IWISH - Intelligent Workflow optimization and Intuitive System interaction in Healthcare

DELIVERABLE D2.2

Synthetic data state-of-the-art report

Project number:	AI2021-066
Document version no.:	v1.0
Edited by:	Simon Brouwer
Date:	24-02-2023

Eureka AI Thematic call

HISTORY

Document version #	Date	Remarks
V0.1	20-01-2023	Initial version by Simon Brouwer
V0.2	10-02-2023	Review by Marijn Vonk
V0.3	20-02-2023	Review by Younes Moustaghfir
V1.0	24-02-2023	Final version by Simon Brouwer

Table of Contents

1 Introduction

Synthetic data is generated by a computer algorithm that cultivates completely new and artificial data points – in contrast to original data, which is collected at its source from real individuals through measurements or registrations. To generate synthetic data, a generative machine learning model trains on the original data to capture its characteristics, relationships, and statistical patterns. This trained model can be used to generate synthetic records that mimics the properties of the original dataset. A generative machine learning model can generate entirely new synthetic data while preserving the characteristics, relationships and statistical patterns of the original data, to such an extent that the synthetic data can be used as if it is original data.

Recent advances in computing power and improved software have expanded the range of methods for generating synthetic data. This report provides an overview of the different types of models used for synthetic data generation, which are grouped into three categories: (1) simulated dummy data; (2) deep-learning methods; and (3) fully conditional specification. Each method will be discussed in a separate chapter.

1.1 Synthetic data in healthcare

Compelling initiatives regarding public synthetic data release have unfolded in the healthcare sector, for example, the synthetic version of the Netherlands Cancer Registry in October 2021[1]. The Integral Kankercentrum Nederland (“IKNL”) provides synthetic data for methodological studies in data analysis and software development, so that researchers can assess if data in the Registry meets their needs, before they submit a request to access real data.

Market-leading research company Gartner has predicted that: “*in 2024, 60% of the data used for the development of AI and analytics solutions will be synthetically generated*”. In particular, they cite the healthcare sector as having “*the biggest interest in synthetic data, since privacy laws are the strictest and non-compliance fines the highest*”.

[1] https://iknl.nl/en/ncr/synthetic-dataset

1.2 Legacy anonymization methods

In the past, legacy privacy measures such as k-anonymity and l-diversity (Machanavajjhala, 2007) (Sweeney, 2002) have been used to protect sensitive data from disclosure. These techniques modify or suppress certain data attributes to ensure the anonymity of the individuals in the data. Techniques used to achieve k-anonymity and l-diversity manipulate the original data to make re-identification harder. Examples of applied techniques are provided below in Table 1 - Classic anonymization techniques.

Example technique	**Original data**	**Manipulated data**
Generalization	27 years old	Between 25 and 30 years old
Suppression / Wiping	info@syntho.ai	xxxx@xxxxxx.xx
Pseudonymization	Amsterdam	hVFD6td3jdHHj78ghdgrewui6
Row and column shuffling	Aligned	Shuffled

Table 1 - Classic anonymization techniques

These legacy anonymization methods always result in a trade-off between privacy and utility (the “privacy utility trade-off”), because the data is manipulated to ensure better privacy. See Figure 1 - Privacy utility trade-off for legacy anonymization methods for an illustration of the privacy utility trade-off for an image, noting that the same principle holds for structured tabular data.

