

Making Sensitive Data Open and Fair Through Synthetic Data Generation

A guidebook



This guidebook is the final result of [Project SENSYN](#), which is funded by the NWO Open Science Fund (Grant number OSF23.1.006).

Authors

Katharina Krüselmann 

Institute of Security and Global Affairs, Leiden University

Jim Achterberg 

Health Campus The Hague, Leiden University Medical Center


Dr. Marcel Haas 

Health Campus The Hague, Leiden University Medical Center

Prof.dr. Marco Spruit 

Institute of Advanced Computer Science, Leiden University

Public Health & Primary Care, Leiden University Medical Center

Prof.dr. Marieke Liem 

Institute of Security and Global Affairs, Leiden University

August 2024

DOI: [add](#)



Contents

| | | |
|----------|---|-----------|
| 1 | Introduction: The promise of synthetic data for open science | 5 |
| 1.1 | Hurdles for datasharing | 6 |
| 1.2 | About this guidebook | 8 |
| 2 | Synthetic data: an introduction | 9 |
| 2.1 | Creating synthetic data from scratch | 9 |
| 2.2 | Creating synthetic data based on real-world data | 10 |
| 2.3 | Advantages and disadvantages of synthetic data | 11 |
| 2.4 | Potential uses for synthetic data in scientific research | 13 |
| 2.4.1 | The use of synthetic data for open science practices | 14 |
| 2.4.2 | Synthetic data compared to other privacy-preserving methods | 16 |
| 3 | Synthesizing data - the process | 19 |
| 3.1 | The preparation phase | 19 |
| 3.2 | The data generation phase | 23 |
| 3.3 | The data evaluation phase | 23 |
| 3.4 | Open-source tools for synthetic data generation | 24 |
| 4 | Privacy and utility of synthetic data | 27 |
| 4.1 | Privacy of synthetic data | 27 |
| 4.1.1 | Privacy risks | 27 |
| 4.1.2 | The issue: lack of standards for privacy | 30 |
| 4.1.3 | Increasing privacy through differential privacy | 31 |
| 4.2 | Utility of synthetic data | 31 |

| | | |
|----------|--|-----------|
| 4.3 | The trade-off between privacy and utility | 32 |
| 5 | Proof-of-concept: synthetic Dutch homicide data | 35 |
| 5.1 | Dataset: The Dutch Homicide Monitor | 35 |
| 5.1.1 | Privacy of homicide data in the Dutch Homicide Monitor | 38 |
| 5.2 | Generating a synthetic version of the Dutch Homicide Monitor | 38 |
| 5.2.1 | First attempt | 38 |
| 5.2.2 | Second attempt | 50 |
| 5.2.3 | Final attempt | 56 |
| 5.3 | Concluding Reflections | 62 |
| 5.3.1 | The Process | 62 |
| 5.3.2 | The Data | 62 |
| 5.3.3 | Open & FAIR data | 63 |
| 6 | Conclusion: Enhancing Open Science Practices through Synthetic Data | 65 |
| 7 | Appendix: Resources | 67 |
| 7.1 | Resources on Open Science | 67 |
| 7.2 | Resources on Open Data and Personal Data | 67 |
| 7.3 | Resources on synthetic data | 67 |



Chapter 1

Introduction: The promise of synthetic data for open science

Open Science is a movement to promote and stimulate open and accessible research practices, not just at the stage of publishing research results, but throughout the research cycle as a whole: from preregistration of research ideas and processes, to open analysis protocols, open code, open data, open peer review, open access publications and all steps in between. These practices have the aim to make research more accessible and transparent, to foster collaboration, strengthen trust in research and increase the scientific and societal impact of research. This movement is gaining momentum globally, driven by the growing recognition of the value of open science to research for advancing knowledge and addressing societal challenges [1].

An integral part of every empirical research cycle is data, which, too, is part of open research practices. Open data and FAIR data are central to the ethos of open science. **Open data** refers to data that is freely available for anyone to use, reuse and redistribute, with as little restrictions as possible. **FAIR data**, on the other hand, emphasises that data should be **f**indable, **a**ccessible, **i**nteroperable, and **r**esuable. The concept of FAIR data was introduced to create an infrastructure around the reuse of scholarly data. While open data focuses on the unrestricted availability of data, FAIR data principles provide a framework for ensuring that data are not only accessible, but also well-organised,

annotated, and structured in ways that make them useful and meaningful to researchers across disciplines. Thus, open and FAIR data are not synonyms: data can be open but not follow FAIR standards, whilst FAIR standards do not require data to be open without restrictions. Specifically, FAIR standards recognise that not all data can be made open and accessible to everyone at all times. Therefore, when it comes to open data, the principle "as open as possible, as restricted as necessary" is commonly applied, which contrasts the ideas of open data [2].

Both open data practices and FAIR principles are associated with several benefits to researchers. By making data openly available, researchers can enhance the visibility and impact of their work, leading to increased citations and greater recognition within the scientific community. FAIR data practices improve data quality and facilitate data sharing and collaboration, enabling researchers to build on each other's work more effectively. This collaborative approach can lead to new insights, help to avoid duplication of data collection efforts, and accelerate the generation of new knowledge. Furthermore, open and FAIR data practices support transparency and reproducibility in research, which in return should enhance trust in the scientific process and results. In addition, more and more journals and funding agencies require researchers to share (meta)data or a justification for why data sharing is not possible.

1.1 Hurdles for datasharing

Despite these benefits, some researchers are reluctant to adopt open and FAIR data practices. There may be concerns about the misuse of data, loss of competitive advantage, and the potential for data to be misinterpreted or misrepresented when made openly accessible. Another concern about open and FAIR data sharing may relate to legal and ethical risks: The processing and sharing of personal data is protected under privacy regulations, such as the European General Data Protection Regulations (GDPR). The GDPR defines personal data as "any information relating to an identified or identifiable natural person ('data subject')" [3], which includes a wide range of information, from names and addresses to someone's gender or age, biometric data, and social identity markers. Some researchers may be faced with even more restrictive regulations, for example when working with highly sensitive data, such as crime-, financial- or health-

data. In the face of these restrictions, researchers may opt out of sharing any part of their data.

Yet, the consequences of a lack of data sharing cannot be understated, in particular in those disciplines that commonly work with sensitive and/or personal data. Pridemore, for example, notes a replication crisis in the field of criminology, which is partially brought about by the unwillingness of researchers to share their data and a lack of a culture that incentivises such practices, with potentially far-reaching consequences for crime and justice policies [4]. When it comes to societal impact of research, a lack of transparency may decrease trust in research and evidence-based policies [5, 6].

