pdpc

PERSONAL DATA
PROTECTION COMMISSION
SINGAPORE

PRIVACY ENHANCING TECHNOLOGY (PET): PROPOSED GUIDE ON SYNTHETIC DATA GENERATION

Published 15 July 2024

Version Number 1.0

JOINTLY DEVELOPED WITH

Agency for
Science, Technology
and Research
SINGAPORE

SUPPORTED BY

INFOCOMM
MEDIA
DEVELOPMENT
AUTHORITY

TABLE OF CONTENTS

I. Introduction to Privacy Enhancing Technology (PET)

Privacy Enhancing Technologies (PETs) are a suite of tools and techniques that allow the processing, analysis, and extraction of insights from data without revealing the underlying personal or commercially sensitive data. By incorporating PETs, companies can maintain a competitive edge in the market through leveraging their existing data assets for innovation while complying with data protection regulations, reducing the risk of data breaches and demonstrating a commitment to data protection. PETs are not just a defensive measure; they are a proactive step towards fostering a culture of data protection and securing a company's reputation in the digital age.

PETs can generally be classified into three key categories[1]: data obfuscation, encrypted data processing, and federated analytics. PETs can also be combined to address varying needs of organisations. The following **Table 1** maps out the current types of PETs in the market and their key applications.

Table 1. Types of PETs and their applications

Categories of PETs	PETs	Examples of applications (non-exhaustive)
Data obfuscation	Anonymisation/pseudonymisation techniques	• Secure storage • Data sharing and retention • Software testing
	Synthetic data generation	• Privacy-preserving AI machine learning • Data sharing and analysis • Software testing
	Differential privacy	• Expanding research opportunities • Data sharing
	Zero knowledge proofs	• Verifying information without requiring disclosure (e.g., age verification)
Encrypted data processing	Homomorphic encryption	• Secure data stored in cloud

[1] Adapted from OECD, "Emerging Privacy Enhancing Technologies: Current Regulatory and Policy Approaches," *OECD Digital Economy Papers* (OECD, 2023).

		• Computing on private data that is not disclosed
	Multi-party computation (including private set intersection)	• Computing on private data that is not disclosed
	Trusted execution environments	• Computing using models that need to remain private • Computing on private data that is not disclosed
Federated analytics	Federated learning	• Privacy-preserving AI machine learning
	Distributed analysis	

II. Synthetic Data

This guide focuses on the use of synthetic data[2] to generate structured data. While synthetic data is generally fictitious data that may not be considered personal data on its own, it is not inherently risk-free due to possible re-identification risks[3]. As such, this guide proposes good practices that organisations may adopt to generate synthetic data to minimise such risks for a set of common use case archetypes. The guide also includes a set of good practices and risk assessments/considerations for generating synthetic data as well as governance controls, contractual process, and technical measures to mitigate residual risks.

The target audience for this guide are CIOs, CTOs, CDOs, data scientists, data protection practitioners, and technical decision-makers who may directly or indirectly be involved in the generation and use of synthetic data.

Synthetic data is a technology that is being actively researched and developed at the time of publication. Hence, this guide is not intended to provide a comprehensive or in-depth review of the technology or its assessment methods. The guide is intended to be a living document, and will be updated to ensure its recommendations remain relevant.

[2] There are two types of synthetic data: fully synthetic data and partially synthetic data. This guide discusses the use of fully synthetic data.

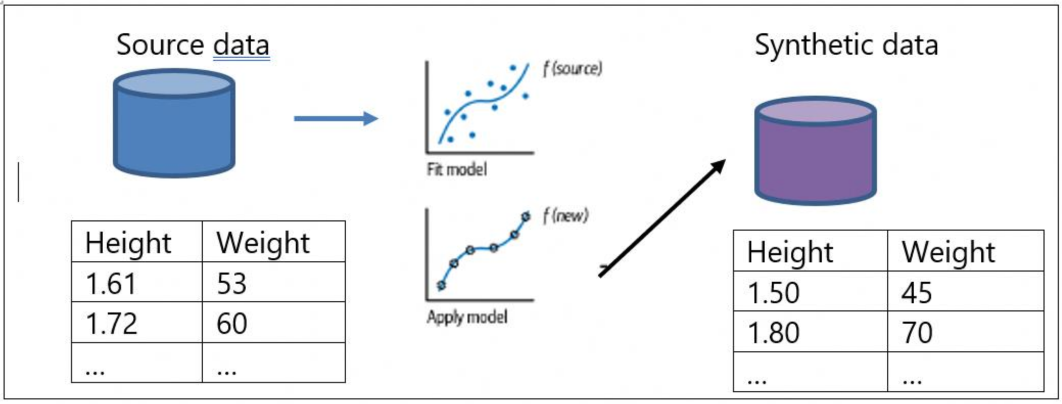
[3] In this guide, we generally refer to privacy risks as re-identification risks.

What is Synthetic Data?

Synthetic data is commonly referred to as artificial data that has been generated using a purpose-built mathematical model (including artificial intelligence (AI)/machine learning (ML) models) or algorithm. It can be derived by training a model (or algorithm) on a source dataset to mimic the characteristics and structure of the source data. Good quality synthetic data can retain the statistical properties and patterns of the source data to a high extent. As a result, performing analysis on synthetic data can produce results similar to those yielded with source data.

Characteristics of synthetic data

Figure 1 shows an example of how synthetic data may look like as compared with the source data. Generated synthetic data will generally have different data points from the source data, as seen from the tabular data. However, the synthetic data will have statistical properties that are close to that of the source data, i.e., capturing the distribution and structure of the source data as seen from the trend lines in **Figure 1**.



Height	Weight
1.61	53
1.72	60
...	...

Height	Weight
1.50	45
1.80	70
...	...

Figure 1: Source data versus synthetic data.[4]

As such, synthetic data may not always be inherently risk-free as information about an individual in the source dataset, or confidential data, can still be leaked due to the resemblance of the synthetic data to the source data. There will also be trade-offs[5] between data utility and data protection risks in synthetic data generation. However, such risks can be minimised by taking data protection into consideration during the synthetic data generation process.

[4] Diagram taken with modification from Khaled El Emam, Lucy Mosquera, and Richard Hoptroff, *Practical Synthetic Data Generation* (O'Reilly Media, Inc, 2020).

[5] Trade-off between data utility and data protection risks is further discussed in *Annex A: Step 1 and Step 3* in this guide.

Under What Circumstances is Synthetic Data Useful?

Synthetic data can be used in a variety of use cases ranging from generating training datasets for AI models to data analysis and collaboration. The use of synthetic data not only can accelerate research, innovation, collaboration, and decision-making but also mitigate concerns about cybersecurity incidents and data breaches, enabling better compliance with data protection/privacy regulations. **Table 2** discusses a few common use case archetypes, their key benefits, and good practices that organisations can focus on when generating synthetic data.

Table 2. Use case archetypes for synthetic data.

Types of Use Cases	Key Benefits	Good Practices to Generate Synthetic Data
Use case archetype 1: Generating training dataset for AI models		
Augmenting data for AI/ML models	• Synthetic data addresses the challenge of the user having to obtain large volumes of labelled data needed for training and testing AI/ML models due to costs, legal regulations, and proprietary rights. • Augmenting training datasets with synthetically generated labelled data can be more ***cost-effective***, especially when the source datasets are sparse.	• Add noise* to or reduce granularity of the synthetic data points. • Such fictitious new data points will generally not be considered personal data. *If the statistical properties/characteristics of the synthetic data is representative of the population in question and not significantly skewed towards a specific individual/group of individuals used as source training data, adding of noise might not be necessary as re-identification risks are generally low.
Increasing data diversity for AI/ML models	• Synthetic data can be used to simulate rare events or augment under-represented groups in training AI models. • Diverse datasets can be useful in ***improving performance*** of AI/ML models	
Use case archetype 2: Data analysis and collaboration		
Data sharing and analysis	• Underlying trends or patterns, and biases of the data are useful for data analytics regardless of whether the data source is real or synthetic.	• Balance the trade-offs between data utility and data protection by incorporating data protection measures

	• Synthetic data can enable data sharing for analysis especially in industries and sectors, e.g., healthcare, ***where the source data can be sensitive***.	throughout the synthetic data generation process, for example: Data preparation • Remove outliers from source data • Pseudonymise source data • Employ data minimisation and generalise granular data Synthetic data generation • Add noise before or after synthetic data generation Post synthetic data generation • Incorporate technical, contractual, and governance measures to mitigate any residual re-identification risks
Previewing data for collaboration	• Synthetic data can be used in data exploration, analysis, and collaboration to provide stakeholders with a representative preview of the source data without exposing sensitive information. • This enables stakeholders to explore and understand the structure, relationships, and potential insights within the data to ***gain assurance of the data quality*** before finalising any agreement or collaboration.	
Use case archetype 3: Software testing		
System development/ software testing	• Organisations can use synthetic data instead of production data to facilitate software development. • Use of synthetic data can help organisations ***avoid data breaches*** in the event of the development environment being compromised.	• Focus on generating synthetic data that follows semantics e.g., format, min/max values and categories, of source data instead of the statistical characteristics and properties.