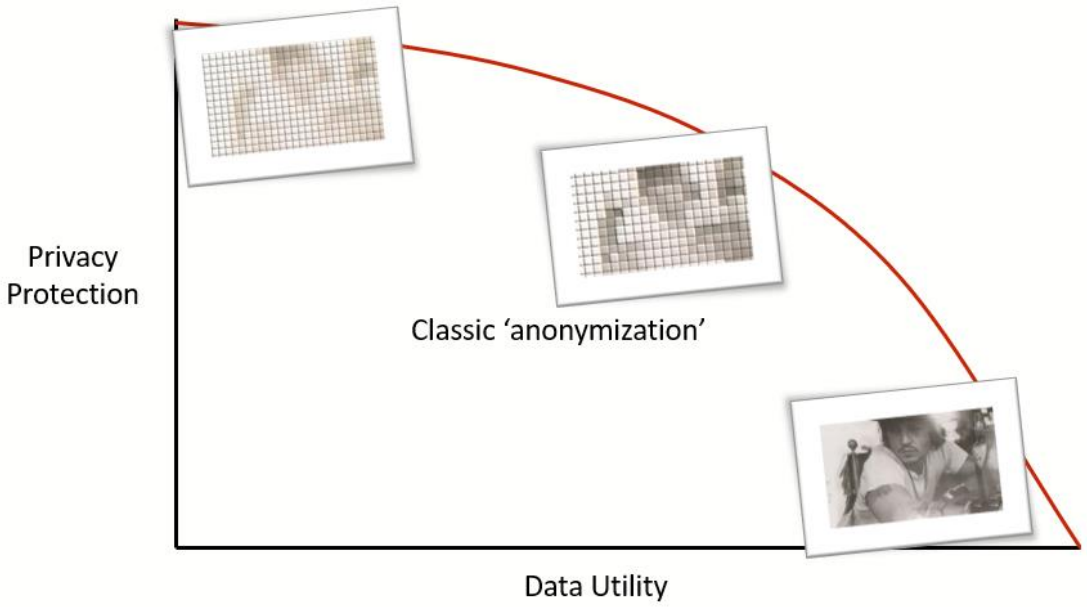

Figure 1 - Privacy utility trade-off for legacy anonymization methods

Studies have demonstrated that these legacy anonymization methods often do not provide privacy guarantees, and can even lead to re-identification (Sweeney, 2000), (Narayanan, 2006). A study by Sweeney established that by combining "anonymized" hospital visit data with a public voter database, it is possible to re-identify the vast majority of the hospital visitors by only their zip code, gender, and date of birth (Sweeney, 2000).

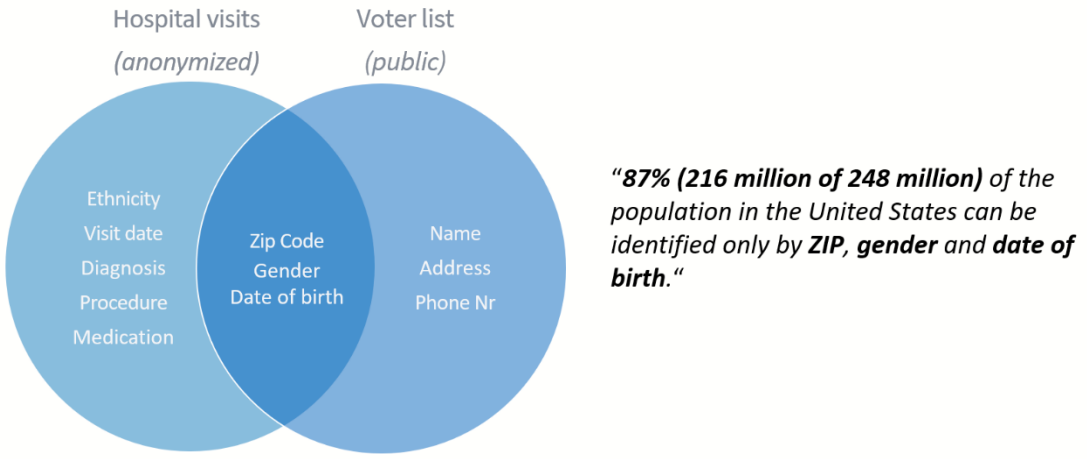

Figure 2 - Results re-identification study (Sweeney, 2000)

In addition, these methods may result in a loss of data utility, meaning that the anonymized data may not be useful for a wide variety of use-cases. As a result,

synthetic data has emerged as a state-of-the-art alternative to traditional anonymization methods.

