Government of **Western Australia**
Department of **Health**

WA Department of Health

Synthetic Data User Guide

health.wa.gov.au

Contents

Purpose

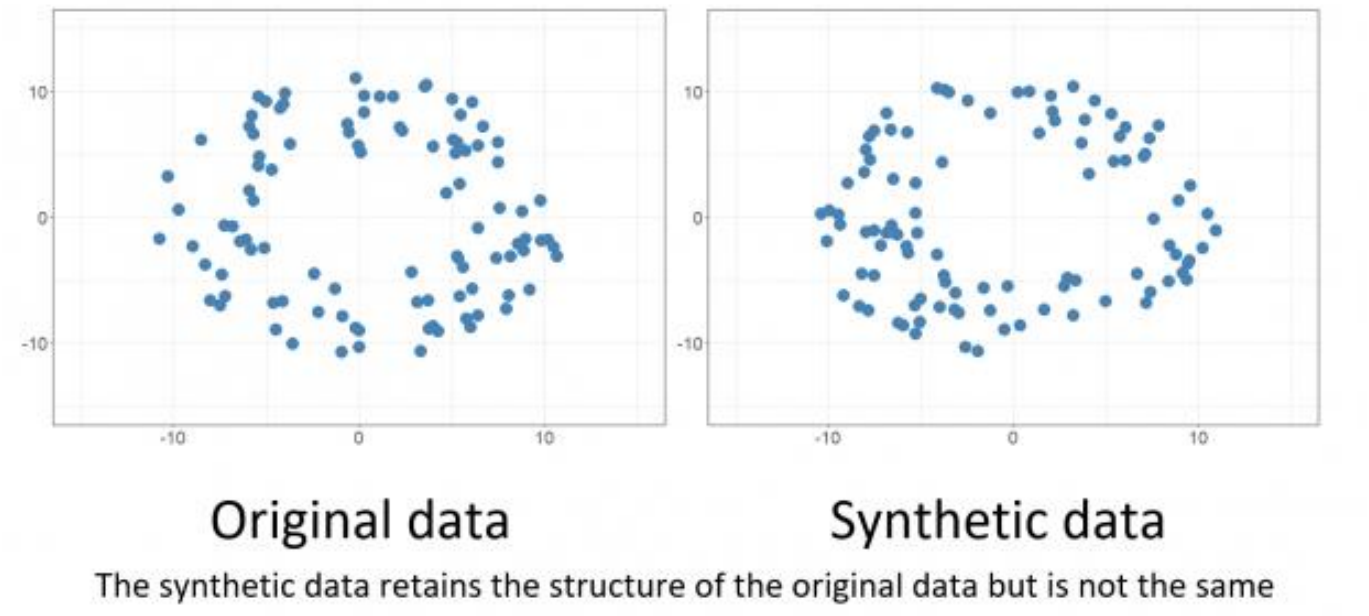
The WA Department of Health (the department) has generated several synthetic datasets from real health datasets as an artificial alternative to real-world data. This document provides potential users with the necessary information to understand the utility of the department's synthetic data offerings and help determine whether they will be suitable for the potential user's intended purpose.

Background

What is synthetic data?

Synthetic data is artificially manufactured data that has been generated using a purpose-built mathematical model (including artificial intelligence (AI) and ML models) or algorithm. It has been derived by training a model (or algorithm) on a real dataset to mimic the characteristics and structure of the real data, however it does not contain any real or identifiable information. Each piece of information in the synthetic dataset is designed to be plausible but is created at random based on the structure of the original, real data.

Figure 1. (Source: UK Government)



Masked data is sometimes offered as an alternative to real data. Masked data is when the original data is amended, modified or transformed. Data that has been de-identified is masked data. Masked data is not synthetic data as it is a modified version of the original data. A key disadvantage of masked data is that there is a risk the masked data could be transformed back into the original data or be re-identified.

Why is synthetic data important?

The Department collects, stores, and delivers extensive volumes of health data which can be used at all levels of the health system to improve access and quality of care. Yet accessing real data comes with challenges related to privacy, security, regulations, and timely data accessibility. Synthetic data offers a solution by enabling the creation of high-quality, privacy-compliant data without the risks associated with real data.

Benefits

The WA health system enables safe information sharing through strong governance and security measures that protect individual privacy. Sometimes, accessing real data takes time to ensure these protections are in place. In such cases, synthetic data provides a high-quality, lower-risk alternative.