Refer to **Annex A** for proposed considerations and good practices to generate synthetic data.

Case Studies

(A) Training AI model for fraud detection in the financial sector[6]

Problem: Since the number of fraudulent transactions in the source data is small compared to normal, non-fraudulent transactions, the source data did not train models very well for fraud detection.

Solution: J.P. Morgan successfully used synthetic data for fraud detection model training. AI models were provided with samples of normal and fraudulent transactions to understand the tell-tale signs of suspicious transactions.

Benefit: Synthetic data proved to be more effective in terms of training models to detect anomalous behaviour. This is because the synthetic data used was designed to contain a higher percentage of fraudulent transactions.

(B) Training AI model for research into AI bias[7]

Problem: Multi-label classification and regression models are frequently utilised at Mastercard for various applications, including fraud prevention, anti-money laundering and marketing use cases for portfolio optimisation. These models, while powerful, require careful attention to proxies of demographic attributes within their training data, which could learn unintended biases. Ensuring the accuracy and fairness of these models is complex due to their multi-label setting, the confidentiality of the demographic attributes, and the challenges in accessing the training dataset for model development.

Solution: Mastercard partnered with researchers to develop new AI bias testing methods adapted to multi-label settings. To protect the privacy of the data shared externally, synthetic data was created to support model training and methodological research into fair multi-label models.

Benefit: Synthetic data was measured to be sufficiently private to be shared with external researchers while capturing real relationships within the source data. Synthetic data enabled new insights that would not have been possible without the privacy protecting characteristics inherent to synthetic data.

[6] J. P. Morgan, "Synthetic Data for Real Insights," Technology Blog, n.d., https://www.jpmorgan.com/technology/technology-blog/synthetic-data-for-real-insights

[7] Contributed by Mastercard

(C) Safeguarding patient data for data analysis[8]

Problem: Prior to utilising synthetic data, Johnson & Johnson (J&J) allowed external researchers or consortia to access healthcare data for research proposals validated by J&J. To safeguard patient privacy, the data was transformed into anonymised healthcare data. However, feedback received indicated that the overall usefulness of the anonymised data, which relied on traditional anonymisation techniques, was not always satisfactory and did not always meet the requirements of the researchers or consortia.

***Solution*:** J&J has introduced high-quality AI generated synthetic data as an additional option to process their healthcare data.

***Benefit*:** Researchers and clients have experienced significantly improved analysis. When employed properly, this form of synthetic data can effectively represent the target population and offer various analytical and scientific benefits.

(D) Facilitating data collaboration[9]

***Problem*:** A pharmaceutical company wanted to purchase heart-related health data from a research institute to test out a new hypothesis. The health data, which was collected by the research institute from consenting subjects, was hosted under a highly regulated environment as required of the healthcare sector. However, this presents significant challenges for many data engagement activities.

***Solution*:** A*STAR was engaged by the pharmaceutical company to build a pipeline to create synthetic copies of the actual data, which can then be brought outside of this regulated environment.

***Benefit*:** This allowed the pharmaceutical company to preview the data and be assured of the data quality prior to the high-value purchase and access to the actual data.

[8] Contributed by Johnson & Johnson (J&J)

[9] Contributed by A*STAR

III. Recommendations

Synthetic data has the potential to drive the growth of AI/ML by enabling AI model training while protecting the underlying personal data. It also addresses dataset related challenges for AI model training, such as insufficient and biased data, through enabling the augmentation and increased diversity of training datasets.

In addition, synthetic data can be used to facilitate and support organisations' data analytics, collaboration and software development needs. An added benefit of using synthetic data in place of production data to facilitate software development is that data breaches can be avoided in the event the development environment is compromised.

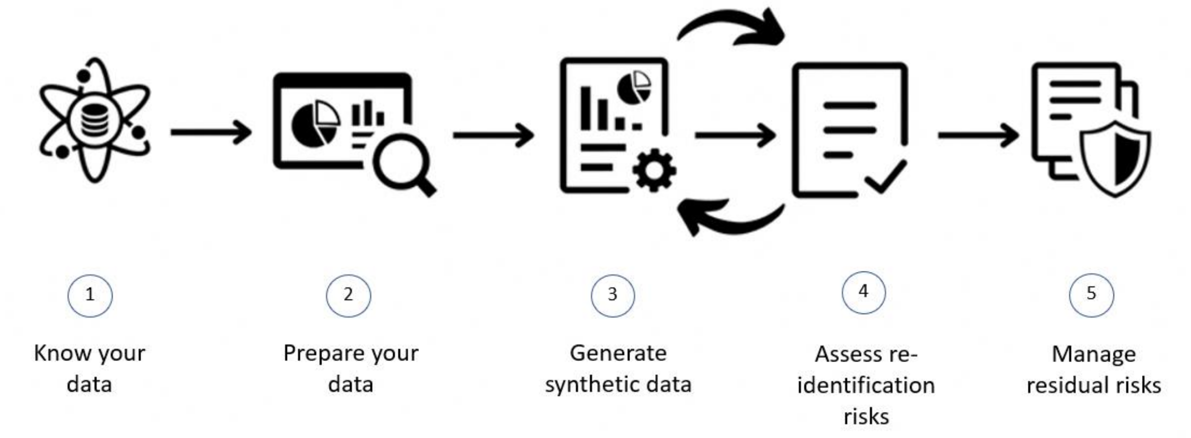
PDPC recommends a set of good practices and risk assessments/considerations for generating synthetic data and to reduce any residual risks from re-identification through governance controls, contractual process, and technical measures (refer to **Annex A**).

Annex A: Handbook on Key Considerations and Best Practices in Synthetic Data Generation

In this handbook, we describe the key considerations and best practices for organisations to reduce re-identification risks of synthetic tabular data through a five-step approach.

For any other complex synthetic datasets that are unstructured, organisations are advised to consider hiring synthetic data experts, data scientists or independent risk assessors to assess and mitigate the risks of the generated synthetic data.

Overview of five-step approach to generate synthetic data



Step 1: Know your data

Before embarking on any synthetic data project, it is necessary to have a clear understanding of the purpose and use cases of the synthetic data and the source data that the synthetic data is to mimic. This will help to determine whether use of synthetic data might be relevant and identify the possible risks of using the synthetic data. Some of the considerations may include:

- Where general trends/insights of source data are sensitive, organisation should take note that the use of synthetic data will not offer any protection to the trends/insights since they will be replicated in the synthetic data.
- Where the synthetic data is intended to be released publicly, organisations may have to prioritise data protection over data utility in such circumstances.

- Where relevant, organisations should also put in place proper contractual obligations on recipients of synthetic data where necessary to prevent re-identification attacks on the data.

With this knowledge, the management and data owner, with the help of relevant stakeholders such as the data analytics team, should establish objectives prior to synthetic data generation to determine an acceptable risk threshold[10] of the generated synthetic data and the expected utility of the data. This will help provide organisations with the appropriate benchmarks to assess any trade-offs between data protection risks and data utility.

These benchmarks may be adjusted appropriately to meet the business objectives, taking into consideration any trade-offs between data utility and data protection risks after the synthetic data generation process, as well as safeguards and controls to mitigate or lower any residual risks posed by the generated synthetic data. The acceptance criteria should be incorporated into the organisation's risk assessments (e.g., enterprise risk management framework[11] if applicable) or a Data Protection Impact Assessment ("DPIA")[12].