In light of these and other issues resulting from a lack of sharing (sensitive) data, the question arises whether there are options to share data of personal and sensitive nature, whilst still protecting the privacy of research subjects and adhering to relevant regulations. With fast developments in artificial intelligence and machine learning over the last few years, promising tools arose that offer new approaches of processing, and eventually sharing, data without violating privacy regulations. One of these tools is the concept of synthetic data, which is, in short, artificially manufactured data that mimics real data without referring to actual persons that are protected through GDPR and other regulations [7–9] (more on synthetic data in the next chapter). Although the idea of synthetic data has gained traction amongst statisticians, computer- and data-scientists and has been applied by some commercial and public organisations, it has not yet reached all academic disciplines (specifically in the social sciences and humanities). One main reason for the lack of attention on synthetic data in those fields is that the majority of discourse on synthetic data has focused on the underlying statistical and technical processes of data synthesis, rather than the application of synthetic data to various fields. In addition, existing literature on synthetic data is somewhat inaccessible as it requires advanced technical and statistical knowledge. As a result, researchers without the technical expertise to engage with existing literature may not be aware of options to making their personal data FAIR and open.

1.2 About this guidebook

This guide aims to introduce and stimulate the use of synthetic data as a promising tool to overcome obstacles to the sharing of sensitive data for public and research use, in particular to a non-technical audience. This guide includes a general introduction to the concept, possible applications, a non-technical explanation of the synthesis process, an introduction to open-source tools for data synthesis and an overview of how the privacy and utility of synthetic data needs to be evaluated.

Researchers from [Leiden University](#) and the [Leiden University Medical Centre](#) created this report in the context of Project SENSYN, which is funded through the [NWO Open Science Fund](#). More information on the project and its outputs can be found on the [Open Science Framework project page](#) or the project's webpage [ENTER HYPERLINK](#).

Chapter 2

Synthetic data: an introduction

In short, synthetic data is artificially manufactured data. Synthetic data can encompass a wide range of data types, mirroring the diversity found in real-world datasets. It can be used for both structured data, such as tabular data where rows represent individual records and columns represent attributes, and unstructured data such as text, images, and audio. Examples of synthetic tabular data are health records, customer information, transaction records, and clinical trial data [7, 8, 10]. In terms of unstructured synthetic data common examples are chat logs, emails, social media posts and clinical notes for text [11, 12], and medical scans and deepfakes for images [13, 14].

Two main types of synthetic data can be differentiated: synthetic data that is created from scratch and not based on a particular real-world dataset, and synthetic data that is generated to mimic a (complete or partial) real-world dataset.

2.1 Creating synthetic data from scratch

Synthetic data generated from scratch is entirely artificial and created without using any real-world data as a template. This type of synthetic data is constructed using mathematical models and algorithms that define specific characteristics of the data. For instance, researchers can use simulation techniques to generate datasets based on theoretical distributions and hypothetical scenarios, or their own domain-specific knowledge. This

approach is particularly useful for testing hypotheses, developing algorithms, and conducting simulations where real-world data is either unavailable or unsuitable. Because it is not derived from actual data, synthetic data from scratch poses very little risk of compromising privacy [7, 15].

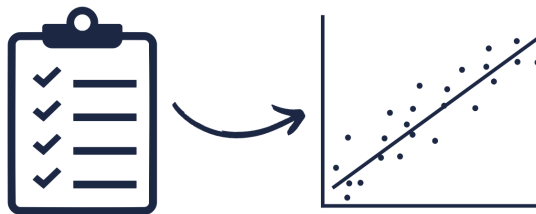


Figure 2.1: Synthetic data from scratch.

2.2 Creating synthetic data based on real-world data

Synthetic data designed to mimic real-world data, on the other hand, is generated using real datasets as a basis. Advanced algorithms, including machine learning and deep learning techniques, are employed to analyse the patterns and relationships within the real data. These patterns are then used to generate synthetic data that closely mirrors the statistical properties and structure of the original dataset. The goal is to produce synthetic data that is realistic enough to be used in place of actual data for analysis, testing, training, or informative purposes while ensuring that no individual's personal information is exposed. This type of synthetic data is particularly valuable in fields like healthcare, finance, criminology or other social sciences, where data privacy is paramount but the need for realistic data is critical for effective research and development [9, 16].

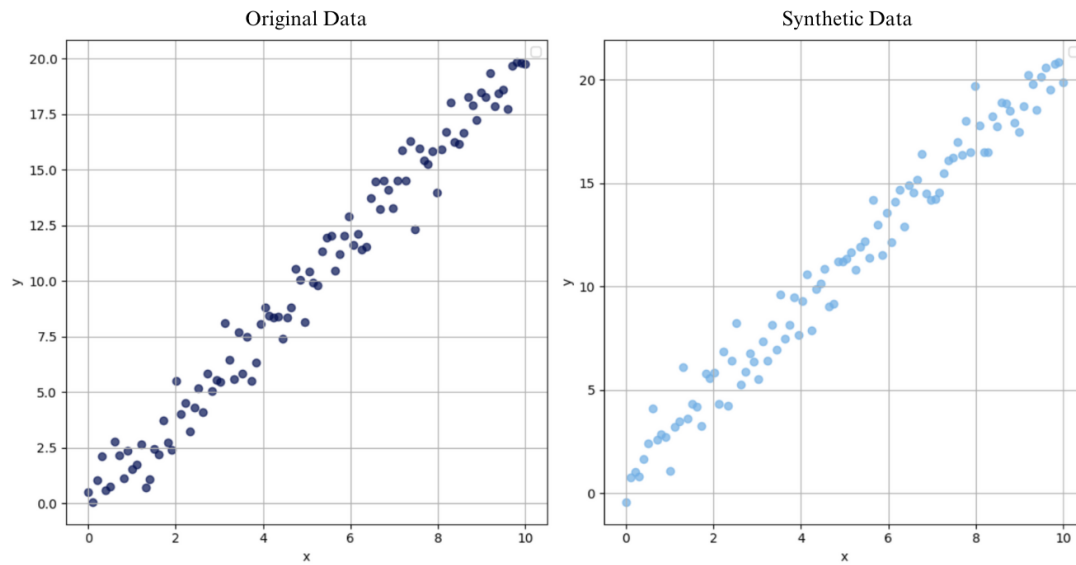


Figure 2.2: Synthetic data mimicking real-world data.