2 (Simulated) dummy data

2.1 Introduction

Dummy data is synthetic data that is generated from scratch, meaning that no direct access to real data is needed to generate it. These methods are useful in scenarios where direct access to real data is unfeasible, or in cases where there is simply no data available yet. Dummy data is often used as a replacement for personal identifiable information (“PII”). PII is data that directly links to an individual person, such as names, addresses, and social security numbers. Dummy data can be created through the use of various methods, including random generation or statistical models that simulate the distribution of real data.

Often, dummy data may not be representative of real data and can lead to inaccurate results or conclusions if used inappropriately. It is therefore essential to ensure that the dummy data is used appropriately or closely approximates the real data, which can be a very time consuming endeavor.

2.2 Methods

2.2.1 Random dummy data

Random dummy data refers to the type of dummy data that is often used to replace PII. To generate this type of synthetic data, random samples are simply retrieved from a database. This type of synthetic data has no value for analytics purposes, and is mainly relevant for demos or software testing & development purposes.

2.2.2 Simulated dummy data

Simulated dummy data uses hand-crafted statistical models or computer simulations to generate synthetic data. For example, in the healthcare sector, the open-source synthetic data generator Synthea[2] (Walonoski J. K., 2018) is used to generate synthetic patients and medical records, providing a realistic representation of patient demographics and health conditions. Synthea demonstrates that simulated dummy

[2] https://github.com/synthetichealth/synthea

data can allow researchers to work with sensitive data without compromising patient privacy, as no actual patient data is ever used.

However, simulated dummy data methods also have limitations, such as the potential for bias and inaccuracies in the generated data. Simulated dummy data may not fully capture the complexities of real-world patient care, leading to limitations in the conclusions that can be drawn from the data.

The quality of the synthetic data generated by such methods is as strong as the assumptions the models are based on. To get these assumptions right can be a very time-consuming and tedious process, as often specific parameters in the models need to be estimated based on limited information.

3 Deep learning methods

3.1 Introduction

Deep learning methods have been widely used to generate synthetic data and are the main recent research focus for synthetic data generation methods. Two key architectures of deep-learning models are used: Generative Adversarial Networks ("GANs") and Variational Auto Encoders ("VAEs"). These models learn the underlying distribution of the original data and generate new data samples by sampling from the learned distribution. The open-source package Synthetic Data Vault ("SDV") [3] developed and maintained by MIT, provides both VAE and GAN architectures to generate synthetic data.

3.2 Methods

3.2.1 Generative Adversarial Networks

GANs (Goodfellow, 2014) consist of a generator network that generates synthetic data samples and a discriminator network that distinguishes between real and synthetic data samples. The two networks are trained simultaneously, with the generator trying to fool the discriminator and the discriminator trying to correctly classify the data samples as either real or synthetic data.