Step 2: Prepare your data

When preparing the source data[13] for generating synthetic data, it is important to consider the following:

- What are the key insights that needed to be preserved in the synthetic data?
- Which are the necessary data attributes for the synthetic data to meet the business objectives?

[10] The re-identification risk threshold represents the level of re-identification risk that is acceptable for a given synthetic dataset. There is currently no universally accepted numerical value for risk threshold. For further details refer to **Step 4** (Assess re-identification risks).

[11] Organisations may refer to ISO27001 for more information on developing an enterprise risk management framework.

[12] An example of this is PDPC's *Guide to Data Protection Impact Assessments*. A DPIA is applicable in the case where personal data is involved. The DPIA may not be relevant in situations where the synthetic data generation does not involve personal data processing.

[13] This step assumes that the source data has been properly cleaned (such as fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data) and is of acceptable quality for the generation of synthetic data.

Understanding key insights to be preserved

To ensure that the synthetic data can meet the business objectives, organisations need to understand and identify the trends, key statistical properties, and attribute-relationships in the source data that need to be preserved for analysis e.g., identify relationships between demographic characteristics of population and their health conditions.

Organisations should consider, at this point, whether outlier trends and insights are necessary to be preserved for the business objectives. Key considerations could include the following:

- If outliers are not necessary to meet the business objectives and the risk of re-identification is high, organisations should consider removing the outliers. This can be done prior to synthetic data generation or at subsequent stages of the synthetic data generation.

- If the objective is to mimic the characteristics of the source data as closely as possible, including outliers, then the organisation may have to preserve the outlier trend/insight to meet the business objectives. In such instance, the organisation should note that the re-identification risks of individuals in the outlier data may be high and hence put in place risk mitigation measures.

- If the business objective is to balance the number of data points in different data categories, then the synthetic data generation process itself can help mitigate the issue of outliers simply by generating more outliers. For example, in a dataset, the number of outlier data points comprising male individuals may be balanced with outlier data points comprising female individuals.

Selecting data attributes

Based on the key insights needed, organisations should apply data minimisation to extract only the relevant data attributes from the source data. Thereafter, remove or pseudonymise all direct identifiers[14] from the extracted data.

Where granular information is not necessary, organisations may generalise or further add noise to the data at this point or at a later step to reduce the risk of re-identification. For example, organisations can generalise exact height and weight

[14] Refer to PDPC's *Guide to Basic Anonymisation* on how to identify direct identifiers in a dataset.

information into height and weight bands to reduce the possibility of height and weight combinations being used to identify any outliers.

Organisations should also standardise and document the details on each data attribute (such as data definitions, standards, metrics etc.) in a data dictionary. This enables the organisation to subsequently validate the integrity of the generated synthetic data to detect anomalies and fix any data inconsistencies. Refer to the following checklist in **Table 3** for key considerations.

Table 3: Checklist for data preparation

Data Preparation Checklist	
Understand key insights	
i.	Identify trends and entity relationships to be preserved for synthetic data generation.
ii.	Remove outliers if such trends/insights are not necessary. This can be performed post generation.
Select data attributes	
iii.	Apply data minimisation to select only data attributes that are necessary to meet business needs.
iv.	Remove or pseudonymise direct identifiers (e.g., name, national identification numbers).
v.	Generalise granular data or add noise (e.g., using differential privacy[15]) to the data/model if such detailed information is not necessary. This can also be performed post generation.
vi.	Standardise and document format, constraints, and categories of source data in data dictionary (refer to **Annex B** for a reference template): Format • Standardise strings to lower or proper case • Data types, column names, structures, relationships • Frequency of data record Constraints • Constraints of values for each data type, e.g., min-max values, non-negative values, non-null values Category • Types of data categories • Expected or valid values for data attributes within each data category. Example of a data category is "country".

[15] The use of differential privacy to add noise to synthetic data is widely discussed as a mechanism to reduce re-identification risks. However, there is currently no universal standard on how to implement differential privacy. Moreover, the noise added may also reduce the utility of the synthetic data, making it less accurate or useful for certain types of analysis.

Step 3: Generate synthetic data

There are many different methods [16] to generate synthetic data, for example, sequential tree-based synthesisers, copulas, and deep generative models (DGMs). Organisations need to consider which methods are most appropriate, based on their use cases, data objectives, and types of data. Please refer to **Annex C** for more information on these synthetic data generation methods. Thereafter, organisations may consider splitting the source data into two separate sets e.g., 80% as training dataset, and 20% as control dataset [17] for assessing re-identification risks of the synthetic data.

After generating synthetic data, it is a good practice for organisations to perform the following checks on the quality of the generated synthetic data:

- Data integrity
- Data fidelity
- Data utility

Data integrity

Data integrity ensures the accuracy, completeness, consistency, and validity of the synthetic data as compared with the source data. Organisations can validate the integrity of the generated synthetic data against the dictionary of the source data.

Data fidelity

Data fidelity examines if synthetic data closely follows the characteristics and statistical attributes of the source data. There are a few metrics for measuring data fidelity and they are typically done by statistically comparing the generated synthetic data directly with the source data. Organisations should use the performance metric(s) for data fidelity[18] (see **Table 4**) that best meet their data objectives.

[16] This guide may not be comprehensive in covering all other synthetic data generation methods such as Bayesian model and variational autoencoders (VAE).

[17] Refer to Approach 2 in Annex E for more details on the assessment and evaluation framework for quantifying re-identification risk.

[18] There are other generic metrics described here in addition to those listed in Table 4. See Khaled El Emam et al., "Utility Metrics for Evaluating Synthetic Health Data Generation Methods: Validation Study," *JMIR Medical Informatics* 10, no. 4 (2022)

Table 4: Performance metrics for data fidelity

Performance metrics generally used for assessing data fidelity	
Histogram-based similarity	Measures the similarity between source and synthetic data's distributions through a histogram comparison of each feature. This ensures the synthetic data preserves important statistical properties such as central tendency (mean, median), dispersion (variance, range), and distribution shape (skewness, kurtosis).
Correlational similarity	Measures the preservation of relationships between features in the source and synthetic datasets. For example, if higher education typically leads to higher income in the source data, this pattern should also be evident in synthetic data.

Data utility

Data utility refers to how well synthetic data can replace or add to source data for the specific data objective of the organisation.

There are different approaches to evaluate the utility of synthetic data. The true test of utility is how it performs in real-world tasks. One common approach to check this is by training identical AI/ML models on synthetic and training data. The performances from the two models are compared with the control dataset, simulating testing in the production environment, to assess the utility of the synthetic data. Examples of performance metrics generally used include "accuracy", "precision", "recall", "F1-Score", or "Area Under the ROC Curve (AUC-ROC)" for classification tasks, and "Mean Absolute Error (MAE)" or "Mean Squared Error (MSE)" for regression tasks[19] (see definition in **Table 5** below). If their compared scores are close, then it indicates that the synthetic data has high utility. In simple terms, a high utility score means that machines trained on synthetic data work similarly to those trained on training data.

When trying to maximise the utility of data, there is often an inherent trade-off between data utility and data protection. Thus, a fine balance between data utility and

[19] There is another performance metric suitable for regression tasks, i.e., replicability, which is used for assessing data utility and is described here in addition to those listed in Table 5. See Khaled El Emam et al., "An Evaluation of the Replicability of Analyses Using Synthetic Health Data," *Scientific Reports* 14 (2024), https://www.nature.com/articles/s41598-024-57207-7

data protection needs to be achieved through an iterative process (Steps 3 and 4) to synthesise data up to an acceptable level for re-identification risks while finding the right balance of data utility.