By understanding and utilising these two types of synthetic data, researchers can overcome many of the challenges associated with data privacy and sharing. Synthetic data from scratch provides a risk-free environment for preliminary research and hypothesis testing, while synthetic data that mimics real-world data enable researchers to perform realistic and meaningful analyses without compromising the privacy of individuals. Together, these approaches offer solutions for enhancing data sharing and collaboration in compliance with privacy regulations.

2.3 Advantages and disadvantages of synthetic data

Synthetic data both provides significant benefits, whilst also presenting unique challenges. This section summarises some of the main advantages and disadvantages of synthetic data. Some of these are discussed more in-depth in the following chapters.

Benefit: Privacy preservation

Synthetic data can be generated without compromising sensitive information, making it

an excellent tool for privacy-preserving data analysis. This is particularly relevant for fields that work with sensitive data, such as healthcare data, financial data or crime data, where data privacy regulations like GDPR and HIPAA are stringent. By using synthetic data, organisations can share and analyse data without exposing personally identifiable information and harming someone's right to privacy.

Benefit: Data augmentation

Synthetic data is invaluable for augmenting small datasets. In fields such as machine learning, large datasets are crucial for training robust models. Synthetic data can fill gaps, balance class distributions, and enhance the diversity of training data, leading to improved model performance.

Benefit: Cost-effective

Generating synthetic data can be more cost-effective than collecting and labelling real-world data. Data collection processes can be expensive and time-consuming, often requiring significant human resources. Synthetic data, on the other hand, can be produced at a fraction of the cost and time.

Benefit: Data- and environment manipulation

Synthetic data allows for the creation of controlled environments where variables can be manipulated to study their effects. This is beneficial in scientific research and simulation studies, where researchers can systematically vary parameters and observe outcomes without real-world constraints.

Disadvantage: Quality concerns

The quality of synthetic data depends heavily on the models used to generate it. Poorly designed synthetic data may not accurately reflect real-world patterns and correlations, leading to misleading conclusions and poorly performing machine learning models. So far, there are no standardised frameworks for quality control of synthetic data.

Disadvantage: Potential biases

If the models generating synthetic data are trained on biased real-world data, these biases can be perpetuated or even enhanced in the synthetic data. This can lead to biased outcomes in analyses and machine learning models, undermining the benefits of using

synthetic data.

Disadvantage: Complexity of data generation

Generating high-quality synthetic data can require advanced knowledge in data science, statistics, and domain-specific expertise for the evaluation of synthetic data. Developing and fine-tuning the synthesis process and associated models can be complex and time-consuming, requiring specialised skills and computational resources. Moreover, in many fields, there are no good examples yet of the generation and application of synthetic data.

Disadvantage: Regulatory and ethical concerns

While synthetic data can help navigate privacy regulations, it is not free from regulatory scrutiny. Misuse of synthetic data, especially when it comes to mimicking sensitive data, can raise ethical and legal issues. Furthermore, reliance on synthetic data must be carefully managed to avoid overestimating its reliability and validity, given possible quality control problems.

2.4 Potential uses for synthetic data in scientific research

The potential of synthetic data for scientific research is plentiful, given its ability to replicate real-world data patterns while preserving privacy and circumventing data access limitations.

In **healthcare**, synthetic data plays a crucial role in medical research and clinical trials. Researchers can generate synthetic patient records that mimic real-world medical histories, allowing for the development and testing of new treatments and diagnostic tools without risking patient confidentiality. For example, synthetic data can simulate patient outcomes in clinical trials, providing a safe environment to assess the effectiveness and safety of new drugs before conducting actual trials. [17–19]

In **finance**, synthetic data is used to model financial markets and economic behaviour. Researchers can generate synthetic stock market data, transaction records, and customer profiles to test trading algorithms and risk management strategies without exposing sensitive financial information. This approach not only safeguards privacy but also allows researchers to conduct extensive stress testing and scenario analysis, which is

essential for understanding market dynamics and preparing for potential economic crises. [20, 21]

Environmental science benefits from synthetic data through the simulation of environmental phenomena and the modelling of ecosystems. For instance, researchers can generate synthetic climate data to study the impacts of climate change under various scenarios. This include simulating temperature variations, precipitation patterns, and extreme weather events to predict their effects on ecosystems and human societies. Synthetic data also aids in the development of predictive models for natural disasters, such as floods and hurricanes, enhancing preparedness and response strategies. [22, 23]

In the fields of **criminology**, synthetic data is utilised to analyse crime patterns and develop predictive policing models. By creating synthetic crime reports and incident data, researchers can study trends and correlations without compromising the privacy of individuals involved in real criminal cases. This helps in identifying risk factors, improving law enforcement strategies, and designing more effective crime prevention programs. [24]

Social sciences also leverage synthetic data to study human behaviour and societal trends. Researchers can generate synthetic survey responses, demographic data, and social interactions to explore phenomena such as voting behavior, social mobility, and public opinion. This enables the examination of various hypotheses and policy impacts without the ethical and practical challenges associated with collecting real-world data. [8, 25]

2.4.1 The use of synthetic data for open science practices

Next to these practical applications, synthetic data offers several compelling advantages for open science practices and the FAIR principles.

Privacy preservation

One of the most significant benefits of synthetic data is its ability to preserve privacy. Open science advocates for data sharing to foster transparency and reproducibility, but sharing real-world data often conflicts with privacy concerns, especially in sensitive fields like healthcare, medicine or social sciences. Synthetic data, which mimics real

data without containing personal information, allows researcher to share datasets freely without risking breaches of confidentiality.

Data accessibility

Synthetic data facilitates broader access to data. Researchers can generate and share synthetic datasets that replicate the properties of restricted or proprietary data. This accessibility aligns with the FAIR principle of making data accessible, enabling a wider community of researcher to engage with and analyse the data, thus promoting inclusivity and collaboration.

Enhanced interoperability

By using standardised methods and tools to generate synthetic data, researchers can ensure that datasets are compatible across different platforms and software. This interoperability is crucial for collaborative efforts where multiple teams might be using various systems and tools, allowing for seamless integration and analysis of data from diverse sources.

Reusability

Synthetic data enhances the reusability of datasets. Because it can be shared without legal or ethical restrictions, synthetic data can be reused in multiple studies and by various research teams. This reusability supports the FAIR principle by maximising the utility of datasets, enabling continuous testing and validation of scientific hypotheses, and facilitating cumulative knowledge building.