- *More timely access to data:* synthetic datasets facilitate efficient data sharing as they are quality assured and pre-approved for certain use cases. This leads to reduced wait times for users seeking timely access to representative healthcare data.

- *Maintained privacy and security:* synthetic data can replicate important statistical properties of real health information without releasing real data, thereby reducing the risk that individuals will be identified in data breaches, data leaks and ransomware attacks.

- *Enhanced collaboration and sharing:* synthetic data supports collaboration by offering a privacy-protected alternative to real data. This accelerates the pace of innovation as data scientists, analysts, and developers can collaborate without facing barriers imposed by data privacy concerns. The provision of quality representative synthetic data for public innovation can support the translation of data into actionable solutions that can help reform the WA healthcare landscape.

- *Assistance for the research community:* synthetic data contains the same field names, value domains, and general trends of real data, which allows researchers to test their methodologies or commence the creation of analytical scripts in preparation for the receipt of real data.

Synthetic data fidelity

The fidelity of synthetic data is dependent on how it is generated and how closely it mirrors the real data from which it was created. High fidelity (representative) datasets share, and deliberately conserve, many of the features of the original dataset from which they are created. This may include complex relationships between different variables. Low fidelity (non-representative) datasets while conserving some key characteristics, does not mirror as closely the real data from which they are created. Fidelity is calculated by comparing it to the real data through statistical and analytical tests. This includes an assessment of how well the synthetic data preserves key statistical properties, such as means, variances, and correlations between variables. The level of fidelity required may differ depending on the use of the data. Not all use cases will require access to high fidelity synthetic datasets.

Synthetic data utility

The utility, or value, of a synthetic dataset indicates its effectiveness for the specific purpose or use case for which it is intended. Fidelity and utility are partially overlapping qualities but can be evaluated separately. A low fidelity dataset could provide high utility for an education or training purpose where the dataset need not be statistically representative, only structurally representative. However, where the dataset requires a more accurate representative of the source data, then high fidelity is required to achieve high utility.

How is personal data kept safe?

The real datasets utilised to generate the synthetic datasets, have previously undergone a process known as de-identification to ensure anonymity. Furthermore, the techniques employed to generate synthetic data results in the formation of fabricated events and identities that do not correspond to any actual circumstance or person.

Representative (high fidelity) datasets have gone through additional assessments to make sure that there is no risk of personal information being inadvertently shared. Non-representative (low fidelity) datasets that do not maintain patterns across the dataset are already very low risk.

Synthetic dataset offerings

The synthetic datasets listed below have been created at the Department and can be requested for access.

- Non-representative*:* the most secure synthetic data product that does not maintain data distributions however retains data structure (improbable pairs are not present in the dataset).
- Representative*:* A data product that offers significant utility while maintaining low risk to privacy. It focuses on individual datasets, ensuring the replication of variable distribution, internal relationships, and data structure.
- Linked Representative: A data product encompassing all the attributes of a representative dataset while incorporating inter-table relationships. It presents a minimal privacy risk alongside significant utility.

The representative synthetic datasets offered by the department, have conserved many, but not all, relationships between variables. Please refer to Limitations of Synthetic Data

Table 1: Synthetic datasets offered by the WA Department of Health

Synthetic Data Offerings	Non-Representative (6-8 business days for access)	Representative (6-11 business days for access)	Linked Representative (9-11 business days for access)
Emergency Department Data Collection 2023	✗	✓	✗
Emergency Department Data Collection 2022	✓	✓	✓
Hospital Morbidity Data Collection 2022	✓	✓	
Melanoma 2010-2020 WA Cancer Registry	✓	✓	✗
Stomach Cancer 2010-2022 WA Cancer Registry	✗	✓	✗

Requests for Access

Requests for access to the Department's synthetic datasets can be submitted through the Synthetic Data Access Request Form. The Data Custodians responsible for their respective synthetic datasets will oversee the approval process for the provision of synthetic data in accordance with the Information Access, Use and Disclosure Policy - MP 0015/16.

Information Release Agreement

All non-WA Health applicants will need to agree to abide by the conditions outlined in the department's Synthetic Data Information Release Agreement. The agreement conditions are outlined in the access request form mentioned above. Applicants will need to accept the conditions before they can submit their access request.

Unsupported Use

Due to various limitations, synthetic data is deemed unsuitable for the following purposes.

