- Australian Institute of Aboriginal and Torres Strait Islander Studies 2012, *Guidelines for Ethical Research in Indigenous Studies*[45]
- Cultural and Indigenous Research Centre Australia 2020, *Aboriginal Privacy Insights Report*, commissioned by the Office of the Victorian Information Commissioner[46]

[41] The Lowitja Institute, Information sheet: 'Indigenous Data Governance and Sovereignty', 2021; https://www.lowitja.org.au/icms_docs/328550_data-governance-and-sovereignty.pdf
[42] https://www.nhmrc.gov.au/about-us/publications/national-statement-ethical-conduct-human-research-2025
[43] https://www.nhmrc.gov.au/about-us/publications/values-and-ethics-guidelines-ethical-conduct-aboriginal-and-torres-strait-islander-health-research#block-views-block-file-attachments-content-block-1
[44] 2018, https://www.nhmrc.gov.au/about-us/resources/keeping-research-track-ii
[45] 2012, https://aiatsis.gov.au/sites/default/files/2020-09/gerais.pdf
[46] https://ovic.vic.gov.au/wp-content/uploads/2020/10/Cultural-and-Indigenous-Research-Centre-Australia-Presentation-on-Aboriginal-Privacy-Insights-Report.pdf

APPENDIX 6: Technical Assessment

The Data Custodian at the Data Provider typically liaises with the responsible Data Requestor about the suitability of the source data to meet the terms of the request. The Data Custodian at the Data Provider may also bring in specialist expertise, for example on issues around statistical modelling.

Data Requestors should provide context or explanation for why they are asking for certain data for a particular use case, so as to facilitate this assessment. Providing the Data Requestor the metadata or a data dictionary/terminology used by the Data Provider can assist in clarifying the scope of the data request, as well as progress discussions about the feasibility of the proposed methodology.

Technical Assessment Checklist

The accountable decision-maker must assess the request and the relevant source data in consideration of the following:

Considerations:	
Data source	Has the most appropriate data source been identified? Is there already an available data asset that could meet this need? Catalogues setting out datasets and repositories held by organisations can assist Data Requestors to understand what data is available.
Availability	Is the data actually available in the requested and appropriately structured format? For complex requests, this may require a detailed specification to be provided, listing all parameters and selection criteria for the data. Data dictionaries/ terminology and metadata can assist Data Requestors to understand what data is available.
Quality	Is the data of high quality? Indicators of data quality are accuracy, timeliness, consistency, validity, uniqueness and completeness.
Reliability	Is the data valid? Is it reasonable to rely on the data requested to provide valid results for the intended purpose?
Context	The assessment of whether data is fit for purpose should reflect the context of the synthetic health data request, including the overall project, expected outcomes and any decision-making that will rely on the data.
Comprehensible	Is the data comprehensible to enable proper understanding of the information? We need to avoid possible misinterpretation or misuse of information.

Granularity	Is the requested level of granularity appropriate? Could the Data Requestor's objectives be achieved by providing aggregate data instead, without having an undue impact on the utility of the data? In order to minimise the risks of re-identification, Data Requestors should be encouraged to practice data minimisation: • only request data that is required to directly answer the key question related to the use case • minimise the number of demographic data fields requested • aim for the broadest geographical level (e.g. LGA rather than postcode) • reduce the timeframe required to hold the requested synthetic health data
Utility	Is the data 'representative enough' for the use case at hand? Will it provide value in terms of the purpose underpinning the proposed use case / the question it is intended to help answer?
Capability	Does the Data Requestor have the appropriate skills to understand and interpret the data?
Methodology	Is the Data Requestor's proposed statistical methodology appropriate?
Clarity	Has a statement of data quality been prepared? What other notes need to be supplied with the synthetic health data to support End Users of the data to interpret it appropriately, including any caveats or statements about data quality? A statement of data quality should be prepared, to assist Data Requestors in determining if the data is fit for their specific use case. The statement of data quality should be made available with the synthetic health data, together with relevant information to assist with interpreting the synthetic health data. Where appropriate, there should be explanatory material to support End Users of the data to interpret it appropriately, including a data dictionary/terminology and any caveats or statements about data quality in relation to the data presented.

Further resources

Relevant organisational policies that may need to form part of an organisation's technical assessment can be set out / linked to here.

E.g. Data quality frameworks and standards, data quality statement templates, guidelines related to the technical creation of synthetic health data, etc.

- WA Department of Health, Synthetic health data: Governance and Technical Guidelines for the Generation and Use of Synthetic health data
- ABS Data Quality Statement Checklist

APPENDIX 7: De-identification techniques and Evaluation of Privacy in Synthetic Data

De-identification techniques

De-identification aims to break the link between a dataset and an individual in the real world, so that the disclosure of a fact (such as that a patient is being treated at this hospital for HIV) cannot be linked back to an identified individual (the patient is Sally Citizen).

The harm being prevented here is known as 'identity disclosure'. Identity disclosure - which occurs when data is re-identified - can arise in one of two ways: by either matching a person to data, or matching data to a person. Checking the robustness of de-identification techniques should involve testing your dataset for both these types of re-identification risk.[47]

> "De-identification is not a single technique, but a collection of approaches, algorithms, and tools that can be applied to different kinds of data with differing levels of effectiveness. In general, privacy protection improves as more aggressive de-identification techniques are employed, but less utility remains in the resulting dataset."[48]

There is no single 'correct' way to de-identify data. It is an exercise in risk management. Re-identification risks will differ according to the type of data, its context, and other factors. Trade-offs need to be made between minimising the risk of re-identification, and maximising the value of the data. The wrong de-identification method can fail to reduce privacy risk, and/or decrease data utility. As such, the information outlined in this section should be considered as general guidance about de-identification techniques, and not as a specific plan to achieve a 'de-identified' synthetic health dataset.

Examples of de-identification techniques include:

- aggregation
- suppression (remove identifiers or other overtly identifying data fields)
- generalisation (e.g. replace exact date of birth with a date range like '35-44 year olds')
- pseudonymisation (replace direct identifiers with statistical linkage keys (SLKs), or encrypt or hash identifiers), and
- perturbation (adding noise, micro-aggregation or data-swapping).

[47] When considering re-identification risks, the GDPR makes clear that the identifiability of data should not be considered in a vacuum. Instead, "account should be taken of all the means reasonably likely to be used … to identify the natural person directly or indirectly", including "objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments"; see GDPR Recital 26.

[48] Simson L. Garfinkel, NISTIR 8053: De-Identification of Personal Information, National Institute of Standards and Technology, US Department of Commerce, 2015, p.1; available at http://dx.doi.org/10.6028/NIST.IR.8053.

- Evaluation of Privacy in Synthetic Data

Evaluation of Privacy in Synthetic Data

Introduction

Privacy protection is the first principle guiding synthetic data generation, even though synthetic data ultimately represents a trade-off between fidelity, utility, and privacy. Privacy evaluation is therefore a critical yet often misunderstood component of synthetic data governance. While synthetic data aims to protect privacy by generating artificial records without direct links to real individuals, residual privacy risks persist. Some generative models incorporate privacy-preserving mechanisms by design (for example, DPGAN and PATEGAN implement differential privacy, while ADSGAN targets re-identification risks [1-3]), yet issues such as model overfitting and the inadvertent retention of sensitive patterns can compromise privacy. Consequently, robust privacy assessment remains essential, as risks such as membership and attribute inference may arise when synthetic data preserves statistical patterns from the original dataset.

This appendix complements the main text by providing a concise overview of current approaches to evaluating privacy in synthetic data. There is no single accepted definition or universal measure of privacy risk; existing metrics capture only partial aspects of it. The aim here is to clarify what these measures assess and what they do not, emphasising that no single score can demonstrate overall privacy safety. The discussion is conceptual rather than implementation-focused; we do not provide algorithmic details or code, although several of the described metrics may be available through open-source R or Python packages.

A wide range of metrics has been proposed to evaluate privacy in synthetic data [4, 5], with various classification schemes presented in the literature. One of the most practical groupings is summarised in Table 1 [6]. These metrics span traditional re-identifiability measures such as *k-anonymity*, *l-diversity* and *t-closeness* to distance-based metrics like Nearest Neighbour Distance Ratio (NNDR) and adversarial approaches, including membership inference attacks. While some studies recommend using individual metrics or combinations of them, others advocate for broader frameworks that assess privacy alongside utility and fidelity [7-9]. Although these methods aim to capture privacy risks, none fully resolve the inherent trade-off between privacy and usability. Developing rigorously validated, context-aware evaluation criteria remains an open challenge [10].

Despite these efforts, there remains no unified standard defining what constitutes adequate privacy protection or how to interpret the results of these evaluations in practice. This lack of consensus is further complicated by the variation of synthetic data generation processes.

Table 1: Categories of metrics used for evaluating privacy in synthetic data.