Cost-effectiveness and efficiency

Generating synthetic data can be more cost-effective and time-efficient than collecting new data. This efficiency allow researchers to quickly produce large, rich datasets necessary for robust scientific inquiry. Additionally, synthetic data can simulate rare or hypothetical scenarios, providing valuable insights that might be difficult or impossible to obtain from real-world data.

Overall, synthetic data significantly advances open science and the FAIR principles by enhancing privacy, accessibility, interoperability, and reusability. It empowers researchers

to share and utilise data more freely and effectively, fostering a more collaborative and transparent scientific environment.

2.4.2 Synthetic data compared to other privacy-preserving methods

The generation of synthetic data is not the first or only mechanism through which researchers can enhance and protect the privacy of their data. Other common mechanisms include anonymisation and differential privacy, yet both come with significant shortcomings.

Anonymisation involves removing or obfuscating personally identifiable information from dataset to prevent the identification of individuals. This method can be as easy as deleting attributes in a dataset that are related to personal information, such as names of individuals. As such, anonymisation is a widely used method. However, anonymisation has significant limitations, particularly in the face of advanced re-identification techniques. Even anonymised datasets can sometimes be linked with other data sources to re-identify individuals, compromising privacy. This limitation is especially pertinent in the era of big data, where vast amounts of auxiliary information are readily available [26]. Additionally, obfuscating information which may be relevant may harm the overall utility of the dataset.

Differential privacy is another privacy-preserving mechanism [27]. Differential privacy is a sophisticated technique that adds controlled noise to datasets, ensuring that the inclusion or exclusion of any single data point does not significantly affect the overall results. For example, noise can be added by altering certain values, such as salaries to obscure the exact, personal information. This method provides strong mathematical guarantees of privacy and is particularly effective in preventing re-identification. However, the added noise can degrade data utility, making it less useful for certain types of analyses. Implementing differential privacy also requires careful calibration to balance privacy and data quality, which can be complex and resource-intensive [28, 29]. Additionally, it should be noted that differential-privacy can also be employed *within* synthetic data generation, as a means to prevent re-identification from a synthetic dataset [30–32].

Each of these privacy-preserving methods comes with its own benefits and challenges. The choice of mechanism depends on the specific requirements and constraints of the research project.

Chapter 3

Synthesizing data - the process

Synthetic data generation is an iterative process that can be divided into three phases: the preparation phase, the data generation phase and the data evaluation phase.

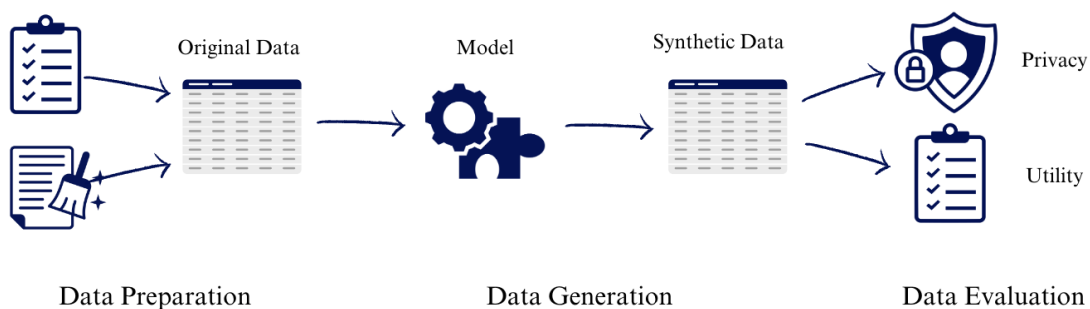


Figure 3.1: The Synthesis Process.

3.1 The preparation phase

The goal of the preparation phase is to determine the ideal characteristics of the synthetic data and to choose a fitting generation method.

Step 1a: When creating synthetic data from scratch: determine the structure and characteristics of the synthetic data

Generating synthetic data from scratch requires meticulous preparation to ensure that the artificial data accurately reflects the intended use case and properties. The preparation phase involves several key steps. When developing synthetic data based on domain knowledge, deep comprehension of the domain is crucial. This includes understanding the key variables, their relationships, and the underlying processes governing the real-world data. For example, generating a synthetic database of a population of prisoners requires knowledge about the characteristics of prisoners. Equally, when the goal is to generate synthetic legal text, knowledge of legal jargon is a requirement. Another key step lies in defining the objectives for generating synthetic data. Will the data be used for testing algorithms, simulating specific scenarios, or conducting replication studies (to name just a few examples)? Depending on the answer to this question, the key features of the synthetic data need to be identified, such as their specific statistical properties: distributions, means, variances, correlations, and other relevant statistical metrics. For textual data, the desired length, complexity, tone or linguistic patterns need to be determined [7, 15].

Step 1b: When generating synthetic data to mimic real-world data: prepare the original dataset

When generating synthetic data based on real-world data, it is still necessary to understand the key features of that data, such as distributions and relationships among variables for structured data, or patterns in language, sentiments, topic and structure for unstructured data, such as texts or audio. This knowledge is later required to evaluate the quality and usefulness of the synthetic dataset. In addition, data cleaning and preprocessing are crucial at this stage to ensure the quality and consistency of the real-world data, which serves as a benchmark for the synthetic data. Any mistakes, biases, outliers, special characteristics or irrelevant information need to be resolved or at least identified before the synthesis, as the synthesis process would mimic or even enhance those mistakes. Finally, any additional requirements for the synthetic data, such as logical rules, need to be determined. For example, if the original dataset contains a variable classifying individuals as either adult or minor, as well as a variable on the specific age of the individual, the model may recognise that individuals classified as minor

have a significantly lower age than those classified as adults, but it may be necessary to create a rule that the specific age of minors in the synthetic dataset must not exceed 17.

Step 2: Choosing the right generation method

There are several methods to generate structured and unstructured synthetic data, and the choice depends on the type and complexity of the data, as well as the specific requirements of the project. Some of these methods may require advanced statistical knowledge, whilst others are more accessible to researchers without advanced expertise in statistics or data science. Some general categories of these methods are presented here.

Rule based methods for synthetic data generation rely on predefined rules, algorithms or simulations that mimic the behaviour and interactions found in real-world systems. These methods are particularly effective when no prior data is available, but domain knowledge is well-understood and can be explicitly encoded into the data generation process. In some disciplines, there may already be mathematical models that can be used for simulations, such as epidemiological models on the spread of diseases that can be used as a guidance for the synthetic data [17, 33]. Rule-based models are also suitable for demographic simulations, if demographic rules and census data offer enough information to create an artificial population. On the one hand, rule-based models can be highly tailored to specific domains, ensuring that synthetic data closely mimics real-world scenarios. Another benefit is that the rules through which the synthetic data is created need to be specified, which makes the data generation process transparent and understandable. On the other hand, it can be complex and time-consuming to create rule sets that accurately capture dynamic real-world behaviours. Additionally, synthetic data from rule-based methods can only include patterns and relationships which are known beforehand, thereby possibly excluding unknown but vital information.