Table 5: Performance metrics for data utility

Performance metrics generally used for assessing data utility	
Accuracy	Measures the overall correctness of the model. It is calculated as the ratio of correct predictions (true positives and true negatives) to the total observations. *E.g., if out of the health data of 100 individuals, the model predicts 90 of these individuals' health status correctly, the accuracy is 90%.*
Precision	Measures the model's ability to identify only relevant instances. It is calculated as the ratio of correct positive predictions (true positives) to all positive predictions (true positives and false positives). *E.g., If out of 100 individuals that are predicted as diseased by the model, 80 of these individuals are correctly identified as diseased, the precision is 80%.*
Recall	Measures the model's ability to find all relevant cases. It is calculated as the ratio of correct positive predictions (true positives) to all actual positives (true positives and false negatives). *E.g., If out of 100 diseased individuals, the model predicts 90 of these individuals as diseased, the recall is 90%.*
F1-score	Balances precision and recall in a single metric. (mathematically, it is the harmonic mean[20] of precision and recall).
Area Under the ROC Curve (AUC-ROC)	Measures the model's ability to distinguish between classes. It is represented by the area under the Receiver Operating Characteristic (AUC-ROC) curve, comparing the true positive rate to the false positive rate at various classification thresholds. *E.g., If the AUC-ROC score is 0.9, it means there is a 90% chance that the model will correctly distinguish between a randomly chosen positive instance and a randomly chosen negative instance.*
Mean Absolute Error (MAE)	Measures the model's errors in predictions by averaging the absolute differences between predicted and actual values, providing a direct measure of average error magnitude without considering error direction. It is calculated as the mean of the absolute differences between actual and predicted values.

[20] A type of average that gives more weight to lower values of precision and recall scores.

Mean Squared Error (MSE)	Measures the model's errors in prediction by averaging the squares of the errors between predicted and actual values. MSE heavily penalises larger errors more than smaller ones, due to squaring the error values. This makes it more sensitive to outliers and large errors. It is calculated as the mean of the squared differences between actual and predicted values.

Use the following checklist in **Table 6** as a reference guide where applicable.

Table 6: Checklist for checking the generated synthetic data

Post-generation Checklist	
i.	Remove outliers if such trends/insights are not necessary to meet business needs.
ii.	Generalise granular data or add noise to the data/model if such detailed information is not necessary.
iii.	Perform data integrity checks on synthetic data by validating data format, structures etc with the earlier documented data dictionary.
iv.	Select relevant metrics that meet data objectives to measure data fidelity.
v.	Select relevant performance metrics that meets data objectives to measure data utility.

Step 4: Assess re-identification risks

After the synthetic data is generated and utility measurement is assessed to be acceptable, organisations should assess and perform the re-identification risk assessment based on their internal acceptance criteria. **Annex D** discusses the widely known re-identification risks for synthetic data. As synthetic data generally does not replicate its training data points, re-identification risk cannot be deduced directly from scrutinising whether the generated synthetic data contains any personal data.

Generally, re-identification (or privacy) risk assessment for synthetic data is an attack-based evaluation. It evaluates how successful an adversary, who carries out re-dentification attacks through singling out attacks, linkability attacks and inference attacks (as described in **Annex D**) on synthetic datasets, can determine if an individual belongs to the source dataset (i.e., membership inference) and/or derive details of an individual from the source dataset which are otherwise undisclosed (i.e., attribute inference). The goal for organisations is to ensure that the re-identification risk levels for the three key re-identification attacks are acceptable. If re-identification risk level is unacceptable, repeat Step 3 to re-generate synthetic data to meet the acceptable risk level. This can be achieved by applying more data protection controls on the source data, e.g., generalising the data or adding noise (see "Checklist for data preparation" in **Table 3**).

Various approaches have been proposed to determine and quantify re-identification risks. Refer to **Annex E** for examples of such approaches. Organisations may need to engage the synthetic data solution provider to perform the re-identification risk assessments.

While there is no universally accepted numerical threshold value for risk level, some organisations[21] have chosen to align their re-identification risk level with existing industry guidelines and recommendations for de-identified/anonymised data (see **Table 7**). However, organisations should take note that the computation method for re-identification threshold in a de-identified/anonymised dataset is very different from that for a synthetic dataset. Nevertheless, the fundamental basis for both is that the re-identification/privacy risk assessment is a probabilistic measurement.

The re-identification risk threshold values in **Table 7** summarises the precedent acceptable risk threshold used by some organisations for assessing de-identified/anonymised data.

Table 7. Existing risk threshold guidelines for de-identification/anonymisation

Risk threshold for de-identification/anonymisation	
European Medicines Agency (EMA)	The European Medicines Agency (EMA) established a policy on the publication of clinical data for medicinal products. The guidelines accompanying the policy recommend a maximum risk threshold of 0.09.[22]
Health Canada	Health Canada implemented the same threshold as EMA, 0.09, for the sharing of clinical data.[23]

[21] Samer El Kababji et al., "Evaluating the Utility and Privacy of Synthetic Breast Cancer Clinical Trial Data Sets," *JCO Clinical Cancer Informatics* 7 (2023), https://ascopubs.org/doi/full/10.1200/CCI.23.00116

[22] European Medicines Agency, "European Medicines Agency Policy on Publication of Clinical Data for Medicinal Products for Human Use," 2019, https://www.ema.europa.eu/en/documents/other/policy-70-european-medicines-agency-policy-publication-clinical-data-medicinal-products-human-use_en.pdf

[23] Health Canada, "Guidance Document on Public Release of Clinical Information: Profile Page," 2019, https://www.canada.ca/en/health-canada/services/drug-health-product-review-approval/profile-public-release-clinical-information-guidance.html

ISO/IEC 27559 Privacy enhancing data de-identification framework.	ISO/IEC 27559 summarises a list of example thresholds providing a range of acceptable values which encompasses 0.09.

Step 5: Manage residual risks

In this final step, organisations should identify all potential residual risks and implement appropriate mitigation controls (technical, governance, and contractual) to minimise the identified risks. These risks and controls should be documented and approved by the management and key stakeholders as part of the organisation’s enterprise risk framework.

Organisations can take into consideration the following risks as part of risk assessment.

New insights derived from synthetic data

New insights may be learnt about the source data by analysing the synthetic dataset alone or in combination with other available datasets. Organisations should assess if these insights may be sensitive or could misinform.

Potential impact on groups of individuals due to membership disclosure

Membership disclosure is when an adversary, using the information in synthetic data, determines that a target group of individuals was included in the source dataset. Potential disclosures or inference of attributes from the synthetic dataset related to groups of individuals may be regarded as confidential in nature e.g., there may be social stigma impact on individuals in a counselling group if their membership were to be disclosed.

In determining the source dataset for synthetic data, it is important to also consider the sampling fraction from the population dataset, which is the ratio of the sample size within the source data as compared to the population size. For example, an adversary will have a lower chance of predicting whether a target group of individuals from a population is included in a synthetic dataset that is trained from a source dataset sampled from 20% of the population, as compared with a source dataset sampled from 90% of the population.

Parties receiving synthetic data

The receiving parties of the synthetic data, including any data intermediaries, may pose data breach compliance risks when handling synthetic data. Organisations should assess the data recipient's ability and motivation to re-identify individuals from the dataset. A data recipient who possesses specialised skillsets or technologies may be able to combine special knowledge or get public knowledge to re-identify any individual from the dataset. Such risks must be accounted for in the risk assessment exercise.

Changing environment

The likelihood of re-identification risks of any given synthetic dataset increases over time, due to increase in computing power and improvement in data-linking techniques.

Model leakage

A model that has been trained using source data to generate the synthetic data can be susceptible to a malicious attack by adversary to reconstruct (parts of) the source data.

Safeguards and best practices

The following **Table 8** lists examples of best practices that organisations can consider implementing to manage residual risks posed by using synthetic data.

Table 8. Best practices and security controls to implement and manage risks

Governance	Access controls	Implement access control for the source data and synthetic data generator model. Apply access control to synthetic data where the re-identification or residual risk is high, especially if the data contains highly sensitive information or insights.
	Asset management	Properly label synthetic data to prevent human error when managing both source data and synthetic data.
	Risk management	Periodically conduct re-identification risk reviews of synthetic datasets, especially if these are publicly released.
	Legal controls	Have in place contractual agreements to outline the responsibilities of third-party

		recipients of the synthetic data and/or models as well as any third-party solution providers who provide the synthetic data generation tools. This includes safeguarding the data/model and prohibiting attempts to re-identify individuals. In situations where the organisation may need to depend on the synthetic data solution provider to perform the risk assessment and mitigations, the solution provider may be required to provide assurance that appropriate controls have been implemented.
ICT Controls	Database security	Segregate storage for synthetic data and source data.
IT Operations	Logging and monitoring	Properly log and monitor usage of source and synthetic data, as well as access to the synthetic data generator model.
Risk Management	Housekeeping of information	Securely delete source data, synthetic data, and synthetic data generator model when they are no longer needed or have reached the end of retention period.