Evaluation Category		Evaluation Method / Metric	Description
I. Non-Adversarial Metrics (often adapted from anonymised datasets.)	**A. Re-identifiability Metrics**	k-Anonymity	Checks if every record is indistinguishable from at least k_1 other records based on quasi-identifiers.
		l-Diversity	An extension of k-Anonymity ensuring that sensitive attributes within each anonymised group have at least l distinct values.
		t-Closeness	Further refines l-Diversity by ensuring the distribution of a sensitive attribute in any group is close to its distribution in the overall dataset.
	B. Memorisation Metrics	Hitting Rate (Common Row Proportion)	Measures the direct percentage of exact matching records (overlapping rows) between the synthetic and source data.
		Close Value Ratio	Assesses the probability of having "blurry matches" or similar values between synthetic and source data, defined by a distance threshold.
		Similarity Ratio (ε-identifiability)	Tests whether less than an ε ratio of synthetic observations are "similar enough" to those in the original dataset, often measured using weighted Euclidean distance.
		Nearest Neighbour Accuracy (Adversarial Accuracy)	Evaluates the proximity of a point in the original distribution (P_R) to its nearest counterpart in the synthetic distribution (P_S); an optimal value of 0.5 suggests indistinguishability.
	C. Distinguishability Metrics	Data Likelihood	Measures the likelihood of synthetic data belonging to the source data distribution, often using Bayesian Networks or Gaussian Mixture Models.
		Detection Rate	Assesses the difficulty of distinguishing source data from synthetic data using machine learning models like logistic regression.
II. Adversarial Metrics (Attack-Based) (Metrics that involve applying actual privacy attacks and measuring the	**A. Singling Out Attacks**	Singling Out Attack (*Univariate Attack)*	Observes the uniqueness of a record or attribute combination in the synthetic data to assess the likelihood that this uniqueness is derived from the original data. (Focuses on rare values of a single attribute (predicate).)
		Singling Out Attack (*Multivariate Attack)*	Involves combinations of multiple attributes (predicates).
	B. Record Linkage Attacks	Public-Public Linkage	Uses the synthetic dataset (S) to establish connections between records found in two separate external datasets (X_1 and X_2).
		Public-Synthetic Linkage	Links rows in the synthetic dataset (S) to a public dataset (X') using matching criteria, serving as a basis for inference attacks.

success ratio, offering a definition of "practical privacy".)	**C. Attribute Inference Attacks (AIA)**	Exact Match AIA	Determines the value of a missing target attribute by precisely matching overlapping quasi-identifiers (Q) between the synthetic data and the target records.
		Closest Distance AIA	Infers the missing sensitive value by identifying the single most similar data point (k=1) in S to the target record (equivalent to a K-Nearest Neighbour model).
		Nearest Neighbours AIA	Deduces the sensitive value by examining the k nearest neighbours (k > 1) in the synthetic dataset.
		ML Inference AIA	Attackers train a predictive machine learning model on S and use it to predict the target attributes of the target records.
	D. Membership Inference Attacks (MIA)	Closest Distance MIA	Infers membership if the target record is significantly more similar to the synthetic data than to unrelated data.
		Nearest Neighbours MIA	Relaxes the criteria of Closest Distance MIA to include proximity to k neighbours (k > 1).
		Probability Estimation MIA	A hypothesis testing method assessing the likelihood that a target record belongs to the synthetic data distribution (and thus the original data distribution).
		MIA Shadow Model	Adversaries create "generative shadow models" using reference datasets (one including the target record, labelled 1; one excluding it, labelled 0) to train a classifier that distinguishes membership.

Acknowledging the absence of universal guidelines with predefined thresholds for privacy assessment, this appendix outlines a practical approach to evaluating privacy in synthetic data. The proposed guidance is based on a published study [11] developed in collaboration with international experts in privacy and synthetic health data, aiming to provide actionable methods while addressing common misconceptions.

The suggested framework stems from a rigorous Delphi consensus process involving 13 global privacy specialists. Through three structured rounds, the panel reached agreement on ten core recommendations. This consensus-driven approach helps bridge the critical gap in standardised methods for assessing privacy in synthetic data.

Fundamental Concepts

Privacy evaluation in synthetic data focuses on three primary disclosure types:

1. **Identity Disclosure**: Occurs when a synthetic record can be linked to a specific individual. Whilst synthetic data generation should inherently protect against this by design, overfitting can still create vulnerabilities. It has been argued that identity disclosure is less relevant for synthetic data, since synthetic records are artificial and do not correspond to real individuals; the more substantive privacy risks arise from membership and attribute disclosure, which can reveal information about real data subjects. Hence, we focus on these two measures in what follows. Attempts to measure identity disclosure often rely on record-level similarity metrics, but these lack clear interpretation and tend to conflate with membership or attribute disclosure depending on whether correctness of sensitive information is considered.

2. **Membership Disclosure**: This occurs when it is possible to determine whether a specific individual's data was included in the training dataset used to generate the synthetic data. This becomes particularly concerning when the training data contains sensitive information (e.g., participants in a clinical trial for a specific condition). Membership disclosure could be treated as a binary classification task where correctly identifying a data member is a "true positive." To quantify this vulnerability, metrics such as the F1 score can be used, combining precision and recall to evaluate the performance of membership inference attacks. A baseline metric, F_{naive}, reflects the expected success of an adversary guessing membership status without using the synthetic data:
    a. **F1 score**: The F1 score is the harmonic mean of precision and recall and therefore measures the balance between correctly identifying positives and avoiding false positives.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

    b. **$F_{naïve}$**: The $F_{naïve}$ score represents the expected F_1 value under random guessing and serves as a baseline to assess whether a model's performance reflects genuine predictive ability rather than chance.

$$F_{naïve} = \frac{2 \times p}{1+p}$$ *(p = sampling fraction of the original dataset from the population)*

[11]

		Prediction by adversary		
		Member	Non-member	
Ground truth	Member	TP	FN	$recall = \frac{TP}{TP+FN}$
	Non-member	FP	TN	$specificity = \frac{TN}{TN+FP}$
		$precision = \frac{TP}{TP+FP}$	$NPV = \frac{TN}{TN+FN}$	$accuracy = \frac{TP+TN}{TP+TN+FP+FN}$

Confusion Matrix in Membership Disclosure Vulnerability. FN: false negative; FP: false positive; NPV: negative predictive value; TN: true negative; TP: true positive.

3. **Attribute Disclosure**: This occurs when an adversary can infer sensitive information about an individual based on the synthetic data. This requires careful distinction between legitimate knowledge generation (learning population patterns) and privacy violations (learning specific information about individuals in the training data).

Other definitions and formulas:

Differential Privacy: Differential privacy (DP) is a mathematical framework that guarantees the output of an analysis is nearly indistinguishable regardless of whether any single individual's data is included in the input. This provides a quantifiable and robust privacy guarantee against a wide range of attacks. Unlike the disclosure risk discussed previously (which are properties of the dataset), DP is an a priori feature of the data generation or analysis process.

Privacy budget (ϵ): The privacy budget is a non-negative parameter that is **set** to control the level of privacy. It's a "budget" in the sense that it's a resource we spend each time we query the data. A randomised algorithm M is **ε-differentially private** if for all pairs of neighbouring datasets D and D' (differing by just one individual), and for all possible outputs $S \subseteq Range(M)$: $\frac{\Pr[M(D) \in S]}{\Pr[M(D') \in S]} \leq e^{\epsilon}$

Common Misconceptions to Avoid

- **Misconception 1: Synthetic Data Is Inherently Private**

Synthetic data generation does not automatically guarantee privacy. Models can overfit, memorising specific details rather than learning general patterns, potentially encoding information about individuals in the training data.

- **Misconception 2: Record-Level Similarity Indicates Privacy Risk**

Unlike anonymised data, the similarity between synthetic and source records lacks precise interpretation for privacy risk. A synthetic record might resemble multiple training records, or similarity in quasi-identifiers might not translate to similarity in sensitive attributes. Stand-alone similarity metrics is inadequate to report privacy in synthetic data.

- **Misconception 3: All Attributes Should Be Considered in Privacy Evaluation**

Assuming an adversary knows all attributes is not necessarily a worst-case scenario. Simulations demonstrate that considering all variables may underestimate disclosure vulnerability. Privacy evaluation should focus on quasi-identifiers relevant to the specific context.

- **Misconception 4: Large Differential Privacy Budgets Ensure Privacy**

The privacy budget ε lacks uniform translation to empirical privacy unless very close to 0. Values ranging from 0.1 to over 18 in published work demonstrate this parameter's inconsistent interpretation. Empirical evaluation remains necessary even for differentially private synthetic data.

Consensus-Based Recommendations

- **R1: Base Evaluations on Quasi-Identifiers**

Disclosure vulnerability metrics should be based on quasi-identifiers (QIs) rather than all attributes. QIs represent the adversary's realistic background knowledge and vary by context. The data controller should determine appropriate QIs, though treating all attributes as QIs remains an option if computationally feasible.

- **R2: Evaluate All Records**

Metrics should be calculated for all records, not pre-selected "vulnerable" subsets. Pre-selection introduces bias and may miss actual vulnerabilities. Maximum vulnerability can only be determined through a comprehensive evaluation.

- **R3: Avoid Stand-Alone Similarity Metrics**

Record-level similarity metrics without connection to membership or attribute disclosure lack meaningful interpretation and should not be used for privacy reporting.

- **R4: Align Membership Disclosure with Threat Models**

  Evaluate membership disclosure only when the assumptions hold, specifically, when adversaries would learn something new about targets from the same population as the training dataset. Misalignment between threat models and metrics can produce misleading vulnerability estimates.