Statistical models use mathematical formulas to replicate the characteristics and relationships found in real data. For example, through random sampling, a model can generate synthetic data points from probability distributions that closely match the distributions observed in the real data. Thus, if the original data follows a normal distribution, synthetic data can be generated by sampling from a normal distribution with a

similar mean and standard deviation as the original data. Other statistical models, such as copulas, may be more suited if one aims to model the dependency structure between multiple variables [34]. The vast variety of statistical models allows for flexible adaptation to different types of data and purposes for generating synthetic data. However, depending on the complexity of the original data, sophisticated statistical techniques and skills may be required to generate useful synthetic data. In addition, synthetic data generated through statistical models require a thorough validation to ensure that it accurately represent the real-world data [35]. Also, statistical models often rely on strong assumptions regarding which distribution the data follows. These assumptions may not always hold true in practice.

Machine/Deep learning models are used for generating more complex and realistic synthetic data. One example of a machine learning model are Generative Adversarial Networks (GANs). GANs involve two neural networks; one that generates synthetic data and one that evaluates it. They "compete" against each other, thereby improving the quality of the synthetic data over time [36]. Another example are Variational Autoencoders (VAEs), that learn to generate data by encoding the original data to grasp its key features and decoding these features into synthetic data points [37, 38]. Bayesian networks require a map about the relationships between variables in the original data beforehand (e.g. based on expert knowledge), based on which they can learn about the probability that certain combination of features in the original data co-occur. This knowledge is then transferred into a synthetic dataset. These methods are powerful for creating high-fidelity synthetic data, but they require substantial computational resources and expertise and are thus not beginner-friendly. For textual data, language models such as GPT-3 are designed to generate and understand human language, such as grammar, syntax and semantics and replicate next text by predicting the next word or sequence of words [39]. The benefit of machine learning models is their flexibility to learn and replicate complex patterns in data without heavily relying on assumptions, next to their ability to quickly scale to large datasets. However, compared to statistical and rule-based models, machine learning models are much more likely to replicate sensitive information in the synthetic data, due to over-fitting. In other words, these models may replicate the original data *too* well. In addition, training these models can require (large amounts of) training data that the researchers needs to have available in addition to the data that

needs to be replicated. Also, training and fine-tuning these machine models can require more advanced knowledge of computations.

3.2 The data generation phase

Step 3: Synthetic data generation

The actual synthesis process is heavily dependent on the chosen model and the type of data that need to be synthesised. In general though, each of these models studies the real data or rules through which new synthetic data needs to be developed to determine the various patterns, relationships, and statistical properties within the original data or rules. Once the model has satisfactorily learned the properties of the real data, it can be utilized to generate new data points. This new data is not copied directly from the real data but is generated based on the patterns the model has learned, or rules that have been fed to the model.

3.3 The data evaluation phase

Step 4: Evaluation of synthetic data

Evaluation is a critical step in the process of generating synthetic data. When synthetic data is created from scratch, theory or domain knowledge, it needs to be evaluated against the rules and requirements set out at the beginning of the process. When synthetic data is generated to mimic real-world data, it needs to be evaluated on its similarity to the original data. When synthetic data is generated to ensure privacy, it needs to be evaluated and checked using controls for privacy. More detailed explanation on the evaluation of synthetic data is described in the following chapter. When the outcome of these evaluations does not align with the requirements for the synthetic data, the process of synthesising data needs to be repeated and adapted where necessary. As such, the synthesis of data becomes an iterative process.

Optional: Further data anonymisation

To ensure privacy, further steps may be taken to verify that the synthetic data does not contain any identifiable information. Even though synthetic data is inherently safeguarding personal information by generating artificial datapoints that do not relate

to real-world persons, additional precautions can be implemented to safeguard against potential privacy risks.

3.4 Open-source tools for synthetic data generation

Synthetic data is not a new invention. Therefore, several open-source tools already exist that make the generation of synthetic data easier, due to their accessibility, flexibility, and the supportive communities that often accompany them. A few note-worthy open-source options are presented here, with a more comprehensive list in the appendix.

Synthetic Data Vault

The Synthetic Data Vault (SDV) is a comprehensive library that offers a variety of methods for generating synthetic data [40]. It supports generating tabular, sequential, and relational data. SDV utilises machine learning models to capture the distributions and relationships in the original data, ensuring realistic synthetic data generation. It is particularly useful for data scientists who need a robust, all-encompassing tool for different types of data. Next to synthetic data generation, it also incorporates various metrics for evaluating synthetic data quality. **Ease of use:** Moderate. SDV offers comprehensive documentation and examples, but understanding its full capabilities may require more extensive data science knowledge. **Programming language:** Python.

SynthCity

SynthCity is a comprehensive library for synthetic data generation and evaluation [41]. It includes a wide variety of models for synthetic data generation, ranging from more classical machine learning algorithms, to Bayesian Networks, to advanced algorithms like GANs and VAEs. It includes methods for both tabular and imaging data generation. Furthermore, it contains a wide variety of metrics for comprehensive evaluation of synthetic data quality. **Ease of use:** Moderate. As in the SDV, SynthCity offers direct plug-and-play capabilities, but understanding its full potential may require more extensive data science knowledge. **Programming language:** Python.

YData

YData Synthetic is an open-source tool that provides a user-friendly interface for generating synthetic data [42]. It leverage GANs to create realistic data, focusing on enhancing

the quality and diversity of synthetic datasets. **Ease of use:** YData Synthetic is designed to be easy to use, with straightforward interfaces and tutorials that make it accessible even for those with limited data science expertise. **Programming Language:** Python.

Faker

Faker is a lightweight library that generates fake data for various uses, including testing and development [43]. It can produce names, addresses, text, and other common data types. While not as advanced as other tools for complex datasets, Faker is highly useful for quickly generating simple, structured synthetic data. **Ease of use:** High. Faker is extremely easy to use, with simple commands to generate various types of fake data quickly. **Programming Language:** Python.