- *Research final analysis and validation* - A research initiative that requires validation or the dissemination of results must obtain access to the source data and seek approval from the appropriate Custodian, HREC and any other approving authority in accordance with the Research Governance Policy, MP 0162/21, or any other application policy or procedure. HREC and Research Governance approvals are project specific. Research related requests are managed by the ISPD Client Service team ISPDClientservices@health.wa.gov.au.

  Synthetic data can however be used as a preliminary tool for data exploration and hypotheses testing (refer to Acceptable Use).

- *Decision-making purposes* - As a result of privacy-preservation methods such as the removal of outliers, synthetic data may not accurately reflect the true distribution of real-world patient data. This may lead to misleading conclusions when used to inform clinical decisions, policy development, or resource allocation. Additionally, biases or limitations in the data generation process can introduce artifacts that do not exist in real populations, potentially undermining the reliability and validity of any insights drawn from such data. As a result, real decisions require real data.

- *Bulking up existing datasets* - Synthetic data is not suitable for bulking up existing datasets because it can introduce biases, distort underlying patterns, and reduce the overall quality of the data. While synthetic data is generated to mimic real-world data, it may lack the complexity and diversity of edge cases found in real datasets. It may reinforce existing biases or overfit certain trends, leading to misleading model performance during training and evaluation. Additionally, synthetic data may not accurately capture rare or unexpected scenarios that are critical for robust generalisation, ultimately compromising the reliability and fairness of machine learning models built on such augmented datasets.

Acceptable Use

The department supports several use cases to access synthetic data. Although an application may align with a supported use case, applications to access will still undergo a technical feasibility assessment and a review by the relevant data custodian, unless

predetermined conditions have already been satisfied. As the synthetic datasets contain both valid and invalid data (refer to 'Limitations of synthetic data'), only certain applications will be approved to avoid incorrect inferences being drawn.

Where the recommended synthetic data level is 'non-representative' and an applicant is requesting a 'representative' synthetic dataset, appropriate justification will need to be provided to demonstrate the necessity for the higher level.

A potential user can still apply for access if their intended use falls outside of the use cases detailed in below. The relevant data custodian will review and may approve if the specified purpose is appropriate, and the synthetic dataset is deemed suitable.

Table 2. Synthetic data use cases supported by the WA Department of Health

Use Case	Key Benefits	Recommended Level
1. Education and Training Synthetic data can be valuable for training and teaching purposes, allowing users to experiment with realistic datasets without sacrificing the privacy of real patient data.	• Allows learners to visualize and experiment with realistic datasets. • Enables demonstration and hands on training in a real-world context without sharing sensitive data with trainers and trainees.	Non-representative (Low Fidelity)
2. Testing and Validation The use of synthetic test data offers a secure, compliant, and efficient alternative to production test data, mitigating the risks associated with handling real patient data in testing environments while fostering enhanced data optimization and enrichment.	• Streamlines the testing process by providing data quickly, reducing the time spent waiting for real data to test different functionalities or systems. • Allows the creation of diverse and complex testing scenarios that might not be possible with real data, enabling more comprehensive testing for different scenarios, including edge cases. • Minimizes the risk of exposing real patient information to potential breaches or unauthorized access during testing procedures.	Non-representative (Low Fidelity)
3. AI and Machine Learning Model Development Synthetic data can be a critical enabler for training high-performing machine learning models. By supplementing real-world datasets, it ensures diverse and balanced training inputs.	• 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 restrictions, and proprietary rights. • Provides a controlled environment that mimics real-world scenarios, so models can be fine-tuned without compromising sensitive information. • Enhances the accuracy and robustness of predictive models, ultimately	Representative (High Fidelity)

	contributing to more informed decision-making.	
4. Data Exploration and Hypotheses Testing (Not to be used for publication of findings) While real decisions are always made using real data, synthetic data can be a valuable resource, enabling researchers to make progress on their projects, validate their methods, and gain insights into the data they expect to work with.	• Speeds up the research process by allowing researchers to begin working on their project before they are granted access to real data. • Helps researchers understand the structure and common properties of the data which enables testing of research hypotheses, conducting of preliminary evaluations and the development of algorithms/search queries. • Assists in the planning processes required to seek ethical approvals or file the data permit applications.	Representative (High Fidelity)
5. Privacy Preserving Data Sharing Synthetic data can facilitate collaboration and provide stakeholders with a representative preview of the source data without exposing sensitive information.	• Facilitates ethically aligned data sharing practices. • Accelerates the pace of innovation as data scientists, analysts and developers can collaborate without the barriers imposed by data privacy concerns.	Representative (High Fidelity)

Limitations of Synthetic Data

While synthetic data can offer advantages, it is essential to be aware of its limitations and carefully consider its appropriateness on a case-by-case basis.