- **R5: Report Prevalence-Adjusted Scores**

  F1 scores for membership disclosure must be reported relative to an adversary guessing membership (F_{rel}) to account for prevalence dependence. This provides a consistent interpretation across different member prevalence levels.

  $$F_{rel} = \frac{F1 - F_{naïve}}{1 - F_{naïve}},$$

- **R6: Limit Attribute Disclosure to Members**

  Meaningful attribute disclosure applies only to dataset members. Accurate predictions about non-members represent legitimate knowledge generation rather than privacy violations.

- **R7: Use Non-Member Baseline**

  Employ relative vulnerability metrics with non-member baselines to distinguish between information learned from dataset membership versus population membership.

- **R8: Apply Dual Thresholds**

  Consider relative vulnerability unacceptably high only when absolute vulnerability also exceeds its threshold. High relative vulnerability with low absolute accuracy doesn't constitute meaningful disclosure.

- **R9: Validate Differential Privacy Empirically**

  Unless ε is close to 0, empirical evaluation using standard metrics remains necessary even for differentially private synthetic data.

- **R10: Report Stochastic Variation**

  Report metrics for both individual synthetic datasets and multiple generations (averaged with variation measures) to account for the stochastic nature of synthetic data generation.

Practical Evaluation Guides

Guide for Membership Disclosure Evaluation

1. **Define the Threat Model**
   - Specify whether targets come from the training population or the broader population
   - Ensure attack dataset prevalence matches threat model assumptions

2. **Calculate Baseline Metrics**
   - Determine naive membership guess ($F_{naïve}$) based on member prevalence
   - Account for prevalence in interpretation
     $F_{naïve} = \frac{2 \times p}{1+p}$ *(p = sampling fraction of the original dataset from the population)*

3. **Compute Relative Metrics**
   - Calculate $F_{rel} = \frac{F1 - F_{naïve}}{1 - F_{naïve}}$
   - Suggested threshold: $F_{rel} \leq 0.2$ (requires contextual adjustment)

4. **Document Assumptions**
   - Explicitly state the target population
   - Clarify what membership reveals in the specific context

Guide for Attribute Disclosure Evaluation

1. **Define Quasi-Identifiers and Sensitive Attributes**
   - Based on realistic adversary knowledge
   - Consider context-specific factors

2. **Establish Non-Member Baseline**
   - Use holdout or external data
   - Calculate prediction accuracy for non-members ($A_{non-members}$)
     $A_{non-members}$: accuracy of predicting sensitive information for individuals who were not part of the Synthetic Data Generation (SDG) training dataset. It is the probability that the presumed sensitive target value is correct given that the attack record is not a member.

3. **Evaluate Member Vulnerability**
   - Calculate prediction accuracy for members ($A_{members}$)
     $A_{members}$ is represents the absolute accuracy of predicting a sensitive attribute for individuals who were part of the Synthetic Data Generation (SDG) training dataset. It is analogous to the probability that the presumed sensitive target value is correct given that the attack record is a member.

- Compute relative accuracy: $(A_{rel} = A_{members} - A_{non-members})$

4. **Apply Dual Thresholds**
   - Absolute threshold: $A_{members} \leq 0.6$ (exceeds poor prediction)
   - Relative threshold: $A_{rel} \leq 0.15$ (meaningful difference)
   - Both must be exceeded for unacceptable vulnerability

Evaluating Multiple Synthetic Datasets

Given the stochastic nature of synthetic data generation:

- **For Model Evaluation**
  - Generate multiple synthetic datasets (minimum 10 recommended)
  - Report average vulnerabilities and standard deviations
  - Document worst-case scenarios

- **For Data Release Decisions**
  - Evaluate the specific dataset(s) to be released
  - Consider maximum vulnerability across evaluations

Implementation Considerations

- **Computational Efficiency**
  - Start with domain-informed QI subsets
  - Progressively expand evaluation if resources permit
  - Balance thoroughness with practical constraints

- **Threshold Determination**
  - Thresholds must be:
  - Based on empirical precedents where available
  - Adjusted for context (data sensitivity, potential harm, consent)
  - Documented with clear justification
  - Subject to refinement as evidence accumulates

- **Reporting Requirements**

4. Comprehensive reporting should include:
   - Absolute metrics: Raw vulnerability measurements
   - Relative metrics: Baseline comparisons
   - Variation measures: Standard deviations across generations
   - Worst-case estimates: Maximum vulnerabilities observed
   - Context documentation: Threat models, assumptions, and adjustments

Future Directions

- **Immediate Priorities**
  - Empirical validation of suggested thresholds across domains
  - Standardisation of meaningful ε values for differential privacy
  - Development of automated evaluation tools

- **Long-term Goals**
  - Context-aware threshold adjustment systems
  - Joint optimisation of utility and privacy
  - Extension to non-tabular synthetic data types

Conclusion

Privacy evaluation in synthetic data requires systematic, evidence-based approaches that move beyond simplistic metrics and unfounded assumptions. This framework provides practical guidance whilst acknowledging that perfect privacy remains unattainable. The goal is not to eliminate all risk but to quantify and manage residual vulnerabilities appropriately.

Organisations implementing synthetic data must recognise that privacy evaluation is not optional but essential. By following these consensus-based recommendations and practical guides, data controllers can make informed decisions about synthetic data generation and sharing, balancing the tremendous potential of synthetic data with appropriate privacy protections.

The framework acknowledges that synthetic data evaluation is an evolving field. As empirical evidence accumulates and new threats emerge, these methods will require refinement. However, the fundamental principles, rigorous evaluation, realistic threat modelling, and empirical validation, provide a robust foundation for responsible synthetic data governance.

References

1. Xie, L., Lin, K., Wang, S., Wang, F., & Zhou, J. (2018). Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*. https://arxiv.org/abs/1802.06739
2. Jordon, J., Yoon, J., & Van Der Schaar, M. (2018, September). PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International conference on learning representations*. https://openreview.net/pdf?id=S1zk9iRqF7
3. Yoon, J., Drumright, L. N., & Van Der Schaar, M. (2020). Anonymization through data synthesis using generative adversarial networks (ADS-GAN). *IEEE journal of biomedical and health informatics*, *24*(8), 2378-2388. https://ieeexplore.ieee.org/abstract/document/9034117
4. Trudslev, F. M., Lissandrini, M., Rodriguez, J. M., Bøgsted, M., & Dell'Aglio, D. (2025). A Review of Privacy Metrics for Privacy-Preserving Synthetic Data Generation. *arXiv preprint arXiv:2507.11324*. https://arxiv.org/pdf/2507.11324
5. Osorio-Marulanda, P. A., Epelde, G., Hernandez, M., Isasa, I., Reyes, N. M., & Iraola, A. B. (2024). Privacy mechanisms and evaluation metrics for synthetic data generation: A systematic review. *IEEE Access*, *12*, 88048-88074. https://ieeexplore.ieee.org/document/10568134
6. Liao, Q., Van Landuyt, D., & Joosen, W. (2025). Pick Your Enemy: A Survey on Privacy Threat Models of Synthetic Tabular Data. https://www.authorea.com/doi/full/10.22541/au.174893956.65391176
7. Folz, J., Vidanalage, M. D., Aufschläger, R., Almaini, A., Heigl, M., Fiala, D., & Schramm, M. (2025). Scoring System for Quantifying the Privacy in Re-Identification of Tabular Datasets. *IEEE Access*. https://ieeexplore.ieee.org/document/10973096
8. Hernandez, M., Osorio-Marulanda, P. A., Catalina, M., Loinaz, L., Epelde, G., & Aginako, N. (2025). Comprehensive evaluation framework for synthetic tabular data in health: fidelity, utility and privacy analysis of generative models with and without privacy guarantees. *Frontiers in Digital Health*, *7*, 1576290. https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgth.2025.1576290/full
9. *Yan, C., Yan, Y., Wan, Z., Zhang, Z., Omberg, L., Guinney, J., ... & Malin, B. A. (2022). A multifaceted benchmarking of synthetic electronic health record generation models. Nature communications, 13(1), 7609.* https://www.nature.com/articles/s41467-022-35295-1?fromPaywallRec=false
10. Pierce, D. V., Li, Y., Greenshaw, A. J., Bailey, T., & Cao, B. (2025). Practical Steps in Implementing Privacy Measures With Synthetic Health Data. *World Medical & Health Policy*. https://onlinelibrary.wiley.com/doi/10.1002/wmh3.70023
11. Pilgram, L., Dankar, F. K., Drechsler, J., Elliot, M., Domingo-Ferrer, J., Francis, P., ... & El Emam, K. (2025). A consensus privacy metrics framework for synthetic data. *Patterns*. https://www.sciencedirect.com/science/article/pii/S2666389925001680

Further Resources

HealthStats NSW: Privacy issues and the reporting of small numbers

CSIRO & OAIC, *The De-Identification Decision-Making Framework*. Available at: https://www.oaic.gov.au/privacy/privacy-guidance-for-organisations-and-government-agencies/handling-personal-information/de-identification-decision-making-framework (the OAIC notes that, as this guide was produced in 2017, certain information it contains may now be out of date)

Office of the Victorian Information Commissioner (OVIC), *The Limitations of De-Identification – Protecting Unit-Record Level Personal Information*, available at: https://ovic.vic.gov.au/privacy/resources-for-organisations/the-limitations-of-de-identification-protecting-unit-record-level-personal-information/