Synthpop

Synthpop is an R package focused on generating synthetic versions of data sets for statistical disclosure control [44]. It is particularly popular in social science research, where it helps in creating synthetic datasets that preserve the statistical properties of the original data while protecting privacy. It relies on decision trees to sequentially populate synthetic features, to ultimately generate a synthetic dataset. **Ease of use:** Moderate. Synthpop provides extensive documentation, but some familiarity with statistical concepts and R programming is necessary. **Programming Language:** R (Python wrapper also available).

Chapter 4

Privacy and utility of synthetic data

4.1 Privacy of synthetic data

Ensuring the privacy of synthetic data is crucial for protecting individuals' sensitive information whilst allowing for useful data analysis.

4.1.1 Privacy risks

In general, we can distinguish between three different privacy risks:

Re-identification risk

Re-identification risk refers to the probability that an individual's identity can be discovered by matching anonymised or synthetic data back to the original dataset [45, 46]. This risk is a major concern in data privacy, particularly when sharing datasets containing sensitive information. Even if direct identifiers such as names and addresses are removed, individuals can often be re-identified through quasi-identifiers - attributes like age, gender, and zip code - when combined with other accessible datasources.

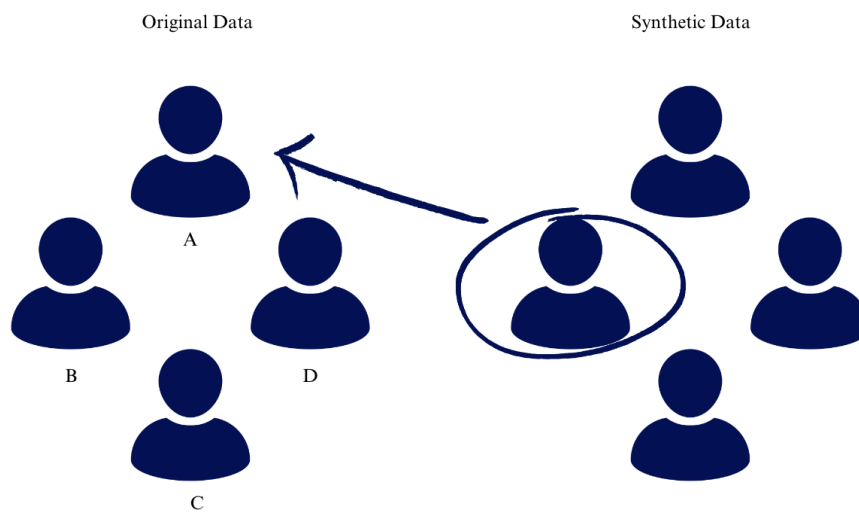


Figure 4.1: Re-identification risk

Membership inference risk

Membership inference risk refers to the probability that an attacker can determine whether a specific individual's data was included in a dataset used to train a machine learning model or generate synthetic data [47]. This risk is particularly concerning because it can reveal sensitive information about individuals, even if the data itself is anonymised. Understanding and mitigating this risk is essential for ensuring privacy in data sharing and open science practices.

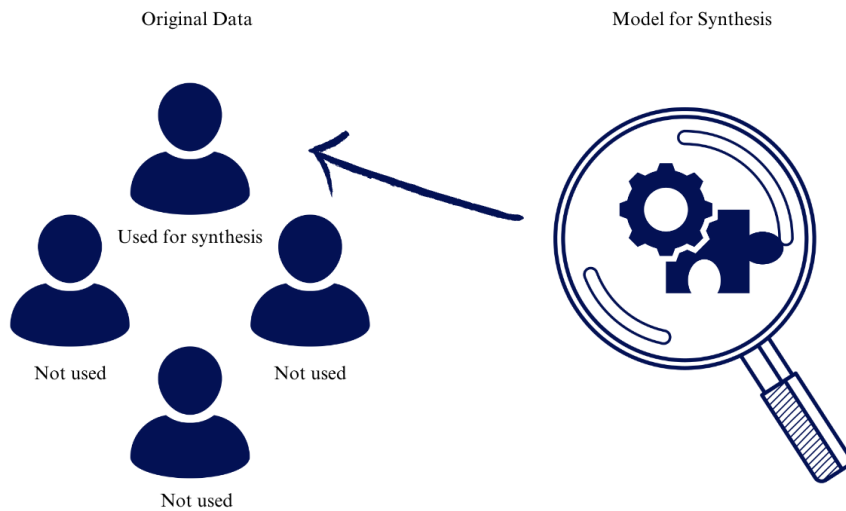


Figure 4.2: Membership inference risk

Attribute disclosure risk

Attribute disclosure risk pertains to the likelihood that sensitive or private information about individuals can be inferred from a dataset, even if direct identifiers are removed. This risk arises when the remaining attributes in the dataset allow an observer to deduce specific details about individuals [48]. For instance, if a dataset includes a combination of age, zip code, and occupation, it might still reveal sensitive information about individuals, such as their income or health status. As such, attribute disclosure risk is similar to the re-identification risk. Yet, whereas the re-identification risk focuses on identifying individuals from data, attribute disclosure risk focuses on revealing sensitive attributes in particular.

Testing for Privacy Risks For each of these risks, various methods exist on how to measure them. These methods range from manually comparing records from the original dataset with records from the synthetic dataset based on quasi-identifiers, or by checking whether outliers in the synthetic dataset can be linked to outliers in the original dataset. Then there are various statistical tests that compare the similarities and differences between both the synthetic and original dataset, for example by calculating statistical distance measures. Examples are Euclidean distance, which connects each

record in the synthetic dataset with the most similar record of the original dataset [49], or Bayesian estimation methods [50]. Another common method for measuring these risks is to simulate attacks on the privacy of the synthetic dataset. In these scenarios, models are trained to act as adversaries that aim to gain sensitive information from the synthetic dataset. These models, controlled by the researchers, can help to identify weak points in the synthetic data and quantify potential privacy risks [48, 51, 52].

4.1.2 The issue: lack of standards for privacy

Although there are a multitude of methods and techniques to evaluate the privacy of a generated synthetic dataset, there are currently several challenges that need to be addressed.

Individualised data generation processes

Each generation of synthetic data involves unique processes tailored to specific use cases, datasets, and goals. Consequently, the techniques and methods used for calculating privacy risks can vary widely. This variability means that there is no one-size-fits-all approach for assessing privacy, and each synthetic data generation scenario may require a distinct set of privacy evaluation methods. This individualised approach complicates efforts to compare and validate privacy measures across different studies and applications.