Incident management

Organisations should identify the risks of data breaches involving synthetic data, synthetic data generator model, and model parameters, and incorporate relevant scenarios into their incident management plans. The following considerations may be relevant for organisations' internal investigations[24]:

Loss of fully synthetic data (for synthetic data that is not intended for public release)

Fully synthetic data that has data protection best practices incorporated in its generation process and has been assessed to have a low re-identification risk is generally not considered personal data. However, organisations should still proceed to investigate the incident to understand the root cause and improve its internal

[24] For data breach reporting to PDPC, organisations will have to assess if it is a notifiable breach based on PDPA's Data Breach Notification obligation.

safeguards against such occurrences in the future. Organisations should also monitor if there is any evidence of actual re-identification and assess if it would be a notifiable data breach to PDPC.

Loss of synthetic data generator model, parameters and/or synthetic data

Both the synthetic data generator model and its parameters can provide useful information to an adversary to perform a model inversion attack. With access to generated synthetic data, it may further enhance the adversary' ability to recover the source data. Organisations should proceed to investigate the incident to understand the root cause so as to improve its internal safeguards. It should also monitor for a possible successful model inversion attack which may result in the reconstruction and disclosure of the source data. Where such reconstruction and disclosure of source data is detected, organisations will have to assess if such breach would be notifiable.

Annex B: Data Dictionary Format

The following is a sample of data dictionary format:

COLUMN	DESCRIPTION	POSSIBLE VALUES	REMARKS
NAME	Name of the column	Gender, date of birth	
DESCRIPTION	Short description of the variable.		
TYPE	General data type. Specifically, how it appears superficially.	Numeric, string, date	Data can often appear in multiple formats. For example, categorical data can often be saved as ordinal integers, Boolean (which appears numerical), or as text, e.g., 'YES' or 'NO', or simply as a unique sequence of numbers, e.g., '3828' and '4271' (which looks numeric but is actually a string). Another example is when dates are saved as strings, or as date-formats using excel, or its numerical equivalent. These two columns help data users navigate this confusion and facilitate development of automated scripts. TYPE indicates how the data appears superficially, whether it should be processed as a number, string, or date when loading or saving the data. SPECIFIC DATA TYPE indicates how the data should be processed in the ideal situation during analyses, for example. a datatype such as education level should ideally be ordinal, though it can be treated as categorical, ordinal, or interval.
SPECIFIC DATA TYPE	The ideal way of processing the variable.	Example 1: If numeric, specify as 'float', 'int', 'boolean', 'ordinal', 'categorical', 'date' etc. Example 2: If string, specify as 'categorical', 'ordinal', 'free text' etc. Example 3: If date, specify as 'date'	

CODINGS		Example 1: If TYPE is 'date', use excel convention to indicate date format, e.g., dd/mm/yyyy, mm-dd-yyyy, etc. Example 2: If SPECIFIC DATA TYPE is 'boolean', 'ordinal', 'categorical', specify exhaustively all possible entries, delimited by '; ', e.g., YES; NO; N.A. OR Male; Female, OR 1; 2; 3; 4; 5 Example 3: If TYPE is 'numeric', specify range. E.g., [0,100] OR (3,4).	Take special notice of capital/small letters to avoid confusion.
FREQUENCY	For longitudinal data. Use to indicate if the variable is collected during a particular visit type.	Example 1: BASELINE; 6 WEEK; 6 MONTH Example 2: VISIT 1; VISIT 2	Leave blank if not longitudinal data.
CATEGORY	Use to group the variable under a specific category.	Example. 1: DEMOGRAPHICS Example 2: ECHO Example 3: LIFESTYLE VARIABLE	
SECONDARY	Whether the variable can be derived from other variables present in the dataset.	Example 1: If yes, 'Y' Example 2: If no, 'N', or leave blank.	For instance, BMI is a secondary variable if it is computed from 'height' and 'weight', and the two variables are also included in the dataset. Other examples are 'age_decade', where subjects of ages between 30 to 40 are grouped together as '30-40' to reduce granularity of the variable, or 'dementia', a diagnosis derived from answers to questions also present in the dataset.

			If yes, explain how the variable was computed from other variables such as bmi formula ir diagnosis standard/criteria etc, either in the CONSTRAINTS or REMARKS column.
CONSTRAINTS	How the variable is dependent on other variables.	Example 1: 'Head_circ' (head circumference) is a variable collected for 'age' <= 6. Leave empty if 'age' > 6. Example. 2: Collected only for data cohort '<COHORT NAME>' or hospital '<HOSPITAL A>'. Example 3: 'Ever_pregnant' only collected for females above age of 12. If 'male' or 'female' below age of 12, recorded as 'N.A.' If 'female' above age of 12, either 'YES', 'NO', or 'UNKNOWN'. Example 4: 'BMI' only computable if 'height' and 'weight' are also collected. Leave blank if either value is blank.	This information will help data users decide if a value is missing/unknown (should be collected but not collected), or not applicable (not collected because of procedure). Note that the value of a variable might be dependent (or conditional) on other variables, but it is not necessarily derived from other variables; CONSTRAINTS and SECONDARY are complementary, but the former does not imply the latter.
REMARKS	Additional comments, such as how the data is encoded, and/or concerns related to the variable.	Example. 1: How categorical variables are encoded as integers: 1=NO, 0=YES, -1=N.A. Example 2: Sensitive OR self-reported variable, etc. Example 3: Metric unit used for collection, 'cm', 'm', 'inches', etc.	It is often necessary to leave a note to remind data owners/users of the difficulties encountered during data collection, the corresponding response, and associated concerns. Some of these remarks can be included in the variable description, or here, if they are deemed miscellaneous.

Annex C: Examples of Methods of Synthetic Data Generation

Statistical Methods

(A) Bayesian Networks

Contributed by Betterdata.ai

Bayesian networks (BN) are probabilistic models that use a directed acyclic graph (DAG) to depict conditional dependencies between variables, enabling the generation of synthetic data statistically similar to the original data. BNs are helpful in sectors like healthcare and finance where accurate data relationships are essential. Typically, BNs require significant domain expertise for precise modelling via an expert-driven approach[25]. Alternatively, they can also be structured through data-driven methods, although these compromise accuracy due to less reliable inferences about the underlying data relationships.

PrivBayes[26] is an example of a BN that addresses moderate-dimensional data while preserving privacy. It constructs a Bayesian network to model relationships among data attributes and approximates the data distribution using low-dimensional marginals. By injecting noise into these marginals to ensure privacy, PrivBayes generates a synthetic dataset that closely mirrors the original while striking an efficient balance between data utility and privacy.

However, the scalability[27] of BNs is limited as their computational complexity can range from polynomial to exponential depending on the number of features and learning algorithms used. Polynomial complexity is achievable with expert-defined structures or by implementing accuracy constraints, such as a limited number of parents per node. Without these constraints, learning becomes an NP-hard problem, causing the complexity to exponentially increase with the number of features. While approximation algorithms can help manage

[25] Anthony Costa Constantinou, Norman Fenton, and Martin Neil, "Integrating Expert Knowledge with Data in Bayesian Networks: Preserving Data-Driven Expectations When the Expert Variables Remain Unobserved," *Expert Systems with Applications* 56 (2016): 197–208, https://www.sciencedirect.com/journal/expert-systems-with-applications/vol/56/suppl/C

[26] Ergute Bao et al., "Synthetic Data Generation with Differential Privacy via Bayesian Networks," *Journal of Privacy and Confidentiality* 11, no. 3 (2021), https://dr.ntu.edu.sg/handle/10356/164213

[27] Ole J. Mengshoel, "Understanding the Scalability of Bayesian Network Inference Using Clique Tree Growth Curves," *Artificial Intelligence* 174, no. 12–13 (2010): 987–1006, https://ntrs.nasa.gov/api/citations/20090033938/downloads/20090033938.pdf

computational demands, they may reduce accuracy. Therefore, BNs are favoured for scenarios that require interpretability but less for high-dimensional datasets where deep learning offers a more practical solution due to its ability to efficiently handle large-scale data.