Office of the Information Commissioner Queensland, *Report on Privacy and Public Data: Managing re-identification risk*, available at: https://www.oic.qld.gov.au/__data/assets/pdf_file/0016/43045/Privacy-and-public-data-managing-re-identification-risk.pdf

ISO/IEC 27559:2022
Information security, cybersecurity and privacy protection – Privacy enhancing data de-identification framework
https://www.iso.org/standard/71677.html

ISO/IEC 27554:2024
Information security, cybersecurity and privacy protection — Application of ISO 31000 for assessment of identity-related risk
https://www.iso.org/standard/71672.html

ISO/TS 14265:2024
Health informatics — Classification of purposes for processing personal health information
https://www.iso.org/standard/83447.html

ISO 25237:2017
Health informatics — Pseudonymization
https://www.iso.org/standard/63553.html

APPENDIX 8: Decision tree for complex synthetic health data scenarios

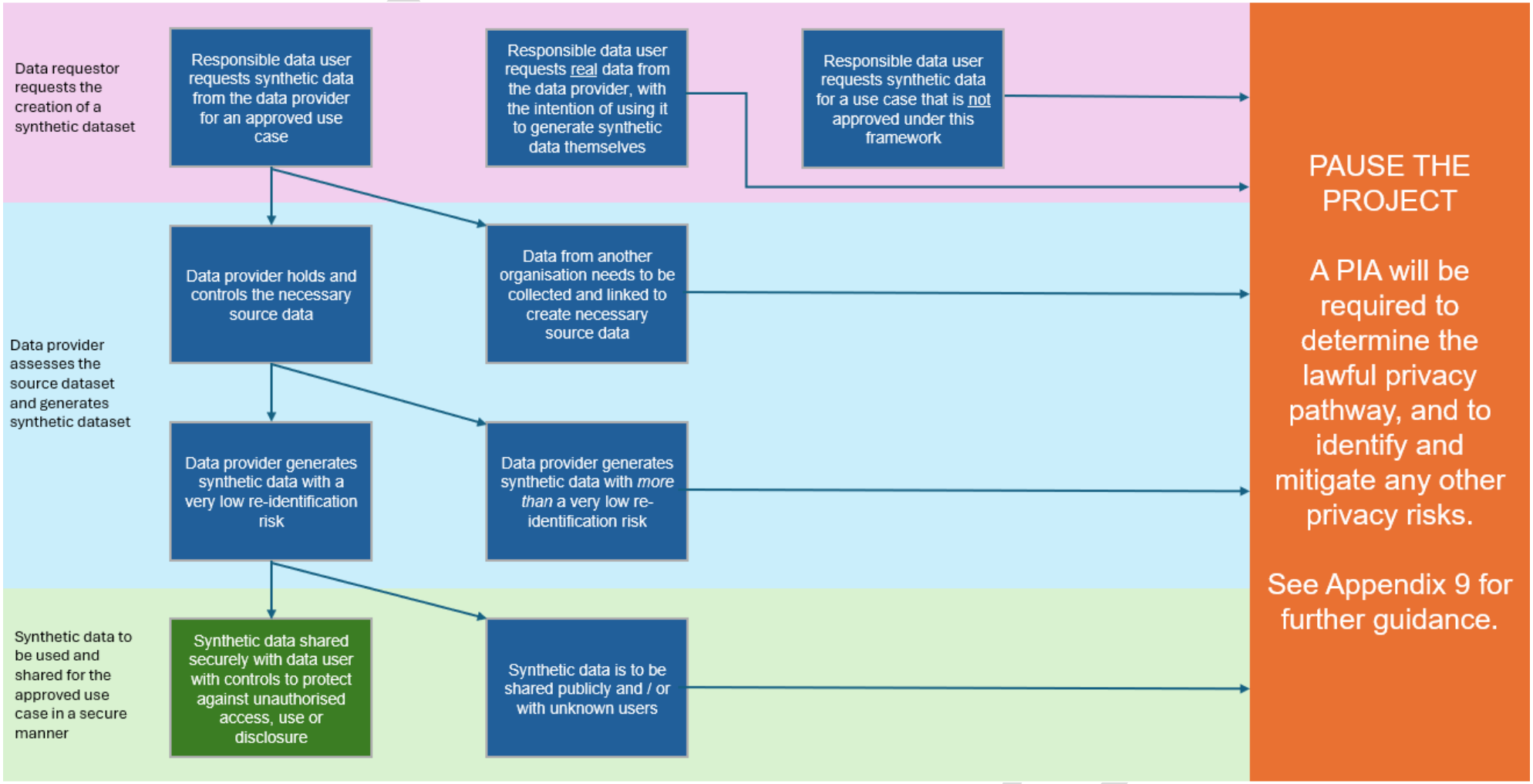
The five-step approach for generating and using synthetic health data outlined in this Framework should be sufficient for managing privacy risks where a synthetic health data request involves a fairly straightforward workflow.

A sample scenario of a straightforward workflow would be as follows:

1. A SynD organisation (e.g. the University of Sydney as the Data Requestor) requests another SynD organisation (e.g. NT Health as the Data Provider) to generate a synthetic health dataset using source data already held by that Data Provider, for an acceptable use case.
2. The source data is fit for purpose, and the resulting synthetic health dataset is considered robustly and effectively de-identified with only a very low risk of re-identification. The data utility is appropriate for the use case.
3. The Data Provider provides the Data Requestor with access to the synthetic health dataset in a secure manner and continues to have an appropriate level of oversight of the handling of the dataset.

However, for scenarios that involve more complex workflows, *or* where any of the steps in this Framework cannot be successfully completed, material privacy and legal compliance risks may emerge. In these cases, additional steps are needed to manage these risks.

The decision tree below anticipates some of these more complex scenarios where additional steps and assessments are required, to ensure the synthetic health data request can proceed lawfully.


Data requestor requests the creation of a synthetic dataset
Responsible data user requests synthetic data from the data provider for an approved use case
Responsible data user requests real data from the data provider, with the intention of using it to generate synthetic data themselves
Responsible data user requests synthetic data for a use case that is not approved under this framework
Data provider assesses the source dataset and generates synthetic dataset
Data provider holds and controls the necessary source data
Data from another organisation needs to be collected and linked to create necessary source data
Data provider generates synthetic data with a very low re-identification risk
Data provider generates synthetic data with more than a very low re-identification risk
Synthetic data to be used and shared for the approved use case in a secure manner
Synthetic data shared securely with data user with controls to protect against unauthorised access, use or disclosure
Synthetic data is to be shared publicly and / or with unknown users
PAUSE THE PROJECT
A PIA will be required to determine the lawful privacy pathway, and to identify and mitigate any other privacy risks.
See Appendix 9 for further guidance.

APPENDIX 9: The lawful pathways explained

While there are differences between the sets of privacy principles that operate in Australian jurisdictions, there are many commonalities, including how the principles both facilitate and restrict the sharing of personal information.

Generally, organisations are permitted to collect personal information where the information is necessary and relevant for their functions or activities (and in some instances, with the individual's consent); this is known as the 'primary purpose' for the collection. Organisations are then permitted to use and / or disclose that personal information for the primary purpose.

If the organisation then wishes to use or disclose the information for a different (secondary) purpose, it will need to satisfy at least one of a limited number of exceptions. Thus, only *some* secondary purposes are allowed.

Under privacy law, using and sharing health data about health consumers for synthetic health data projects will always be for a *secondary* purpose, as the primary purpose for collecting this information was to provide health care services to individuals (see 'Step 1: Assess the use case' above).

synthetic health data requests will generally involve handling personal information (specifically health information, which is a special subset of personal information) at two possible stages, each of which will require a lawful privacy pathway in order to proceed:

- When selecting and handling the source data from which synthetic health data will be generated (i.e. the source data will be considered personal information).
- When the re-identification risk associated with a synthetic health dataset is more than very low (i.e. in these circumstances, the synthetic health dataset will be considered personal information).

Multi-party projects

Where an organisation needs to collect personal information from other organisations in order to generate synthetic health data, the privacy and legal risks associated with these activities must first be identified and appropriately managed. This could include, for example, a scenario where one health department will disclose health information to another health department, which will create a linked health dataset for synthetic health data generation. Or it could include a scenario where a health department will disclose health information to a university research team, who will use it to generate a synthetic health dataset for research-related purposes.

In these cases, a Privacy Impact Assessment (PIA) must be completed in order to ensure that each organisation in these scenarios can *disclose* and/or *collect* the personal information under their own privacy obligations. The PIA should identify the most appropriate lawful privacy pathway/s for *each* participating organisation, and *each* part of the data journey.

The possible lawful pathways for using and disclosing personal information for a secondary purpose are explained below. Where organisations need to assess the possible lawful pathways that apply to their use case, they should also refer to the text of the Use and Disclosure privacy principles in the privacy laws that apply to them (see Appendix 3, The policy and legal framework underpinning this Framework, for a description of and links to the different privacy laws and privacy principles that could apply to an organisation).

'Directly related' and 'within reasonable expectations'

Privacy laws allow for health information to be used and shared outside of an organisation for a secondary purpose, if that secondary purpose is *directly related* to the primary purpose for which the information was collected. The secondary purpose must also be within the reasonable expectations of the individual, and in some cases, the organisation must have no reason to believe that the individual concerned would object to the use or disclosure.