[3] https://github.com/sdv-dev/SDV

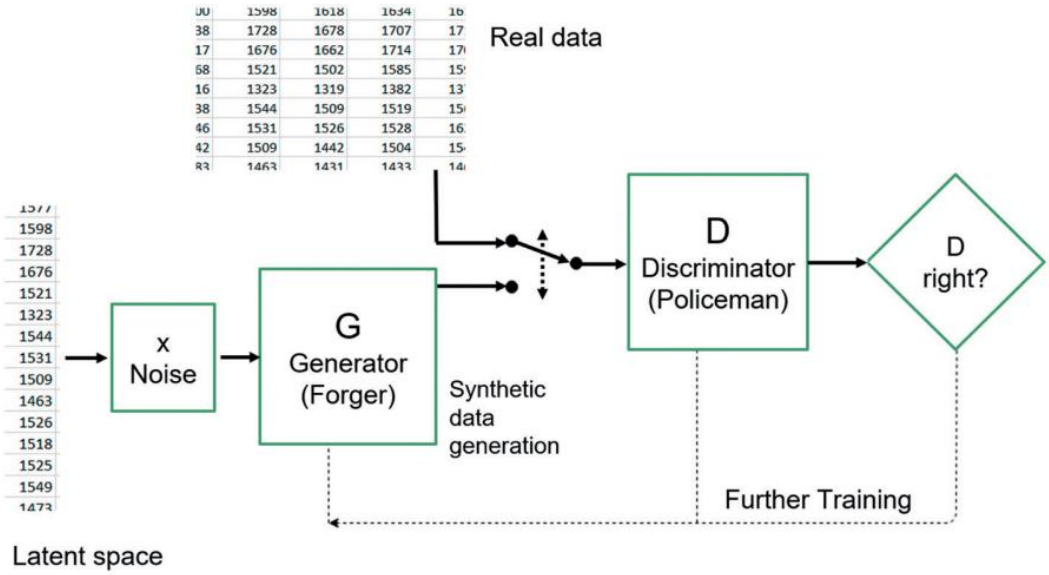

Figure 3 - Illustration of training of Generative Adversarial Network (Kaloskampis, 2020)

Conditional GANs ("CTGAN") (Xu, et al., 2019) are a variation of GANs that incorporate additional conditional input information into the generator and discriminator networks. This additional information allows for more targeted synthetic data generation and better synthetic data quality. CTGAN is an architecture that is specifically designed for tabular data.

PATE-GAN (Jordon, 2019) provides strong privacy guarantees, by utilizing the principles of differential privacy. It works by adding noise during the training process. This noise is generated using the Private Aggregation of Teacher Ensembles ("PATE") framework, which provides differential privacy guarantees.

DoppelGANger (Lin Z. J., 2019) is a generative model that is specifically designed for generating synthetic time-series data. It uses a combination of GANs and Recurrent Neural Networks ("RNNs") to learn the temporal dependencies in the original time-series data and generate synthetic time-series data that resembles the original time-series.

3.2.2 Variational Auto-encoders

Variational auto-encoders are based on the principle of encoding and decoding. The encoder network maps the input data to a latent space, and the decoder network generates synthetic data samples by mapping from the latent space back to the input

space. VAEs use a regularization term to ensure the latent space to follows a specific distribution.

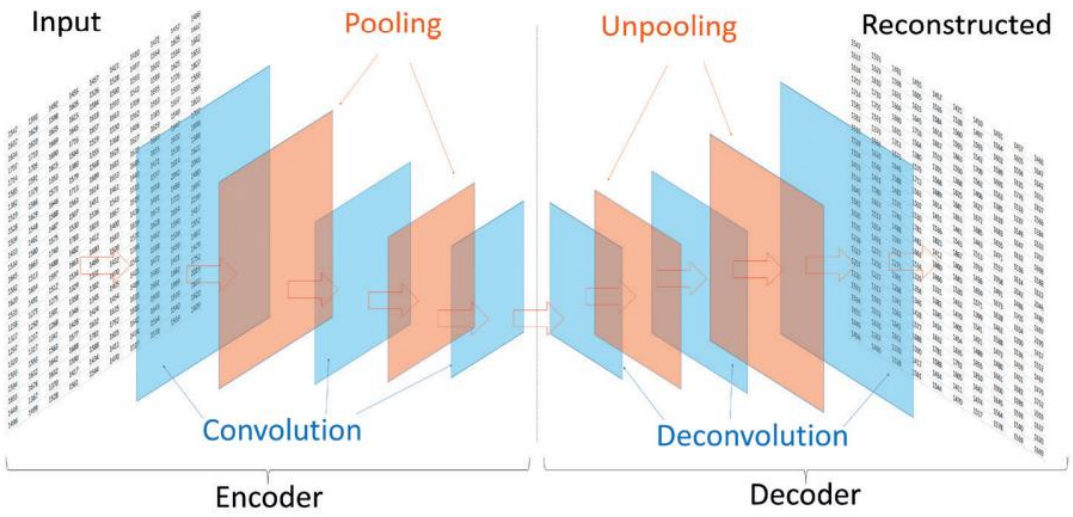

Figure 4 - Illustration showing auto-encoder architecture (Kaloskampis, 2020)

4 Fully Conditional Specification

The fully conditional specification ("FCS") method is inspired by earlier advanced imputation methods (Van Buuren, 2006), as synthetic data generation can be understood as an imputation problem for which all data is missing. FCS uses multiple underlying models to create a fully conditional specification of the data. A popular open-source implementation is Synthpop[4] which is written in R.