(B) Conditional-Copulas

Contributed by the Agency for Science, Technology and Research (A*STAR)

Conditional-Copulas are best suited for synthetic data generation when the training datasets are moderately sized, often generating time-efficient and robust replication of the required data joint distributions. As compared to relatively costly machine-learning methods, which as a data-driven process is much reliant on the cardinality and size of the available training data, copulas provide a cost-effective alternative that balances data availability with prior expert knowledge, generating diverse sample sets based on pre-determined conditions for methodology testing and algorithm training.

The elliptical-copula-centric framework for synthetic data generation is a simple two-step process. In the first step, one estimates the marginal distributions of the input variables, followed by their pairwise-correlation parameters, then combining both to reproduce a statistical estimate of the joint distribution of the training dataset. The second step is relatively straightforward; one simply samples from the learned joint distribution to produce any number of synthetic sample points one requires, quite assured of its statistical properties with reference to what has been learned. The conditional-copula framework further enhances the former by fine-tuning the learning process; using the learned joint distribution as a baseline, one splits the training dataset into meaningful subsets based on identified conditions such as age groups, gender, races, etc., via reiteration of the learning process, one essentially resamples the generated datapoints with renewed conditional distributions (and conditions). This additional enhancement improves the flexibility of the copula-centric method and adapts it to complex training datasets with multi-modal distributions or even non-monotonic, non-linear relationships.

Detailed implementation and performance of this method is available at https://github.com/BiomedDAR/copula-tabular.

(C) Marginal-Based Data Synthesis

Contributed by Prof Xiao Xiaokui, School of Computing, National University of Singapore

Marginal-based data synthesis is a widely used approach for synthesising tabular data. This approach involves selecting a set of marginals from an input table *T*, each being a project of *T* onto a subset of its attributes. For instance, consider the following table *T* with 4 attributes: Age, Gender, Education, Occupation, and Income.

Age	Gender	Education	Occupation	Income
…	…	…	…	…

Table *T*

Here are a few examples of possible marginals of *T*:

Age	Gender	Education
…	…	…

Marginal of *T* on {Age, Gender, Education}

Age	Occupation	Income
…	…	…

Marginal of *T* on {Age, Occupation, Education}

Gender	Occupation
…	…

Marginal of *T* on {Gender, Occupation}

After choosing a set of marginals, the approach constructs a statistical model (e.g., a Bayesian network[28]) to capture the correlations among the attributes within the marginals. This model is then used to generate synthetic data that preserves attribute correlations.

Marginal-based data synthesis has three advantages:

- Simplicity: the concept is simple and easy to grasp.
- Effectiveness: when the chosen marginals cover all important attribute correlations, the synthetic data could preserve the statistical properties of the original data.

[28] "Bayesian Network," Wikipedia, 2024, https://en.wikipedia.org/wiki/Bayesian_network

- Privacy: the data synthesis process could offer strong privacy protection, if noise is carefully introduced during the selection and construction of marginals and the training of the statistical model.

Marginal-base data synthesis has gained widespread adoption in practical applications. Representative methods include PrivBayes[29], MST[30], and PrivMRF[31]. In particular, PrivBayes and MST were among the winners of the 2018 NIST Differential Privacy Synthetic Data Challenge[32], while PrivMRF won the first place in the 2020 edition of the challenge[33]. In addition, PrivBayes has been implemented in SAP's Data Intelligence Cloud[34] as well as numerous open-source data synthesis tools[35].

(D) Sequential Tree-based Synthesisers (SEQ)

Contributed by Dr Khaled El Emam, University of Ottawa

One way to generate synthetic data is to apply decision tree sequentially built on commonly used regression and classification trees ("CART") algorithms, although variants (e.g., boosted trees) of these can also be used. The principle

[29] Jun Zhang et al., "PrivBayes: Private Data Release via Bayesian Networks," in *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, 2014, 1423–34, https://dl.acm.org/doi/10.1145/2588555.2588573

[30] Ryan McKenna, Gerome Miklau, and Daniel Sheldon, "Winning the NIST Contest: A Scalable and General Approach to Differentially Private Synthetic Data," *Journal of Privacy and Confidentiality* 11, no. 3 (2021), https://journalprivacyconfidentiality.org/index.php/jpc/article/view/778

[31] Kuntai Cai et al., "Data Synthesis via Differentially Private Markov Random Fields," Github, n.d., https://github.com/caicre/PrivMRF

[32] National Institute of Standards and Technology, "Disassociability Tools," NIST, 2023, https://www.nist.gov/itl/applied-cybersecurity/privacy-engineering/collaboration-space/focus-areas/de-id/tools#dpchallenge

[33] National Institute of Standards and Technology, "2020 Differential Privacy Temporal Map Challenge," NIST, 2022, https://www.nist.gov/ctl/pscr/open-innovation-prize-challenges/past-prize-challenges/2020-differential-privacy-temporal

[34] SAP Community, "SAP Data Intelligence: Data Synthesizer for Machine Learning Operator," Technology Blogs by SAP, 2021, https://community.sap.com/t5/technology-blogs-by-sap/sap-data-intelligence-data-synthesizer-for-machine-learning-operator/ba-p/13501498

[35] "Reprosyn: Synthesising Tabular Data," Github, 2022, https://github.com/alan-turing-institute/reprosyn; "Synthcity," Github, 2024, https://github.com/vanderschaarlab/synthcity; "DataSynthesizer," Github, 2023, https://github.com/DataResponsibly/DataSynthesizer; DataCebo, "SDGym," Github, 2024, https://github.com/sdv-dev/SDGym; "DPART | Differentially Private Auto-Regressive Tabular," Github, 2024, https://github.com/hazy/dpart

is to sequentially synthesise variables using classification and regression models.[36]

The process can be thought of as initially fitting a series of models. These models make up the generator. Then, these models can be used to generate data. When a model is used to generate data, we sample from the predicted terminal node to get the synthetic values. The distribution of in the node can be smoothed before sampling.

Refer to the footnote for relevant papers on this method.[37]

Deep Generative Models

(E) Generative Adversarial Networks (GANs), contributed by Betterdata.ai

Generative Adversarial Networks (GANs) are deep generative models that excel in synthesising complex, high dimensional datasets. Through an adversarial process, the generator creates synthetic data which a discriminator evaluates for realism, prompting a continual improvement in the synthetic output. This iterative refinement enables GANs to produce synthetic data that closely resembles the original, outperforming non-deep learning techniques in complex real-world datasets.

GANs also demonstrate the ability to handle different data structures commonly found in enterprise settings. The development of specialised models like CTGAN and CTABGAN+ for static tabular data, TimeGAN for time series data and IRG for relational data highlights the adaptability of GANs in diverse data settings.

(F) Language Models

Originally developed for natural language processing tasks, Transformers and large language models (LLMs) have also proven to be effective in synthesising tabular data. These models use the attention mechanism to understand

[36] For more information, refer to Khaled El Emam, Lucy Mosquera, and Richard Hoptroff, "Evaluating Synthetic Data Utility," in *Practical Synthetic Data Generation Balancing: Privacy and the Broad Availability of Data* (O'Reilly Media, Inc, 2020).

[37] Khaled El Emam, Lucy Mosquera, and Chaoyi Zheng, "Optimizing the Synthesis of Clinical Trial Data Using Sequential Trees," *Journal of the American Medical Informatics Association* 28, no. 1 (2020): 3–13.

complex relationships within data, making them ideal for creating synthetic datasets that mirror the complexity of the real world.

LLMs also excel when original data is limited. Leveraging extensive pre-trained knowledge to fill in gaps in sparse original data and generate rich data in data scarce environments. However, while LLMs offer remarkable capability in tabular data synthesis, they require substantial computational power and time to train, presenting a trade-off.

Another significant benefit of deep generative models is the ability to integrate Differential Privacy Stochastic Gradient Descent (DP-SGD), which introduces noise during training to ensure data privacy with provable guarantees. However, it is important to note that while DP-SGD enhances privacy, it can also limit the utility of the generated data, presenting a trade-off between privacy protection and data usefulness.

Annex D: Re-identification Risks

As synthetic data generally tries to retain the statistical properties and characteristics of its source data, adversaries can attempt to re-identify or extract sensitive information about an individual from the synthetic data. The following describes the different types of re-identification attacks (commonly referred to as privacy attacks) on synthetic datasets.