For example, if information is collected in order to provide a health service to the individual, the use of the information to send an appointment reminder to the individual is for a secondary purpose that is directly related to the primary purpose, which an individual should reasonably expect.

The NSW Privacy Commissioner has advised that a directly related purpose "would be the type of situation that people would quite reasonably expect to occur with their personal information".[49]

Examples of disclosures of health information that the NSW Privacy Commissioner considers appropriate[50] under this test include:

- providing information to a person or organisation involved in the ongoing care of the patient
- investigating and managing adverse incidents or complaints about care or patient safety
- monitoring, evaluating, and auditing the provision of a particular product or service that the organisation has provided, or
- managing a legal claim made by the person.

[49] Privacy NSW, *A Guide to the Information Protection Principles*, 1999, p.35.
[50] NSW IPC, *Statutory guidelines on the management of health services*, p. 6.

If the individual has not been made aware (such as through a collection notice included on a patient form) that their personal information will be used or disclosed for the secondary purpose, there is a greater risk that they would not 'reasonably expect' the disclosure to take place, and this pathway may not be available.

Using or processing personal information for the purpose of de-identifying it to the extent it is no longer personal information may, in some circumstances, be considered a 'normal business practice' that is incidental or directly related to the primary purpose of collection.[51]

From a practical perspective, using health consumer information to generate synthetic health data can meet the 'directly related' and 'within reasonable expectations' where:

5. the use case is for a clear 'public benefit' purpose related to providing health services, and where the expected benefits from the use case are related to consumer health or health system outcomes;
6. the aim in creating and managing the synthetic health dataset is to achieve a 'de-identified' dataset for the use case, that significantly minimises the risk to individuals compared to if the source dataset was used for that use case; and
7. steps have been taken to set expectations with health consumers about how their health information will be used.

These requirements are reflected in the Use Case Assessment checklist in Appendix 4.

These conditions are explained in more detail above under 'Step 1', and form the basis for determining whether a particular synthetic health data use case can proceed under this Framework.

With consent

Personal information can be used or disclosed for a secondary purpose if an organisation has the consent of the individual/s to whom the information relates.

To be a valid consent under privacy law, such as to authorise a use or disclosure of personal information that would not otherwise be allowed, consent must be:

8. Voluntary (i.e. the individual opted in, had the opportunity to change their mind later, and has not since withdrawn their consent)
9. Informed (i.e. the individual was told about this data use proposal in a comprehensible manner before they chose to participate)
10. Specific (i.e. the consent was specific to this data handling proposal, not bundled in with other topics)
11. Current (e.g. given within the last two years), and

[51] See the OAIC's guidance, *De-identification and the Privacy Act*, March 2018, available at: https://www.oaic.gov.au/privacy/privacy-guidance-for-organisations-and-government-agencies/handling-personal-information/de-identification-and-the-privacy-act

12. Given by a person with capacity (e.g. parents can consent on behalf of children).

To rely on consent as a lawful pathway, the individual's consent must specifically cover the use or disclosure for the purpose being proposed, with transparency around which organisations will be handling their information.

From a privacy risk management perspective, express consent is typically easier to establish and evidence than implied consent (which relies on non-action or silence). This means an individual has been given the option to consent, and responds affirmatively – for example, by ticking a box, signing a consent form, or saying 'yes'. Organisations seeking to rely on 'with consent' as their lawful pathway also need mechanisms in place to manage consent – for example, to record consent and to have processes in place where an individual wishes to withdraw their consent.

Many data assets containing (or derived from) consumer health information are *collected* with the consent of the individual. However, if the individual did not consent at that time for their information to later be *used* or *disclosed* for a specified purpose, the organisation wishing to use or share the information for a synthetic health data project will need to either obtain the individual's specific consent to use and / or share their information or will need to rely on an alternative lawful pathway to authorise the use or disclosure.

From a practical perspective, obtaining and managing valid consents is not a suitable lawful pathway to pursue for large-scale synthetic health data projects, particularly where organisations wish to use or share data that has already been collected from individuals.

Seeking ethics approval: management of health services & research projects

Most privacy laws carve out special exceptions for using and sharing health information for purposes related to the management of health services and / or research. Although not well defined, privacy laws draw a distinction between 'research', and other activities, including management of health services, although the legal tests for each purpose are often similar, and both commonly require obtaining approval by an appropriate Human Research Ethics Committee (HREC) (depending on the jurisdiction).

When seeking approval from an HREC, organisations must ensure all aspects of the proposed information flows associated with a synthetic health data project are approved appropriately by the HREC. This includes any *disclosure* of information by one organisation to another organisation that is *collecting* it to prepare source data for synthetic health data generation, as well as the subsequent *use* of the information to generate the synthetic health data.

Sharing of data across jurisdictions and sectors with HREC approval relies on meeting tests set out under multiple pieces of privacy law and different statutory guidelines. In order to

seek a *single* ethics review for a project that involves data sharing across multiple jurisdictions and sectors, the organisations should apply to a registered HREC that has been nationally certified, and is recognised under the National Mutual Acceptance (NMA) scheme. This is because the NMA scheme (in which all state and territory jurisdictions participate) supports a single ethical review for multi-centre projects across state and territory public sector organisations.

Where organisations need to *share* real health data in order to prepare a synthetic health dataset, or where a synthetic health dataset is still considered 'personal information' due to the level of re-identification risk, seeking approval and a waiver of consent from an HREC will most likely be the most appropriate and practical lawful pathway to support the use and sharing of this data (although this will depend on the jurisdiction and the other lawful pathway options that may be available to an organisation).

This is the same lawful pathway that many organisations would most likely follow when seeking to collect and handle *real* data about people for their projects (e.g. for a clinical research project).

Required or authorised under another law

A collection, use or disclosure of personal information may be required or authorised under another law. If such an action is 'required' under law, the organisation handling the personal information would have no choice but to handle the information as directed under the applicable legislation. (An example is when an investigative or law enforcement body uses its compulsion powers to compel production of documents.)

If such an action is instead 'authorised' under another law, the organisation would be permitted to handle the information as set out in the legislation but would have a choice as to whether or not they do so.

The OAIC also describes the need for clear and direct language when seeking to rely on this exception as a lawful pathway:

> "An act or practice is not 'authorised' solely because there is no law or court/tribunal order prohibiting it. Nor can an act or practice rely solely on a general or incidental authority conferred by statute upon an agency to do anything necessary or convenient for, or incidental to or consequential upon, the specific functions and powers of the agency. The reason is that the purpose of the APPs is to protect the privacy of individuals by imposing obligations on APP entities in handling personal information. A law will not authorise an exception to those requirements unless it does so by clear and direct language.[52]"[53]

[52] See *Coco v The Queen* (1994) 179 CLR 427.
[53] OAIC, *Australian Privacy Principles (APP) Guidelines*, December 2022, B.135.

This pathway could be relevant where the principal laws that govern the functions or services of an organisation require or authorise a particular collection, use or disclosure of information that is applicable for a synthetic health data project. Other laws may also authorise using and sharing data for secondary purposes under certain circumstances, for example, the *Data Availability and Transparency Act 2022* (Cth) (DAT Act).

Effectively de-identified to be safe for sharing

Only robustly and effectively de-identified data will be considered safe to share under this pathway, which is one of the key aims in creating and handling synthetic health datasets with a very low re-identification risk.

In the context of privacy law, the term 'de-identification' must be understood by reference to the meaning of 'personal information'. 'De-identified' data therefore means that a person's identity is no longer apparent, or cannot be reasonably ascertained, following the application of one or more de-identification techniques to 'personal information'.[54] Examples of de-identification techniques are discussed in more detail in Appendix 7.

In theory, this means that 'de-identified' data is no longer 'personal information' for the purposes of regulation by privacy laws.

However even if direct identifiers such as name and address, or individual healthcare identifiers or unique reference numbers, are removed from a data set, a person's identity may still be 'reasonably ascertainable'. In fact, the Australian Privacy Commissioner has warned that de-identification "can be effective in preventing re-identification of an individual, but may not remove that risk altogether", for example if "another dataset or other information could be matched with the de-identified information".[55]

This means that if the surrounding context, and other available information used in combination with the data to be shared, could be used to ascertain a person's identity, the data should be assumed to be re-identifiable. In other words, the data should still be considered 'personal information', and privacy protections applied accordingly.

Further, 'identifiability' in law does not necessarily imply that a person's name or legal identity can be established from the information. The Australian Privacy Commissioner has said that:

[54] See the NSW Information & Privacy Commission Fact Sheet 'De-identification of personal information', 2020, https://www.ipc.nsw.gov.au/fact-sheet-de-identification-personal-information

[55] OAIC, *De-identification of data and information*, April 2015; was previously available at https://www.oaic.gov.au/information-policy/information-policy-resources/information-policy-agency-resource-1-de-identification-of-data-and-information but has since been replaced by OAIC, *De-identification and the Privacy Act*, 21 March 2018, available at https://www.oaic.gov.au/privacy/guidance-and-advice/de-identification-and-the-privacy-act/

> "Generally speaking, an individual is 'identified' when, within a group of persons, he or she is 'distinguished' from all other members of a group."[56]

De-identification should be seen as a methodology to manage 'personal information', in order to *reduce* the likelihood of any individual being identifiable. Data should not be considered completely 'de-identified' (and thus outside the regulation of privacy laws) unless it has been tested for re-identification risk, and found not to pose such a risk. It is also important to note that some privacy laws (e.g. the WA PRIS Act) still regulate some aspects related to the handling of de-identified data (such as limiting the transfer of de-identified information outside of Australia, ensuring the security of de-identified information, and prohibiting the re-identification of the data).