For training the underlying models, the multidimensional joint distribution represented by the data is decomposed into a series of conditional and univariate distributions on which each model is trained. The training proceeds by modelling and generating one variable at a time, with each model conditional on the previous one.

The underlying model used in this process should be chosen carefully, as the appropriate model to use depends on the datatype of the targeted variable. Classification and Regression Trees ("CART") (Breiman, 1984) are advantageous because they do not require extensive parameter tuning, and because they handle irregular distributions and non-linear relationships between variables well (Reiter, 2005) as compared to, e.g., parametric models.

In practice, FCS can be challenging to scale and is hard to work with on larger datasets. Moreover, FCS can be prone to overfitting, especially if CART models are used, which could pose a privacy risk.

[4] https://github.com/cran/synthpop

5 Privacy evaluation

The goal of privacy-preserving synthetic data generation is to generate data that is statistically similar to the original data, but does not disclose sensitive information about individuals. To evaluate the privacy of synthetic data, researchers often simulate an attacker who tries to re-identify individuals in the original data using information from the synthetic data. In this chapter the key industry-standard distance-based privacy evaluation metrics are discussed (Platzer, 2021).

5.1 Identical match ratio

The Identical Match Ratio (IMR) is a privacy evaluation metric that measures the proportion of records in the synthetic data that exactly match a record in the original data. The IMR can be used to evaluate the risk of record linkage attacks, where an attacker tries to link records in the synthetic data to corresponding records in the original data. Record linkage attacks are a common privacy threat, and protecting against them is an important goal in synthetic data generation.

To compute the IMR, first pairs of records in the synthetic and original data that correspond to the same individual are identified. This can be done by comparing attributes such as postal code and birthdate. Once these pairs are identified, the proportion of records in the synthetic data that exactly match a record in the original data are computed.

A high IMR value can, but does not necessarily indicate, that the synthetic data is more susceptible to record linkage attacks. However, it can also be the result of an original dataset having many duplicate records, which are reflected in the synthetic data. To infer whether the IMR values pose a privacy risk, or not, it is recommended to use a holdout dataset (see Section 5.3).

5.2 Distance to Closest Record

The Distance to Closest Record (DCR) is another common evaluation metric used in synthetic data generation. It measures the distance between each record in the synthetic data and its closest neighbor in the original data. The DCR measures the ability of an adversary to re-identify individuals in the original data using information from the synthetic data.

To compute the DCR, the first step is to calculate the distance between each record in the synthetic data and its closest neighbor in the original data. A commonly used distance measures is the Gower distance, as this distance measure can deal the mixed datatypes, both continuous and categorical.

A low DCR value, in the privacy evaluation context, indicates synthetic records are closer to the original data, but this does not necessarily indicate individual-level information has leaked in the synthetic data, nor whether there is a high risk of re-identification. Instead, to infer whether the synthetic data is adequately protecting the privacy of individuals in the original data, it is recommended to use a holdout dataset (see Section 5.3).

5.3 Using a holdout dataset with distance-based metrics

Using a holdout dataset in combination with the IMR and DCR can provide a more comprehensive evaluation of the quality and privacy of synthetic data. A holdout dataset is a portion of the original dataset that is not used during the training of the synthetic data generation model. This dataset can be used as a validation set to evaluate model overfitting risk on data the model has not been trained on.