(A) Singling Out attack

Singling out attack is generally conducted for outliers, e.g., unique attribute(s), rare data attribute(s) or unique combination of attributes. As the generated synthetic datapoints attempt to reflect or capture the presence and characteristics of such outliers, they offer a heightened possibility of singling out unique data records, and outliers are especially susceptible. While singling out may not represent a re-identification risk by itself, it may allow the adversary to gain information about the data record through using related datasets or other background information (see example in linkability attack).

(B) Linkability attack

For a linkability attack to occur, the adversary is assumed to have access to two sets of data i.e., (i) synthetic data and (ii) other publicly available data or private datasets where the adversary has privileged access. In a linkability attack, the adversary attempts to determine if any data points from the two data sets belong to the same individual, or group of individuals.

For example, an adversary might conclude that in the synthetic dataset of patients in a community hospital (through singling out) that there is a high possibility of exactly one individual who is male, above 80 years of age, has diabetes, and is earning an annual income of $100,000 to $200,000. A successful attack occurs when the adversary correctly guesses that the synthetic data is trained from a dataset containing the data record of an 86-year-old male patient with diabetes, from the patient's social media account linked to the community hospital and private knowledge that there is no other male diabetic patient above the age of 80 years old in that community hospital. The adversary now has additional knowledge about this patient, i.e., he has an annual income of $100,000 to $200,000.

The linkability attack assessment examines if the additional availability of synthetic data improves an adversary's ability to form linkages between different datasets. Intuitively, the adversary’s chances of a successful attack are likely to improve when data utility of the generated synthetic data increases, i.e., the closer it resembles the statistical characteristics of the source data, the higher the chance of a successful attack.

Therefore, it is important that data protection practices are incorporated during data preparation process for synthetic data generation. It is also imperative for any assessment of re-identification risks to differentiate between improvements of data utility that are desirable, i.e., to resemble general population trends that do not betray an individual’s involvement in a source dataset, or undesirable, i.e., resulting in increased re-identification risks of some individuals in the source dataset.

(C) Inference attack

The adversary is assumed to have access to a set of data attributes common to the source dataset and uses the information present in the synthetic data to infer sensitive attributes (e.g., other medical complications) about individual(s) in the source dataset.

For instance, a successful attack occurs when an adversary can infer with high confidence that an 86 years-old male with diabetes (from the source dataset of the community hospital) has other medical complications such as hypertension.

In an inference attack, we are examining if the additional availability of synthetic data would lead to a higher probability of successful inference with regards to sensitive attributes about the individual(s) in the source dataset. As before, any synthetic data with sufficient utility could be expected to improve an adversary's success rate.

Importantly, this observation can apply to any person belonging to the same distribution (e.g., males above 80 years of age with diabetes), even when his data has never been used for training.

Therefore, an inference attack assessment should measure the re-identification risk and then compare the incidence of successful attacks against some established baseline, for instance, people outside the source dataset. In such a scenario, one is measuring if the probability of successfully inferring data of someone in the source dataset is higher/lower than inferring data of someone not in the source dataset, so as to isolate and quantify the privacy leakage to individuals on top of identified population trends.

Annex E: Examples of Approaches to Evaluate Re-identification Risks

This annex introduces different approaches to evaluate re-identification/privacy risks adopted by three industry members. These approaches can be applied to synthetic data regardless of the generation method used.

(A) Approach 1

Contributed by Dr Khaled El Emam, University of Ottawa

Attribution disclosure. This is an extension of the traditional notion of identity disclosure to synthetic data. It considers the similarity of a synthetic record to a real record, conditional upon the identity disclosure risk of the original (real) dataset).[38] Conceptually, this evaluates the extent to which an adversary would learn something new about an individual by finding a record that looks like them (i.e., has the same values on the indirect identifiers) in the synthetic data. Attribution disclosure can be interpreted as a probability and an acceptable value of 0.09 (within the range defined in ISO/IEC 27599) is often used.

For computation of the attribution disclosure, the following article describes the process in depth: https://www.jmir.org/2020/11/e23139/.

Membership disclosure. This evaluates the extent to which an adversary would learn that an individual from the same population as the real data was included in the training dataset for the generative model.[39] Knowledge that someone is in the training dataset can reveal something about the target individual if the training dataset has a defining characteristics (e.g., they were all people with a particular disease). This can be defined as a relative F1 score measuring accuracy in determining membership corrected for a naïve determination, with a typical value of relative F1 = 0.2 used as a threshold.

For the membership disclosure, the following article describes the details of the calculation: https://academic.oup.com/jamiaopen/article/5/4/ooac083/6758492?searchresult=1.

[38] Khaled El Emam, Lucy Mosquera, and J. Bass, "Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation," *Journal of Medical Internet Research* 22, no. 11 (2020): e23139.

[39] Khaled El Emam, Lucy Mosquera, and Xi Fang, "Validating A Membership Disclosure Metric For Synthetic Health Data," *Journal of the American Medical Informatics Association* 5, no. 4 (2022): 00ac083.

References

Emam, Khaled El, Lucy Mosquera, and J. Bass. "Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation." *Journal of Medical Internet Research* 22, no. 11 (2020): e23139.

Emam, Khaled El, Lucy Mosquera, and Xi Fang. "Validating A Membership Disclosure Metric For Synthetic Health Data." *Journal of the American Medical Informatics Association* 5, no. 4 (2022): 00ac083.

Emam, Khaled El, Lucy Mosquera, Xi Fang, and Alaa El-Hussuna. "Utility Metrics for Evaluating Synthetic Health Data Generation Methods: Validation Study." *JMIR Medical Informatics* 10, no. 4 (2022).

Kababji, Samer El, Nicholas Mitsakakis, Xi Fang, Ana-Alicia Beltran-Bless, Greg Pond, Lisa Vandermeer, Dhenuka Radhakrishnan, and Khaled El Emam. "Evaluating the Utility and Privacy of Synthetic Breast Cancer Clinical Trial Data Sets." *JCO Clinical Cancer Informatics* 7 (2023). https://ascopubs.org/doi/full/10.1200/CCI.23.00116

Yang, S. "Process Mining the Trauma Resuscitation Patient Cohorts." In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, 29–35, 2018.

(B) Approach 2

Contributed by A*Star

In Approach 2, the assessment and evaluation framework for quantifying privacy risk in synthetic data are detailed in https://github.com/statice/anonymeter, wherein the proposed set of methods will generate three privacy risk threshold scores based on attack-based evaluations for the three major risks, mainly (i) singling Out attack, (ii) linkability attack, and (iii) inference attack.

As compared to ML-based alternatives, the framework provides a computationally efficient and statistically robust method for measuring privacy risks. As a first step, we split the source dataset into two disjoint subsets, namely (i) control (20-30%) and (ii) training (70-80%), with the former performing the role of a pseudo set of public individuals and the latter as the training dataset. Clearly, the control dataset can never be used for training, but instead, serves as a privacy attack baseline measuring the ability of the synthetic dataset to "infringe" the privacy of individuals it has never used. In other words, the control dataset represents the population that has not contributed to the synthetic data generation process, and successful knowledge gained from attacking the control dataset is, to a certain extent, some measure of the synthetic dataset's "utility".

To that end, two separate attacks were performed, namely the (i) control attack and the (ii) main attack. The control attack targets the control dataset and measures patterns common to the whole population; the main attack targets the training dataset and measures patterns common to the whole population and possible biases towards the training dataset. The computed asymmetry between the two attacks provides a fair measurement of how effective the synthetic data is in differentiating individuals in the training dataset from the larger population, while grounding the obtained privacy-risk metric with some reasonable baseline from which to make further interpretations.

Lastly, the framework also measures a "naïve" baseline that assumes no prior knowledge of the synthetic dataset and is therefore, entirely dependent on luck. This closes a loophole where one might erroneously assume that the generated synthetic dataset is risk-free because it has extremely poor fidelity/utility and/or when the designed inference/linkabililty attacks or synthetic data is insensible in the first place. In these scenarios, the "naïve" attack might outperform the other two attacks, indicating that the test is flawed.

The computed asymmetry between the main and control attack is normalised to obtain a privacy risk leakage metric, known as "R". This value is bounded between 0 and 1 and increases with the risks of privacy leakage. It is reasonable to first decide on an acceptable threshold value of "R" before generating the synthetic data; reversal of this process exposes one to considerable latitude in justifying one's product. The said threshold can be fixed based on policy, and further mitigated based on the sensitivity of the training dataset and the availability of the generated synthetic dataset.

