The Framework

Guiding principles for handling synthetic data requests

As discussed in the Overview, organisations involved in synthetic data requests are encouraged to recognise the significant benefits of generating and using synthetic data over real data across a wide range of important use cases, including:

- Enhanced privacy protection for individuals
- The opportunity to strengthen trust
- Cost and time savings
- Increased scalability and volume of datasets needed for analysis
- Customisation of datasets to ensure they are fit-for-purpose, including addressing risks around bias, under-representation and / or gaps or errors in real data
- Modeling and creating data where real data does not exist ('making the invisible, visible')

This framework is designed to help data custodians and other stakeholders progress synthetic data requests with the confidence that they are doing so in a manner that ensures compliance with their privacy obligations. Where the risk of re-identification associated with a synthetic dataset is validated to be at a 'very low' level, organisations will be able to use the synthetic dataset for use cases that may not otherwise be permitted if real data was to be used (for example, without further HREC approvals).

While there will be use cases (or aspects of use cases) where real data is required or preferable over synthetic data, organisations will approach synthetic data requests in line with the following guiding principles:

- Where synthetic data *can* be used for a particular use case, it *should* be used in preference to real data
- Organisations that hold real data are committed to supporting and facilitating synthetic data requests given the material benefits to a wide range of stakeholders, including health system consumers
- Organisations will bring an 'innovation' mindset to projects, and look for ways to take advantage of the opportunities provided by synthetic data so that the *value and benefits* of the data can shared in faster, meaningful ways with communities represented by the data
- [For SynD members: please add any other guiding principles]

Steps for generating and accessing Synthetic Data

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

At a high level, synthetic data requests will involve 5 key steps. Each step must be successfully completed before moving on to the next.[14] Depending on the outcome of each step, some steps may need to be iterated before a request can progress.

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

Step 1: Assess the use case

Synthetic data will not be suitable for all use cases. For example, synthetic 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 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 dataset, will need to comply with the ‘Use’ principle/s in the applicable privacy law. This is discussed further in Appendix 9.

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,

[14] This five-step approach is based on the Personal Data Protection Commission Singapore’s *Privacy Enhancing Technology (PET): Proposed Guide On Synthetic 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

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 2 and 8 for a further discussion of privacy law).

Using health data to generate synthetic data is a use of information about health consumers for a secondary purpose. In order to use health data for this secondary purpose, unless an exception applies, the use case for the synthetic data request must be one that is *directly related* to the primary purpose of collection, and individuals would *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 data requests under this framework if all three tests below can be met:

- 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.

- The stated aim for creating and managing the synthetic 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 dataset with only a very low risk of re-identification will be suitable for use cases to proceed under this framework.
- 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 some **public communication about using synthetic health data** for public benefit projects before organisations facilitate synthetic 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 created with health consumers as the audience. This 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.

Under this framework, a use case will not be suitable to proceed if it does not meet these tests. 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 data for sharing. The Data Custodian at the Data Provider should be comfortable that *using real data* to create the synthetic 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 data request can proceed on this basis, an alternative lawful privacy pathway must be determined (see Appendices 7 and 8 for guidance on dealing with complex synthetic 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 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 data needs to be considered in terms of whether the non-legal 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 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 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 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 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 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 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 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 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 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 dataset (if they haven’t already been removed or treated).

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 data request to the next steps.

Step 3: Generate the synthetic data

There are various methods for generating synthetic 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 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 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 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 correlations between key variables that must be preserved for the use case. These correlations will need to be taken into account when generating the synthetic data, while also aiming to keep statistical disclosure risk to a minimum.

Once generated, the Data Provider should check the synthetic 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 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 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 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 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 data should be generated, and warns:

> "Synthetic data may not represent outliers present in the original personal data. You will need to assess whether the personal data on which the synthetic data was trained can be reconstructed. Further additional measures (e.g. Differential Privacy) may be required to protect against singling out".[15]

('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

[15] 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/

'personal data', or 'personal information' as it is known for the purposes of privacy law in Australia.)

Thus again, unless synthetic 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 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 simultaneously enables adversaries to extract sensitive information about individuals."[16]

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 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 dataset would pose a legal compliance and reputational risk.

Once a synthetic dataset has been created, there will be additional legal compliance issues if the data in the synthetic dataset could contain 'personal information'. Following the creation of the synthetic 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

[16] See Theresa Stadler, Bristena Oprisenu and Carmela Troncoso, "Synthetic 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.

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 dataset must be considered and tested via a robust Re-identification Risk Assessment.

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

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.[17] 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".[18]

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 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 dataset must be considered 'personal information', and privacy law obligations will continue to apply to the way it is handled. This means the Data Custodian cannot use or share the synthetic 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.

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:

[17] 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

[18] 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

> "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."[19]

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 dataset will be shared with the Data Requestor (e.g. if the full synthetic 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 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 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

For SynD members: we recommend as a next step that SynD develop an agreed methodology for conducting re-identification risk assessments, which is then linked to this framework. The methodology should use a unform approach for benchmarking re-identification risk so organisations can have confidence the 'very low risk' threshold is standardised and appropriate regardless of which organisation is carrying out the de-identification process.

We can connect SynD with one of our associates who is an expert in re-identification risk assessments if assistance is needed. There are various testing methodologies, each with their own strengths and limitations (e.g. see the following paper about two key methodologies for testing synthetic data for re-identification risk: https://arxiv.org/abs/2505.01524). See also: A consensus privacy metrics framework for

[19] 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

synthetic data (https://www.sciencedirect.com/science/article/pii/S2666389925001680) and Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation (www.jmir.org/2020/11/e23139/)

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

[For SynD members: this type of data utility / fidelity assessment was suggested earlier as feedback. Similar to the re-identification risk assessment, does the SynD community 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 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 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 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 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 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 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 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 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 dataset is not approved, the accountable decision-maker must provide reasons and further context where appropriate.

All synthetic data requests and their outcomes must be documented. The Request and Assessment Outcomes form (attached at Appendix 110) should be used to document synthetic data requests, assessment results and approvals.

Both the Data Requestor and the Data Provider will have responsibility for maintaining synthetic 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 data and its parameters
- statement(s) of data quality
- any conditions the requester has been asked to meet
- for complex synthetic data requests, documentation of the lawful privacy pathway to create and share the synthetic data (see Appendix 8 for further guidance on complex requests)
- documentation of the accountable decision-maker's approval to create and share the synthetic data
- any agreed modifications
- metadata describing the synthetic data provided
- any Privacy Impact Assessment completed in connection with the synthetic data request
- the Re-Identification Risk Assessment completed in connection with the synthetic 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 data is not created or provided, the reason for that decision

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

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

Data Requestors may also wish to use a synthetic 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 dataset is fit for purpose in light of the use case (Step 2)
- Consideration should be given to whether the synthetic 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 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 data is protected (Step 5)