About this Framework

Scope

This 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 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 data can be realised while ensuring the associated privacy risks are identified and managed.

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

Creating, using and sharing synthetic data can raise questions not only about whether these activities are carried out lawfully, but also the reliability of the synthetic 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 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 data can be granted:

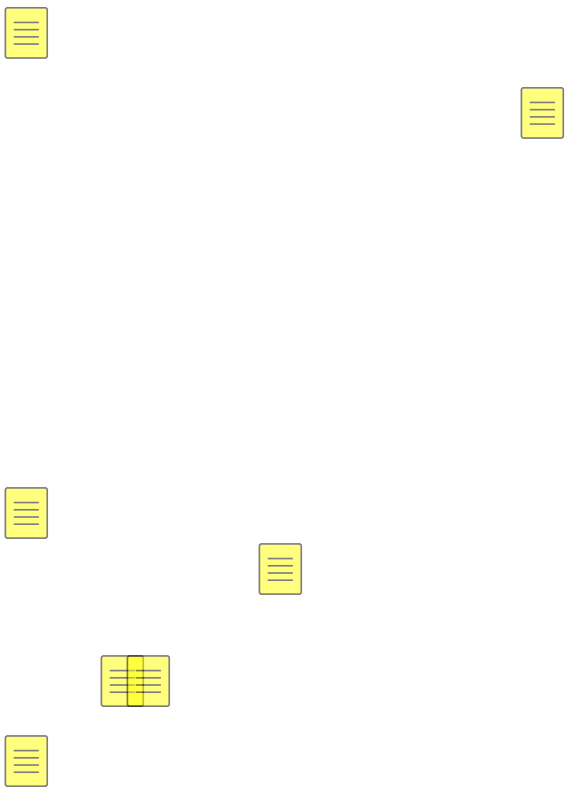
1. Assess the use case	2. Assess and prepare source data	3. Generate synthetic data	4. Assess re-identification risks	5. Take steps to manage residual privacy risks

Steps required under this framework	Assessments to be completed
Step 1: Assess the use case *This step uses a risk-based approach for data custodians approving the creation of synthetic data and / or synthetic data use cases based on legal and ethical considerations, and explains different lawful pathways for progressing high risk and / or complex synthetic data requests*	☐ Use case assessment (Appendix 4) ☐ Impact assessment (Appendix 5) • Decision tree for complex synthetic data requests (Appendix 8) • Further guidance on different lawful pathways (Appendix 9) • Privacy obligations regulating the use and disclosure of health information (Appendix 12)
Step 2: Assess and prepare the source data This step helps *Data Custodians answer the question – 'is this data fit for purpose?'*	☐ Technical assessment (Appendix 6)
Step 3: Generate the synthetic data *A range of synthetic data generation methods may be suitable under this framework. Data Custodians, with support from Data Scientists and synthetic data experts, will need to determine a suitable approach that produces a synthetic 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 proceeding with synthetic data requests.*	☐ *[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 / fidelity assessment, reference here and include in an appendix – see Step 4 for more detail]* • De-identification techniques (Appendix 7) • Decision tree for complex synthetic 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 data safely.*	☐ Safety assessment (Appendix 6)

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 data requests*	☐ Request and Assessment Outcomes Form (Appendix 11)

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

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

Responsibility for decision-making about synthetic data

Different types of synthetic 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 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 data is appropriate for the particular use case, as well as which method / model will be used to generate the synthetic data. For synthetic 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 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 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 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 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 data. The Data Requestor will be able to explain the use case for the synthetic 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 dataset can be generated, re-used or re-purposed, and what attributes are required. The Data Requestor is responsible for ensuring the synthetic data is processed and handled in a secure manner, and that the integrity of the dataset is maintained. Where synthetic data is *transferred* by a Data Provider to the Data Requestor (as opposed to the Data Provider *providing access to* a synthetic dataset), the Data Requestor who receives the data is responsible for its secure storage and handling. Data security measures for storing and accessing synthetic data are discussed further in this framework, and will depend on the level of re-identification risk associated with the synthetic data and the particular use case.
End Users	**End Users** are individuals such as data analysts or researchers who will access and use synthetic 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.

Responsibilities under this framework – who to consult

Requests for synthetic data will require collaboration between organisations that both hold relevant real data needed to generate synthetic data, and those who wish to access and use synthetic data. There may also need to be broader collaboration with other organisations that have synthetic data expertise but are not otherwise involved in handling data within scope of the synthetic 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 data, and opportunities to refine what is needed data-wise to support a particular project prior to an organisation receiving a request for synthetic data. In line with this framework, organisations should be supportive of synthetic data generation and use, given the range of benefits to multiple stakeholders and the privacy protective nature of synthetic data compared with real data.

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) Human Research Ethics Committee (HREC) approval under a research exemption because there are material privacy risks in a synthetic dataset that cannot be further de-identified.

While the assessments in this framework must be followed and documented for each synthetic 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 data projects, as required under an organisation's own frameworks and policies.[12] Where this is the case, both the information gathered and the assessments completed under this framework will provide valuable inputs for these activities.

[12] 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)