It is crucial to note that the privacy risks are evaluated with respect to the individuals in the training database, and not the wider public. As such, privacy is compromised when an adversary finds it easier to (i) determine if an individual belongs to the training database and (ii) derive details of an individual from the training database otherwise undisclosed.

References

For more details, a description of the framework and the attack algorithms can be found in the paper by M. Giomi et al. "A Unified Framework for Quantifying Privacy Risk in Synthetic Data." *In Proceedings on Privacy Enhancing Technologies Symposium (PETS 2023),* 2023.

(C) Approach 3

Contributed by Betterdata.ai

Approach 3 audits the privacy integrity of the end-to-end Synthetic Data Generation (SDG) pipeline and not only the generated synthetic data. This approach is based on Differential Privacy (DP)[40], which provides a mathematical quantification of individual privacy. DP quantifies the risk of someone deducing that specific personal data was included in the training dataset by looking at the synthetic data. The privacy loss is calculated using two parameters $\varepsilon>0$ and $0\leq\delta\leq1$, where ε represents the maximum allowable privacy loss and δ represents the acceptable tolerance of this privacy loss being exceeded which is generally kept close to zero.

The audit process begins by identifying the most sensitive outliers in the source dataset. These outliers have a 50% chance of being randomly excluded from the data used in the SDG pipeline. The next step involves conducting membership inference attacks on the synthetic data to attempt identification of these outliers. The success of these attacks, particularly how it exceeds a 50% baseline (random guess), indicates a potential privacy leakage. This sets the lower bound for the actual privacy loss. Using the award-winning privacy audit analysis developed by Steinke, Nasr, and Jagielski[41], we convert membership attack statistics into high-confidence lower bounds on the privacy budget ε for tolerance δ.

For detailed methodologies on conducting membership inference attacks, refer to the TAPAS framework[42].

Organisations can tailor their privacy budget, ε, to their specific requirements, reflecting the sensitivity of their data and aligning with industry benchmarks. The following are examples of publicly reported privacy budgets used in real-world applications:

[40] Cynthia Dwork et al., "Calibrating Noise to Sensitivity in Private Data Analysis," in *Theory of Cryptography. TCC 2006. Lecture Notes in Computer Science, Vol 3876*, ed. S. Halevi and T. Rabin (Berlin: Springer, 2006); Cynthia Dwork and Aaron Roth, "The Algorithmic Foundations of Differential Privacy," *Foundations and Trends® in Theoretical Computer Science* 9, no. 3–4 (2014): 211–407.

[41] Thomas Steinke, Milad Nasr, and Matthew Jagielski, "Privacy Auditing with One (1) Training Run," in *NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems*, ed. A. Oh, T. Naumann, and A. Globerson (Curran Associates Inc., 2023), 49268–80, https://dl.acm.org/doi/10.5555/3666122.3668265

[42] TAPAS, "Welcome to TAPAS's Documentation!," tapas, 2022, https://tapas-privacy.readthedocs.io/en/latest/index.html

Organization Name	**Data Type**	**DP Budget (ε)**	**Collection Period**	**Purpose of Data Collection**
Apple [5,6]	Health Data Safari Emoji QuickType	2.0 4.0 4.0 8.0	2017-2024	Analytics
2020 US Census Data [7,8]	Housing Unit Data Person's File	2.47 17.14	2020	Deciding Fund Distribution, Assisting States

For more details, please refer to Betterdata.ai URL at *How it works* (betterdata.ai).

References

Abowd, John M. "The US Census Bureau Adopts Differential Privacy." In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.

Apple Privacy Team. "Differential Privacy." Apple.com, n.d. https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf

Dwork, Cynthia, F. McSherry, K. Nissim, and A. Smith. "Calibrating Noise to Sensitivity in Private Data Analysis." In *Theory of Cryptography. TCC 2006. Lecture Notes in Computer Science, Vol 3876*, edited by S. Halevi and T. Rabin. Berlin: Springer, 2006.

Dwork, Cynthia, and Aaron Roth. "The Algorithmic Foundations of Differential Privacy." *Foundations and Trends® in Theoretical Computer Science* 9, no. 3–4 (2014): 211–407.

Houssiau, F, J Jordon, SN Cohen, O Daniel, A Elliot, J Geddes, C Mole, and C Rangel-Smith. "Tapas: A Toolbox for Adversarial Privacy Auditing of Synthetic Data." *arXiv Preprint ArXiv:2211.06550*, 2022.

National Conference of State Legislatures. "Differential Privacy for Census Data Explained." National Conference of State Legislatures, 2021. https://www.ncsl.org/technology-and-communication/differential-privacy-for-census-data-explained

Steinke, Thomas, Milad Nasr, and Matthew Jagielski. "Privacy Auditing with One (1) Training Run." In *NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems*, edited by A. Oh, T. Naumann, and A. Globerson, 49268–80. Curran Associates Inc., 2023. https://dl.acm.org/doi/10.5555/3666122.3668265

Tang, Jun, Aleksandra Korolova, Xiaolong Bai, Xueqiang Wang, and Xiaofeng Wang. "Privacy Loss in Apple's Implementation of Differential Privacy on MacOS 10.12." *ArXiv:1709.02753*, 2017. https://arxiv.org/abs/1709.02753

TAPAS. "Welcome to TAPAS's Documentation!" tapas, 2022. https://tapas-privacy.readthedocs.io/en/latest/index.html

ACKNOWLEDGEMENTS

The PDPC and Infocomm Media Development Authority (IMDA) sincerely extends their appreciation for the editorial contributions in the development of this publication from the following:

- Betterdata.ai
- BioMedical Data Architecture & Repository (DAR), BMRC, A*STAR
- Dr Khaled El Emam, University of Ottawa
- School of Computing, National University of Singapore (NUS)

PDPC and IMDA also express their appreciation and acknowledgment for all the valuable feedback received from the following organisations:

- Data Privacy and Protection Capability Centre, Government Technology Agency (GovTech)
- Johnson & Johnson (J&J)
- Mastercard
- Ministry of Health (MOH)
- Nanyang Technological University, Singapore (NTU Singapore)

The following guides/articles were referenced in this guide:

Agencia Espanola Proteccion Datos. "Synthetic Data and Data Protection." Blog, 2023. https://www.aepd.es/en/prensa-y-comunicacion/blog/synthetic-data-and-data-protection

Financial Conduct Authority (U.K.). "Exploring Synthetic Data Validation – Privacy, Utility and Fidelity." Publications, 2023. https://www.fca.org.uk/publications/research-articles/exploring-synthetic-data-validation-privacy-utility-fidelity

Giomi, Matteo, Franziska Boenisch, Christoph Wehmeyer, and Borbala Tasnadi. "A Unified Framework for Quantifying Privacy Risk in Synthetic Data." In *Proceedings on Privacy Enhancing Technologies Symposium Issue 2*, 312–28, 2023. https://petsymposium.org/popets/2023/popets-2023-0055.php

Information Commissioner's Office (U.K.). "Chapter 5: Privacy-Enhancing Technologies (PETs)." ICO call for views: Anonymisation, pseudonymisation and privacy enhancing technologies guidance, 2022. https://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/ico-call-for-views-anonymisation-pseudonymisation-and-privacy-enhancing-technologies-guidance/

———. "G7 DPAs' Emerging Technologies Working Group Use Case Study on Privacy Enhancing Technologies." UK GDPR guidance and resources, n.d. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/privacy-enhancing-technologies/case-studies/g7-dpas-emerging-technologies-working-group-use-case-study-on-privacy-enhancing-technologies/

Jordon, James, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N. Cohen, and Adrian Weller. "Synthetic Data -- What, Why and How?" ArXiv:2205.03257, 2022. https://arxiv.org/abs/2205.03257

END OF DOCUMENT

JOINTLY DEVELOPED BY

pdpc
PERSONAL DATA
PROTECTION COMMISSION
SINGAPORE

Agency for
Science, Technology
and Research
SINGAPORE

The contents herein are not intended to be an authoritative statement of the law or a substitute for legal or other professional advice. The PDPC and its members, officers, employees and delegates shall not be responsible for any inaccuracy, error or omission in this publication or liable for any damage or loss of any kind as a result of any use of or reliance on this publication.