Limitations at the WA Department of Health

At the department, the safeguarding of privacy and the preservation of fidelity are both of high importance. However, by adopting a risk adverse approach, synthetic data generated at the department is subject to specific limitations to balance the trade-off between privacy and fidelity.

Table 3: Identified limitations of the synthetic datasets generated at the WA Department of Health

Limitations of synthetic data at the WA Department of Health	
Outlier exclusion	Real data often contains outliers or edge cases that can be important to understanding a dataset. The process of synthetic data generation at the Department commences with outlier detection and implementation of privacy preservation techniques that may result in the deliberate exclusion of some outliers based on statistical metrics such as k-anonymity. The removal of outliers can introduce bias,

	resulting in the under-representation of minority groups or events prior to synthesis and subsequently in the generated synthetic data.
Bias and assumptions	The generation of synthetic data often relies on assumptions and predefined models or algorithms. These assumptions may introduce biases into the synthetic data, leading to inaccuracies and unrealistic representations. A synthesis model might assume that the real data precisely represents all real-world scenarios. If the source data fails to encompass specific representations due to restricted input, the resulting synthetic data will not accurately mirror real-world situations.
Inability to capture complex dependencies	Synthetic data generation at the department has undergone extensive evaluation by Data Scientists and subject matter experts. This process ensures that key within-table complex multidimensional patterns, between-table nonlinear relationships and patient trajectories have been preserved. However, synthesis techniques employed may not capture all the complex dependencies inherent in the data, which can result in inaccurate representations in certain scenarios.
Model dependency	The quality of synthetic data is heavily dependent on the machine learning models and algorithms used to generate it. Based on knowledge obtained during Phase 1 of the departments Synthetic Data Generation project, a benchmarking algorithm was selected as the synthesis model for linked synthetic data generation. Advancements in the field, will lead to the integration of new models into the departments synthetic data generation pipeline, potentially rectifying any shortcomings of earlier models.

Validation and Quality Assurance

The department has implemented a robust validation process to assess the effectiveness of its synthesis models and ensure the quality and reliability of the synthetic datasets generated. As a result, several common limitations of synthetic data have been mitigated.

Table 4: Synthetic data limitations mitigated by the WA Department of Health

Mitigated limitations of synthetic data at the WA Department of Health	
Threat of fake data	Ensuring that synthetic data accurately represents real-world scenarios is challenging. The department has defined fidelity metrics to differentiate fake data from synthetic data. The application of the fidelity metrics categorises the synthetic datasets into levels and identifies them as either representative or non-representative of the source data. The categorization is clearly depicted in the labelling of the synthetic dataset to minimise the potential for misinterpretation.

Limited transferability	The synthetic datasets generated by the department have undergone extensive validation to ensure the preservation of many of the complex relationships between variables. This approach optimises the applicability and transferability of synthetic data across different use cases. Furthermore, various levels of synthetic datasets have been generated to ensure that the most appropriate dataset is provided for each intended use.
Model collapse	The department have trained their models on de-identified real datasets removing the risk of model collapse. Model collapse occurs when models trained predominantly on synthetic data begin to lose their ability to generalize to real-world scenarios. As a model repeatedly learns from artificially generated patterns, it may overfit to the specific characteristics of the synthetic data, failing to capture the full complexity and variability of genuine data.
Lack of contextual information	The department specifically selected two linked health datasets for synthesizing with the aim of supporting health operations. The selected data elements are relevant to improving the quality of the public health service and facilitating the delivery of contextual information impacting KPIs of healthcare, for instance, discharge time of Emergency Department data and separation time of Hospital Morbidity data. Other health datasets or un-selected data elements within these two datasets are out of scope at this stage of synthesis.
Overreliance risk	There is a potential risk of overreliance on synthetic data, leading to a disconnect from real-world data and its evolving patterns. To mitigate this risk, the department has defined use cases and guidelines which prohibit the use of synthetic data for decision-making purposes. Real decisions require real data, with synthetic data being used in the preliminary stages only.

This document can be made available in alternative formats on request for a person with disability.

health.wa.gov.au