Only data that has been de-identified to the point that there is only a very low chance of re-identification or 'singling out' will it be considered 'safe' to use or share under this lawful pathway.

It should also be noted that even where health information has been effectively de-identified to be safe for use and / or sharing, where it is being used for a research project, additional ethical considerations may need to be made even if the research may be eligible for 'lower risk research' ethics review pathways on the basis no personal information will be used.[57]

This Framework seeks to rely on this pathway to support various use cases involving synthetic health data on the basis it has been de-identified to the point of no longer being 'personal information'. If synthetic health data has not been robustly and effectively de-identified, this pathway will not be suitable to support the use or sharing of synthetic health data, and an alternative pathway must first be settled (e.g. seeking ethics approval and a waiver of consent from an HREC).

Organisations should be aware that data considered 'de-identified' in one context may not remain de-identified in another. For example, unit record synthetic health data in a restricted, protected environment (such as a secure data enclave) and subject to specific governance controls may be considered 'effectively de-identified' within that environment and context. However, if the same data were published and made publicly available – or if it was subject to a data breach – the risk of re-identification may increase, and the data may become 'personal information' again (and so will become subject to privacy obligations again).

Other factors may also impact or heighten the risk of re-identification, such as if the source dataset (or a closely related dataset) and / or the trained model used to generate the synthetic health dataset were exposed (e.g. were made available on the dark web) or made available to users who have access to the synthetic health dataset. As such, re-

[56] Office of the Australian Information Commissioner, *What is personal information?*, May 2018. p.8, available at https://www.oaic.gov.au/agencies-and-organisations/guides/what-is-personal-information

[57] 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

identification risk should not be considered 'static'. Organisations will need to re-assess re-identification risk where there are changes relating to the data, the environment, or other relevant factors.

Other exceptions

Privacy laws may provide other exceptions which support handling personal information for synthetic health data projects. For example, the NSW Privacy Commissioner can issue a temporary Public Interest Direction (PID), or the NSW Minister of Health can issue a permanent Health Privacy Code of Practice (HPCOP) that may also authorise organisations subject to the NSW HRIP Act to collect, use or disclose health information in a manner not otherwise permitted under the NSW HPPs. Similar schemes are available under privacy laws in Queensland, the Northern Territory, Tasmania and Western Australia, as well as the Privacy Act. However, Victoria (with respect to health information) and the ACT do not have these mechanisms.

Privacy laws also provide other exceptions which can authorise the collection, use and disclosure of personal information, although they will not be relevant for synthetic health data projects. These exceptions include, for example, for law enforcement purposes, where there is a serious threat to life, health or safety, and in the cases of missing persons. Specifically for the handling of health information, there are also exceptions which pertain to the provision of healthcare services to the individual about whom the information relates. For example, there are exceptions that allow a health service provider to share the patient's health information amongst a treatment team, or with family members in certain circumstances.

From a practical perspective, the range of other exceptions which can authorise organisations to collect, use or disclose health information in a manner not otherwise permitted under their privacy obligations are not relevant for synthetic health data projects – particularly where privacy laws already have special exceptions designed for research projects and management of healthcare activities.

Further Resources

- OAIC Australian Privacy Principles Guidelines (see in particular 'Chapter 6: APP 6 – Use or disclosure of personal information')
- See also discussion in Appendix 5 (Impact Assessment) in relation to Aboriginal HRECs
- NHMRC Guidelines approved under Section 95A of the Privacy Act 1988
- NHMRC National Statement on Ethical Conduct in Human Research 2025
- NSW IPC Statutory Guidelines on Research – HRIP Act
- NSW IPC Statutory Guidelines on the management of health services – HRIP Act

- See Appendix 3, *The policy and legal framework underpinning this Framework*

APPENDIX 10: Safety Assessment

Safety obligations when sharing data

Legal requirements to protect data

The SynD organisations have legal obligations to maintain data quality, manage data security, ensure data is disposed of appropriately, and take accountability for the data shared internally or externally.

These legal obligations come from the Commonwealth and State and Territory privacy laws, as well as the range of State Records and Archives laws.

If personal information is leaving a state or territory

If data about identifiable individuals is to be disclosed to a recipient outside of the state or territory where it is currently held, the organisation must take reasonable steps to ensure that the information will not be held, used or disclosed by the recipient of the information inconsistently with the organisation's legal privacy obligations.[58] This is typically satisfied by way of a contract / agreement or confidentiality deed.

Data about identifiable individuals (such as source data, or synthetic health data that is still 'health data') should be held by the organisation who collected it as a rule (including when stored on secure cloud platforms) unless the SynD organisations have completed a risk assessment and agreed on an alternate storage location.

Keeping data secure in transit

There are multiple techniques for 'sharing' data: data might be 'shared' by uploading data using a secure file transfer protocol or a system such as Kiteworks, building a dashboard report via an API, providing a Data Requestor with direct access to a data warehouse, or by providing a Data Requestor with a data extract.

However, whether or not a particular technique can be considered 'safe' will depend on the context. Some techniques, such as emailing records without further security safeguards such as encryption or password protection, will not be considered safe. The security of the proposed method or technique for sharing data should be assessed prior to the sharing taking place to ensure it is suitable and meets any additional organisational requirements.

Disposing of data appropriately

The rules for data retention and disposal are set by the State Records and Archives laws and applicable privacy laws working together. Organisations must set retention periods for any synthetic health datasets generated under this Framework. As soon as possible after that period has expired, the data must be deleted or securely destroyed from all production and non-production environments (unless going into permanent government archives).

[58] IPP 12 (s.19(2)(g) of the PPIP Act), and HPP 14

When sharing synthetic health data amongst the SynD organisations, the Data Provider will need to work with the Data Requestor and specify how long the Data Requestor is allowed to keep the data, and what their obligations are in relation to either returning or destroying the data at the end of that period.

Assurance

Once synthetic health data is approved for sharing, steps should be taken to ensure that the conditions required under the Data Sharing Agreement are being met, and to address any deficiencies. Types of assurance activities that could be planned and carried out include:

- Reviewing relevant current security certifications or system specifications
- Confirming that data is stored and maintained in the approved system and that IT controls are in place and effective
- Verifying any required training has been conducted (e.g. privacy and security training for End Users, or data analytics training)
- Ensuring that access to the data is being managed appropriately, and that End User access is being revoked in a timely manner when no longer need
- Checking data access logs periodically for any unusual behaviour or unauthorised accesses
- Checking any outputs arising from the analysis to confirm they are aligned with the approved use case
- Confirming that any required assessments (e.g. technical security assessments) have been completed and the results are acceptable
- Confirming that any third-party service or supplier arrangements are appropriately managed
- An audit of an organisation's compliance with the Data Sharing Agreement

Organisations with formal assurance policies and processes should apply these to the data sharing arrangement.

Taking a 'five safes' approach to generating, using and sharing synthetic health data

Every instance of generating and handling synthetic health data carries some privacy risks, so the benefits of each use case need to substantially outweigh those risks.

The Five Safes Framework[59] offers a useful way of thinking about how to control for *two* particular types of privacy risks:

[59] Australian Computer Society, *Data Sharing Frameworks: Technical White Paper*, September 2017; available at https://www.acs.org.au/content/dam/acs/acs-publications/ACS_Data-Sharing-Frameworks_FINAL_FA_SINGLE_LR.pdf

- Inadvertent disclosure (also known as 'statistical disclosure'), such as *re-identification risk*, and
- Misuse of data by authorised users (data recipients) – in other words, *misuse risk*.

The Five Safes Framework was designed to manage these two types of privacy risk *in a particular context*: the sharing of data in a controlled environment, in which a party will perform analytical operations on the data, and then share the results of the analysis.

The Five Safes Framework is not a legal requirement, and it does not override our legal obligations. It is not a way of measuring the level of 'identifiability' of data. The Five Safes Framework is not an assessment of whether it is lawful, or ethical, to engage in generating and handling synthetic health data. That is why working through *all* steps and assessments in this Framework is critical. However, the Five Safes Framework is useful as a way of thinking about risk management when handling data.

Each 'safe' refers to an independent but related aspect of managing these two types of privacy risk, in the context of handling data in a controlled environment:

- Safe data – Has appropriate and sufficient protection been applied to the data?
- Safe projects – Is the data to be used for an appropriate purpose?
- Safe settings – Does the access environment prevent unauthorised use?
- Safe people – Is the requestor or end user appropriately authorised to access and use the data?
- Safe outputs – Are the statistical results non-disclosive?

These five elements are intended to be viewed wholistically to create an *overall* level of safety, in which the different elements may balance each other out. In other words, if one type of control has been 'dialled up', it may be safe to 'dial down' another control.

The type of synthetic health data project and its expected outputs will have an impact on which risk management controls can be dialled up or down. For example, when releasing data to the public at large, we must assume there is zero safety in terms of the 'people' or 'settings'. In such cases, to achieve an overall level of risk management to enable safe sharing, it will be critical to 'dial up' other elements such as the safety of the data, through extremely stringent de-identification techniques.