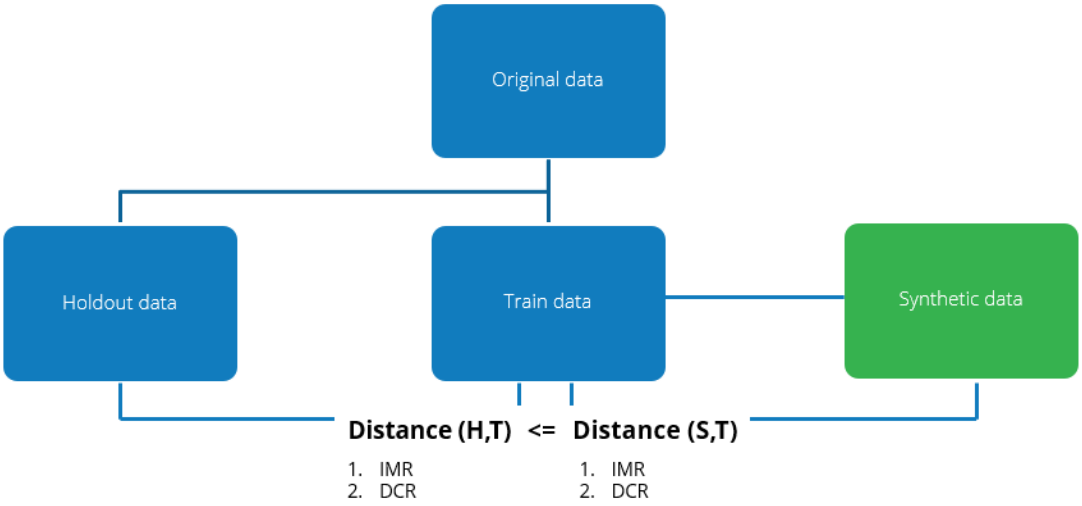

Figure 5 - Illustration showing holdout data split

Now the IMR and DCR are not only computed between the synthetic data and the train data but also between the holdout dataset and the train data. This allows us to assess the potential risk for overfitting.

If the IMR and DCR values between the synthetic data and the train dataset are similar to the values between the holdout data and the train data, it suggests that the synthetic data generation model is producing high-quality synthetic data that is representative of the underlying data distribution. On the other hand, if the IMR and DCR values between the synthetic data and the train dataset are significantly lower from the values between the holdout data and the train data, it suggests that the model may be overfitting to the train data, which poses a potential privacy risk.

Overall, using a holdout dataset in combination with the IMR and DCR provides a more robust evaluation of synthetic data quality and privacy, and helps ensure that the synthetic data is representative of the underlying data distribution and provides adequate privacy protection.

6 References

Breiman, L. (1984). Classification and regression trees. *Routledge*.

Goodfellow, I. P.-A.-F. (2014). Generative adversarial networks. *Communications of the ACM*.

Jordon, J. Y. (219). PATE-GAN: Generating synthetic data with differential privacy guarantees. *International conference on learning representations*.

Kaloskampis, I. J. (2020). Synthetic data in the civil service. *Significance*.

Lin, Z. J. (2019). Generating high-fidelity, synthetic time series datasets with doppelganger. *arXiv*.

Machanavajjhala, A. K. (2007). l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*.

Narayanan, A. &. (2006). How to break anonymity of the netflix prize dataset. *arXiv*.

Nowok, B. R. (2016). Synthpop: Bespoke creation of synthetic data in R. *Journal of statistical software*.

Platzer, M. &. (2021). Holdout-based empirical assessment of mixed-type synthetic data. *Frontiers in big Data*.

Reiter, J. P. (2005). Using CART to generate partially synthetic public use microdata. *Journal of official statistics*.

Sweeney, L. (2000). Simple demographics often identify people uniquely. *Health San Francisco*.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*.

Van Buuren, S. B.-O. (2006). Fully conditional specification in multivariate imputation. *n. Journal of statistical computation and simulation*.

Walonoski, J. K. (2018). Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*.

Xu, L. S.-I. (2019). Modeling tabular data using conditional gan. *Advances in Neural Information Processing System*.