Absence of standardised risk thresholds

Possibly due to the complexity and variety of synthetic data generation methods and use cases, there are currently no universally accepted standards for defining what constitutes *high* risk or *low* risk in the context of privacy measures. Different techniques for evaluating privacy, such as re-identification risk or membership inference risk, may use varying thresholds to categorise risk levels. This lack of standardised thresholds creates inconsistencies in privacy assessments, making it challenging to interpret results and apply them consistently across different research contexts.

Evolving nature of synthetic data generation

The field of synthetic data generation is rapidly advancing, with new methods and techniques continually emerging. As a result, established privacy evaluation methods

may become outdated or less effective over time. This constant evolution complicates the development and maintenance of standard practices for privacy assessments, as researchers must continually adapt their approaches to keep pace with technological advancements and emerging privacy risks.

4.1.3 Increasing privacy through differential privacy

Differential privacy is a mathematical framework used to provide privacy guarantees when releasing information about a dataset. It ensures that the output of a synthetic dataset does not significantly change when any single individual's data is added or removed, thus protecting individual privacy. In this framework, privacy is measured through two parameters: ϵ (epsilon) and δ (delta). ϵ measures the privacy loss, whereas δ represents the probability of the privacy guarantee being violated. In practice, differential privacy can be applied by determining the desired levels of privacy based on the sensitivity of the data and the acceptable risk and applying the associated values for both ϵ and δ . Depending on the chosen values for ϵ and δ , noise is added to the synthetic dataset, meaning that random modifications are made to the data to obscure the contributions of individual data points. While differential privacy is a robust technique for protecting individual privacy in datasets, it has notable drawbacks, in particular in relation to the framework's complexity. Implementing differential privacy requires careful tuning of privacy parameters, which can be complex and context-dependent. This complexity can make it difficult for researchers without extensive expertise in privacy-preserving techniques to apply differential privacy effectively [27]

4.2 Utility of synthetic data

Specifically in the context of open science, the utility of synthetic data is paramount, as it must not only preserve privacy but also effectively support research objectives. Utility measures are used to evaluate how well synthetic data replicates the original data's characteristics and supports scientific analysis. These measures ensure that while data privacy is maintained, the synthetic data remains valuable for the intended use cases. What constitutes *valuable* may differ per use case and goal for the synthesis. The following measures are some of the most common utility measures:

Statistical accuracy

One of the primary utility measures is statistical accuracy, which assesses how well synthetic data reflects the statistical properties of the original dataset. This is often also called fidelity. Specifically, this means that synthetic datasets should accurately capture the distribution and relationships of key variables. Techniques such as the Kolmogorov-Smirnov test or Chi-Square, or simple visualisations, can be used to test and compare distributions between synthetic and real datasets, ensuring that synthetic data replicates these characteristics accurately. Correlation coefficients, like Pearson's or Spearman's rank correlation, help verify that relationships between variables are preserved. Accurate statistical representation is crucial for ensuring that synthetic data can be used effectively in a wide variety of scientific research where data integrity is vital [16, 53, 54].

Utility in data analysis and decision-making

Synthetic data should support meaningful analysis and visualisation. For example, replication studies with synthetic data of studies using the original data should generate the same general results [19]. Researchers can also apply clustering or classification models to synthetic data and compare the results with those obtained from real data. Consistent findings suggest that synthetic data is useful for generating insights and conducting robust scientific analysis. Furthermore, synthetic data should lead to valid and actionable conclusions, reinforcing its utility in practical application.

User experience and Interpretability

Finally, the ease of use and interpretability of synthetic data are essential utility measures. For open science, synthetic data must integrate smoothly into existing research workflows and tools. It should be formatted in a way that is compatible with FAIR standards.

4.3 The trade-off between privacy and utility

Specifically in the context of open science, researchers may generate synthetic data with the goal to create a dataset that both preserves the privacy of the original data, whilst also creating a synthetic dataset that strongly mimics the original data, to enhance utility for statistical analysis, informational purposes or other use cases. However, researchers have to take into account a potential trade-off between utility and privacy.

The core of the trade-off lies in the fact that increasing utility often comes at the cost of decreased privacy protection, and vice versa. Specifically, such a trade-off could appear if the model through which the synthetic data is generated is *overfitting*, thus mimicking the original data so well that - by chance - one or more records in the synthetic dataset have the exact same characteristics as ones in the original data. Conversely, as privacy safeguards become more stringent, for example by adding noise to the synthetic dataset, the synthetic data may become less representative of the original data, leading to potential inaccuracies in research findings. Finding an optimal balance requires careful consideration of both privacy needs and objectives for the synthetic data. The acceptable level of privacy and utility may vary per use case and context. Ultimately, the goal is to achieve a synthesis of data that adequately protects privacy while still providing valuable insights and supporting robust scientific research. This balancing act is essential for advancing open science while ensuring that individual privacy remains protected.

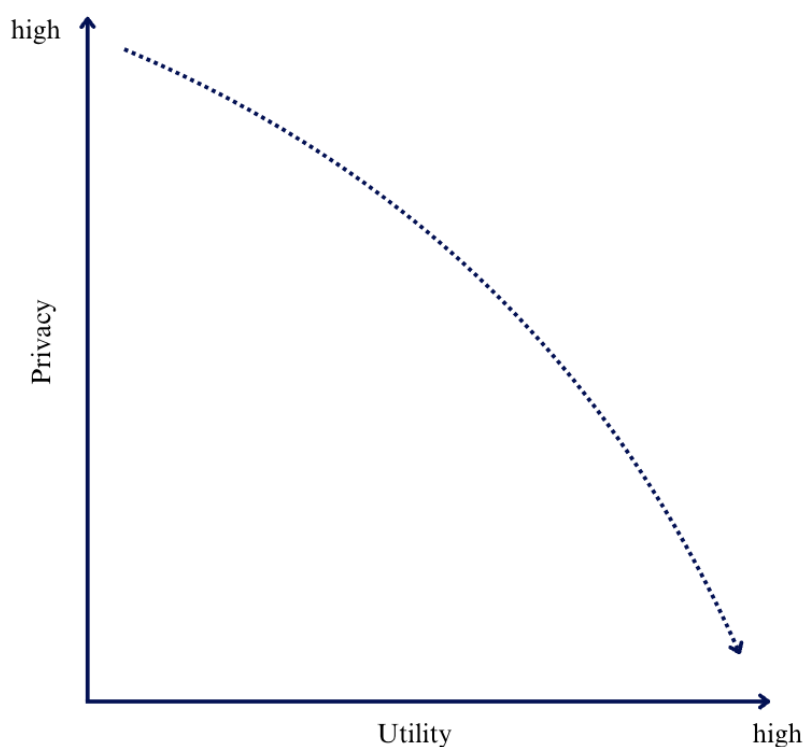


Figure 4.3: Privacy-utility trade-off.