The Safety Assessment Checklist below reflects the different contexts in which data sharing might take place, while still using the broad 'five safes' concept.

Safety Assessment Checklist

The accountable decision-maker must assess the request in consideration of the following:

Considerations:	
Safe data	Has the data been presented so that it can be clearly understood, appropriately footnoted, with data sources acknowledged? Will the integrity and quality of data be maintained by the Data Requestor? For example, will the data remain segregated from other data held by the requestor, and data lineage (including that the data originated at the Data Provider) maintained? Has re-identification risk been tested? Are there any data fields which could raise 'safe data' concerns, due to their identifiability? Examples could include: • Date of birth • Combinations of demographic fields such as age, gender, postcode, Aboriginality, ethnicity or health status • Attribute data that could lead to re-identification by rendering an individual unique in the dataset, e.g. geolocation data, or longitudinal data • Text-based fields / unstructured data
Safe projects	Has a Privacy Impact Assessment (and any other relevant assessment such as a data linkage assessment) been completed, where applicable? Is the data to be used for an appropriate purpose? Reasons for concern could include if: • There is a Data Owner outside the organisations (e.g. data provided by another organisation) • The data was indirectly collected, generated or inferred • A new metric or indicator needs to be built for this proposal • The synthetic health data request does not have a clear objective or methodology
Safe settings	What controls will be in place to prevent unauthorised access or unauthorised use? For example, the accountable decision-maker must be satisfied that: • The Data Requestor can meet all of the Data Provider organisation's requirements including security requirements

	• The data will be transferred, stored, managed and disposed of securely and appropriately, to prevent unauthorised or accidental access, modification, loss, and damage or copying. Systems should allow for user access to be controlled, monitored and audited. • The system within which the data will be used and stored, and any transfer mechanisms, must be subject to a technical security assessment to ensure it is sufficiently secure in the circumstances. What is considered 'sufficiently secure' should take into account the type and format of data, the level of privacy risk associated with the data, and the potential impacts (both legal and non-legal) in the event of a data breach or misuse. For example, unit record data with a 'more than very low' re-identification risk may need to be stored and accessed via a trusted research environment or secure data enclave. Other types of secure storage options may be suitable where data only has a very low re-identification risk, and other reasonable safeguards are in place. • The data has been labelled according to any relevant data classification policies • Obligations relating to the return or disposal of the data have been agreed • If approved for sharing, there will be a legally binding Data Sharing Agreement (DSA), Data Use Agreement (DUA), or other appropriate written agreement in place between the organisations, and with End Users, with appropriate confidentiality and privacy provisions. (Organisations should seek legal support on questions regarding DSAs and DUAs. See discussion on DSAs and DUAs below.) • Where contracted service providers are involved in the synthetic health data project (for example, a cloud platform provider), the contracting organisation has a service contract in place with appropriate data and privacy protection clauses, and carries out appropriate onboarding and oversight of the contracted service provider's performance. • The data must be held in Australia, and by an organisation that is subject to legal privacy obligations and oversight by a body (e.g. a statutory regulator such as a Privacy Commissioner or Information Commissioner) who can enforce these obligations. As an example, private sector organisations (with an annual turnover of less than $3 million) and South Australian public sector agencies are not subject to legal privacy obligations. Any proposed exception to this requirement must first undergo a legal risk

	assessment, and appropriate clauses in the Agreement must be included.
Safe people	Has the Data Requestor provided a list of people who will be authorised to access / use the data, and those people have been approved? • Consider whether an end user or recipient of the data possesses specialised skills or technology, or has access to relevant data, which may increase the risk of re-identification. How can these risks be controlled?
Safe outputs	How will the Data Requestor use or publish the data later, or outputs from analysis? What guarantee is there that their actions will not lead to re-identification of individuals, or other forms of harm to cohorts or the community?

Data Sharing Agreements and Data Use Agreements

Following the completion of the steps and assessments required under this Framework, and where the accountable decision-maker has approved the synthetic health data request, the participating organisations (i.e. the Data Provider(s) and Data Requestor(s)) will need to enter into a **Data Sharing Agreement (DSA) prior any data being shared**. Entering into a DSA has the benefit of documenting what the organisations agreed to regarding the synthetic health data request and setting out requirements for data handling and security. DSAs can also contain clauses around ongoing assurance and enforcement rights once the data has been shared (e.g. whether the Data Provider has a right to inspect or audit compliance by the Data Requestor).

End Users at the Data Requestor organisation who will access the synthetic health dataset for the approved use case will also be required to complete a **Data Use Agreement (DUA) before they are granted access to the data**. A DUA can be used to notify End Users of their responsibilities and obligations when accessing and using synthetic health datasets.

While a legal expert should be involved in drafting DSAs and DUAs, organisations should consider incorporating the following:

- Identifying the parties involved in the data sharing
- Agreement expiry date
- A description of the synthetic health data to be shared under the agreement, including the nature of the data and the source dataset from which it is drawn. This should include a list of specific data fields (and may be contained in data specification document attached as an appendix)
- Confidentiality clauses

- Requirements that parties will comply with Australian and any relevant state or territory privacy legislation to the extent the shared data includes personal information
- Approved purposes for which the data may be used and shared by the parties
- Details of the sharing mechanism to be used, including frequency
- Details of the security requirements for transfer and storage of data (including any relevant standards or policies)
- Data storage location – expected locations may be the ACT, NSW, Northern Territory, Queensland, Tasmania, Victoria and Western Australia (Western Australia will only be a suitable location after the substantive privacy provisions in the PRIS Act become effective in July 2026, unless the source data originated in Western Australia), unless parties agree on another location (which must be subject to a legal risk assessment first)
- Restrictions on who may access the data (i.e. limited to authorised and identified individuals based on their roles or functions within the organisation)
- Conditions on use, release, or publication of the data
- Restrictions on using synthetic health data, or combining synthetic health data with other data, in a manner that could reasonably re-identify an individual or that generates personal information about an individual or otherwise increases the risk of re-identification
- Prohibitions on attempts to re-identify the data
- Conditions on the use or release of outputs (for example, outputs from data analysis, results, insights, statistics, or other information or data generated from the synthetic health data), including any review or approvals from the participating organisations, and whether the data should be identified as originating from the Data Provider. The Agreement should specify who is responsible for the management, storage and the destruction of the outputs. Outputs should not include any personal information, or other information that could reasonably be used to identify an individual (either alone or in combination with other information or knowledge)
- Requirements around sharing the data with (including providing access to) a third party, including notification and/or approval by the parties.
    - For example: if a third-party service provider (e.g. a cloud platform provider) will have access to the shared data, requiring service contracts between a party to the Agreement and the service provider to first be in place which contain appropriate privacy protections (including restrictions on any use of the data, security obligations, and data breach and incident reporting responsibilities).
    - Also, where a party is required by law to disclose the data to a third party (for example, to comply with a court order), they may need to (where lawful) first notify the other parties of the disclosure.

- Obligations relating to the quality of the data to ensure it is fit-for-purpose. For example, that the data is provided in the agreed format; that it is accessible; and that it meets the agreed description in terms of accuracy, completeness, reliability and currency. Data being shared should be limited to what is necessary to achieve parties' objectives in line with approved purposes. Parties may wish to include a requirement that records be maintained about what data was provided and/or combined for traceability and verification purposes – such as source systems, description, date of extraction, etc. Parties must notify the Data Custodian and/or Data Providers of any data quality issues.
- Obligations relating to the return or disposal of the data, including by what date. A Data Retention Schedule can be useful for setting out retention timeframes, and obligations for data retention and destruction (for example, if certain datasets must be retained for certain timeframes to meet legal requirements).
- Obligations relating to reporting / escalation pathways in relation to any privacy incident, inadvertent inclusion of personal information in the dataset, privacy complaint, access or correction request, or suspected or actual data breach. This includes incidents or breaches involving the model (or aspects of the model) used to generate the synthetic health dataset. The Agreement should specify which party is responsible for managing these events.

A DUA for End Users will not typically need to be as detailed as the overarching DSA, and will be a one-way agreement or acceptance of certain terms and conditions completed by the End User. DUAs may incorporate:

- A statement / declaration agreed to by the End User about accepting certain responsibilities and obligations regarding the data
- Relevant elements from a DSA, such as:
    - Confidentiality clauses
    - Restrictions on accessing and using the data only for an approved purpose and during an approved timeframe
    - Restrictions or conditions on how outputs from analysis can be handled (e.g. releasing or publishing outputs)
    - Compliance with any relevant policies or guidelines when accessing the data, such as security policies that apply to devices and / or relevant ethical guidelines
    - Requirements to escalate an incident to an appropriate designated person
    - Prohibitions on activities which may impact re-identification risks

If the SynD organisations wish to develop a template DSA and DUA, the above suggested content can be removed and the Framework can instead point to these templates.

Further resources

Relevant organisational policies that may need to form part of an organisation's safety assessment can be set out / linked to here

E.g. information security policies; data governance policies; data breach and incident management policies, data release policies, acceptable IT use policies, etc.

APPENDIX 11: Synthetic health data request and assessment outcomes form

Request and Assessment Outcomes Form

Project name:	
Project description:	
Who is the Data Requestor organisation?	
Who is the responsible data user / lead at the Data Requestor?	
Who is the Data Provider organisation?	
Who is the accountable decision-maker at the Data Provider who will review / approve the synthetic health data request?	
Any other key personnel involved in gathering information and / or conducting the assessments required under this Framework:	

Who is responsible? *DR = Data Requestor, DP = Data Provider*	
Step 1: Use Case Assessment	
DR	1. Describe the use case
DR / DP to confirm	2. Is this use case for a clear 'public benefit' purpose related to providing health services, and where the expected benefits from the use case are related to consumer health or health system outcomes?

DR / DP to confirm	3. Do you anticipate a synthetic health dataset will be suitable for this use case? *With a synthetic health dataset, there must only be a very low risk that the dataset can identify or disclose information about individuals.*
DP	4. Has the Data Provider communicated publicly and broadly to their health consumers that they will use synthetic health data about consumers for public benefit projects, such as those related to improving health outcomes for consumers and for the health system?
DP	5. Is the proposed use case acceptable under this Framework? *Data Provider must complete the **Use Case Assessment** (Appendix 4) to assist with answering this question.*
DP	6. Are there other impacts or ethical reasons that mean the request should not proceed? *Data Provider must complete the **Impact Assessment** (Appendix 5) to assist with answering this question.*
Step 2: Assess and prepare source data	
DP / with assistance from the DR	7. What data is being requested and is it available? *Identify and describe the data elements or datasets being requested, including the type of data and any other parameters or characteristics (e.g. geography, date/time ranges, any important features or limitations). A data specification form may be used.*
DP / with assistance from the DR	8. Is the source data being requested relevant for the use case?
DP / with assistance from the DR	9. Does the Data Provider hold and control the source data? *If 'no', the request should be considered 'complex' and will require further assessment and action before the request can proceed under this Framework. See Appendices 8 and 9 for further guidance.*

DP / with assistance from the DR	10. Does the source data need to be enriched or linked to datasets held by *other* organisations? *If 'yes', the request should be considered 'complex' and will require further assessment and action before the request can proceed under this Framework. See Appendices 8 and 9 for further guidance.*
DP	11. Are there any other limitations or restrictions on using the source data to generate synthetic health data?
DP	12. Have you identified the system and/or repository that the data will need to be extracted from?
DP	13. What format will be used for the synthetic health data?
DP / with assistance from the DR	14. Are you satisfied that the data being requested is the minimum amount of information needed for the use case? *A subset of the source data may need to be prepared to only include what is needed to generate the synthetic health dataset.* *All data and fields containing directly identifying information (such as names, addresses, phone numbers, date of birth, date of death, unique identifiers such as patient numbers, Medicare numbers or drivers licence numbers) must be removed or appropriately modified or obscured to reduce the risk they will be 'leaked' via the synthetic health dataset. If these fields cannot be removed or appropriately modified or obscured, organisations must be aware of the risk of data leakage and the potential for heightened re-identification risk that must be assessed and managed prior to any use or sharing.*
DP	15. Have you created a data quality statement to be supplied to the Data Requestor that addresses the accuracy, completeness, reliability and currency of the source data which will be used to generate the synthetic health data?

DP	16. Can you provide metadata and/or other material (such as a data dictionary) to help the Data Requestor to understand the nature of the source data and the resulting synthetic health data?
DP / with assistance from the DR	17. Do the Data Requestor's personnel possess the technical requirements and knowledge to effectively use the data for the identified purpose?
DP / with assistance from the DR	18. Is the source data 'fit for purpose' for the use case? *Data Provider must complete the **Technical Assessment** (Appendix 6) to assist with answering this question.*
Step 3: Generating the synthetic health data	
DP / with assistance from the DR if needed	19. What model will be used to generate the synthetic health data?
DP / with assistance from the DR if needed	20. Who will generate the synthetic health data? *If third-party expertise is required, specify the third party and the arrangement (e.g. expertise provided via a contracting arrangement with the Data Provider).* *If the Data Provider will release the source data to another organisation (whether or not that is the Data Requestor) to generate the synthetic health data, the request should be considered 'complex' and will require further assessment and action before it can proceed under this Framework. See Appendices 8 and 9 for further guidance.*
DP	21. Have you documented the model and details of the parameters used to train the model?

DP	22. After the synthetic health data has been generated, will the model be stored separately in a secure manner or otherwise destroyed?
Step 4: Assess and manage re-identification risks	
DP	23. Has the dataset been reviewed and treated for re-identification risks? *See Appendix 7 for further guidance on de-identification techniques.*
DP / with assistance from the DR if needed	24. Has a Re-Identification Risk Assessment been completed? What was the resulting level of re-identification risk? *Where the risk level is more than very low, the request should be considered 'complex' and will require further assessment and action before the request can proceed under this Framework. The project must be paused until a lawful pathway to proceed has been determined. See Appendices 8 and 9 for further guidance.*
DP / with assistance from the DR if needed	25. Has a Data Utility Assessment been completed? What was the outcome and can the use case proceed? *If the Data Utility Assessment indicated the utility of the synthetic dataset was too low for the use case, a new synthetic dataset with different variables or parameters may need to be generated. If this is the case, the Data Provider will need to re-do the Re-Identification Risk Assessment at step 24 and determine the resulting level of re-identification risk.*
Step 5: Manage residual privacy risks	
DP & DR	26. Where and how will the data be stored? *For example:* • *In the Data Provider's storage system (on-premises)* • *In the Data Requestor's storage system (on-premises)* • *In a storage system provided by a third party (e.g. third-party cloud platform provider) (specify the third party, the storage system and which organisation is responsible for the system)*

DP & DR	27. Who is responsible for the security of the storage system (including access management)? Provide details.
DP or DR	28. Describe the data security measures that will be in place to protect the data (provide details) *Describe the security measures that will be put in place to protect the data during storage and access, including from misuse, interference and loss, as well as unauthorised access, modification or disclosure.* *Measures can include technical and organisational controls. For example/if applicable:* • *access is limited to those who have been approved by the Data Custodian at the organisation that will hold the synthetic health data* • *periodic audits of who has access, and removal when access is no longer warranted* • *user login credentials and minimum password requirements* • *maintenance of access logs and audit trails* • *any additional authentication measures* • *data encryption* • *data to be handled in accordance with information security policies (name and link to (if possible) any key policies)* • *system restricts users from downloading or saving data to local drives (if applicable)* • *staff training*
DP or DR	29. In which state or territory will the data be stored? *If the data will be stored in a system held or controlled by a SA public sector agency, a WA public sector agency prior to 1 July 2026, or held by a private sector entity with an annual turnover under $3 million, a legal risk assessment must first be carried out and appropriate contract clauses must be used in the Data Sharing Agreement to ensure compliance with applicable privacy principles (see Appendix 3, 'The policy and legal framework underpinning this Framework' for further information).*

DP	30. What mechanism will be used for sharing the data securely with the Data Requestor? *For example:* • *Secure online data transfer site (e.g. Secure file transfer protocol/SFTP)* • *Secure end-to-end transfer system with password protection (e.g. OneDrive)* • *Other (specify)*
DP and DR	31. Will any third parties (including contracted service providers) be permitted access to the data? If third parties will be permitted access, describe the controls in place to protect the data. *For example, vendor due diligence prior to onboarding, appropriate privacy and security contractual clauses (including data breach notification requirements), ongoing performance and relationship management*.
DR	32. What are the expected outputs from the analysis of the synthetic health dataset? Will the outputs be shared outside of the Data Requestor organisation?
DP or DR	33. How long will the synthetic health dataset be retained? Provide a rationale for the retention period. *Retention periods could include:* • *For the duration of the project* • *For a specified time after the end of the project* • *Indefinitely (include a justification/rationale for the retention period)* *Specify any legal obligations to retain data and the required timeframes*
DP or DR	34. What will happen with the data at the end of the retention period? *For example, data to be destroyed, returned, retained (or 'other') at the end of the retention period. Include any other details or requirements regarding destruction or return of data (e.g. such as disposal methods, transfer*

	methods for return, evidence or attestation of destruction, or applicable policies or standards).
DP or DR	35. What controls will be put in place to protect the data after sharing? *Examples include:* • *Data Sharing Agreement* • *Data Use Agreements* • *Assurance activities, e.g. audit of compliance with Data Sharing Agreement, compliance attestations, etc.*
DP	36. Are you satisfied that the synthetic health data can be safely shared in the circumstances? *Data Provider must complete the **Safety Assessment** (Appendix 10) to assist with answering this question.*
DP and / DR	37. Any additional comments, conditions or recommendations regarding the synthetic health data request:
DP	38. Is the synthetic health data request approved?

To be completed by the accountable decision-maker at the Data Provider			
Signature:		Date:	
Name / title			

Document Control		
SynD Synthetic Health Data Governance Framework		
Version Number		v1.0
Effective Date		XX/XX/XXXX
Date of last review:		XX/XX/XXXX
Owned by	Digital Heath CRC	
Review Schedule	*The document should be reviewed periodically by SynD. Any changes or proposed amendments to the Framework should be circulated with all SynD and participating organisations, to ensure all organisations are working from the same version of the Framework.* *If the Framework has not been reviewed or updated within a **two-year period**, there is a risk the laws and policy frameworks which underpin the Framework may have been amended and the guidance in the Framework may no longer be up to date. In these circumstances, Digital Heath CRC should initiate a review of the Framework for currency and relevancy.*	