Chapter 5

Proof-of-concept: synthetic Dutch homicide data

In order to showcase the applied use, advantages and limitations of synthetic data, this chapter will elaborate on the generation of a synthetic dataset based on the Dutch Homicide Monitor. This proof-of-concept covers the entire process of synthetic data generation (as described in Chapter 4), from data preprocessing and model selection to the generation and evaluation of synthetic data, using step-by-step discussions on the choices made during that process.

5.1 Dataset: The Dutch Homicide Monitor

The Dutch Homicide Monitor is a dataset administered by researchers at Leiden University in the Netherlands (authors of this report). Through 25 nuclear attributes, the dataset captures detailed information on homicides committed in the Netherlands, including case-level information (such as the time of day and type of crime scene) as well as individual-level data of victims and (suspected) perpetrators, such as the age, gender, occupation and country of birth. The information included in the Dutch Homicide Monitor is derived from six sources: public sources, such as the annual homicide list collected by Elsevier Magazine, news articles and public court ruling, as well as non-

public sources, such as police data, information from the public prosecution service and forensic records. As of June 2024, the Dutch Homicide Monitor captures all homicide committed between 1992 and 2023, which covers 5563 cases, and information on 5920 victims and 7542 (suspected) perpetrators.

Structure of the data

The Dutch Homicide Monitor consists of 25 nucleus variables. Most of these variables (21) are categorical; only four variables (number of victims, number of perpetrators, age, date, description of incident) are non-categorical. Categories within a variable range from two (male/female) to 35 categories (victim-perpetrator relationship). In addition to these 25 variables, the dataset has three variables that act as identifying variables: a case number (which represents a case), a serial number (which represents each individual) and a variable that identifies one principal victim and perpetrator (if known) per case. These variables are defined in a [coding manual](#), which is shared with other researchers in the European Homicide Monitor network. Furthermore, in order to allow cross-matching between multiple data sources, the address at which the homicide took place, the name and the social security number of the victim and perpetrator are included as additional variables.

| Case-level | Individual-level |
|-----------------------------|----------------------------------|
| Case number | Serial number |
| Description | Crimescene |
| Number of victims | Modus operandi |
| Number of perpetrators | Type: victim or perpetrator? |
| Spatial region | Gender |
| City | Age |
| Year homicide was committed | Profession |
| Date of homicide | Country of birth |
| Time of day | Alcohol (ab)use |
| | Drug (ab)use |
| | Violent history |
| | Context of homicide |
| | Victim-perpetrator relationship |
| | Main motive of perpetrator |
| | Suicide (attempt) by perpetrator |
| | Stage in judicial system |

Table 5.1: Variables included in Dutch Homicide Monitor

The variables are embedded in the columns. Each record in that dataset represents an individual - either a victim or (suspected) perpetrator of the homicide. A variable 'Type' classifies the individual as either a victim or perpetrator. One homicide case can involve multiple victims and/or perpetrators, meaning that one case can span from two to six or even ten rows. Information stored on case-level variables is the same across the rows that belong to the same case, whereas individual-level information can vary for each row.

| Case number | Serial number | Time of day | Type | Gender | Age | .. |
|-------------|---------------|-------------|-------------|--------|-----|----|
| 1 | 1.1 | 6am-12pm | Victim | F | 22 | .. |
| 1 | 1.2 | 6am-12pm | Perpetrator | M | 24 | .. |
| 2 | 2.1 | 6pm-12am | Victim | M | 46 | .. |
| 2 | 2.2 | 6pm-12am | Perpetrator | M | 55 | .. |
| 2 | 2.3 | 6pm-12am | Perpetrator | M | 43 | .. |

Table 5.2: Example of DHM structure

5.1.1 Privacy of homicide data in the Dutch Homicide Monitor

The Dutch Homicide Monitor contains sensitive personal information on victims and (suspected) perpetrators of homicide in the Netherlands. Specifically sensitive information includes the gender, age, country of birth, occupation, address of the incident and more. Although some of this information can be found in public sources, such as news articles, overall, this information is protected through the European Union's GDPR regulation, as well as agreements made with the Dutch National Police and Public Prosecution Office of the Netherlands on the handling of their data. The sensitive nature of the data limits the use of the Dutch Homicide Monitor: students, journalists or other researchers who want to use the disaggregated dataset need to apply for access to the data which requires approval by the Dutch National Police. This process can take up considerable time and efforts, making it infeasible for certain use case. In addition, implementing FAIR standards to the dataset is limited, given that the dataset is neither accessible, interoperable, nor reusable.

5.2 Generating a synthetic version of the Dutch Homicide Monitor

5.2.1 First attempt

To begin with, we set a list of requirements for the to-be-generated synthetic dataset. We want a:

- a) fully synthetic dataset
- b) dataset with the same structure as the original data that allows for analysis on case-, victim- and perpetrator-level
- c) statistical accuracy that allows for replication of previous studies using the original data
- d) safeguarding of sensitive information to adhere to GDPR and additional regulations associated with the original data

Through these requirements, we aimed to generate a synthetic dataset that is complete and can be used for several purposes, such as student exercises, as a source of information for journalists and other stakeholders, as well as a source for future academic studies. In short, a dataset that can be registered and used for any possible future purpose.

Step 1: Preparations

Given that the requirements do not ask for much changes to the original data, the preparations were minimal. As with any data analysis, we checked the data for internal inconsistencies, for accidental mistakes in the coding (e.g. 998 instead of 999), duplicates and so on. In addition, we excluded three identifying variables: the name and social security number of victims and (suspected) perpetrators and the address of the homicide. Although these variables could be synthesised, they would carry no meaning in the synthesised dataset, but may increase the risk for privacy breaches. The identifying variables (case number, serial number and the variables that differentiate victims from perpetrators) have been kept in the to-be-synthesised dataset.

Another concern during this phase was data quality. The fraction of missing data for certain cases can be high, e.g. due to a lot of missing information in the homicide, such as the identity of the perpetrator, or due to a lack of details in the available data sources. For example, police reports on earlier homicides may not have as much detail as police reports on more recent cases. In order to have a high-quality synthetic dataset of the Dutch Homicide Monitor, we decided to create a dataset that only contains ten years of homicide data. The chosen ten years of data have the highest data quality in the Dutch Homicide Monitor. After extracting these ten years of high-quality data from the Dutch

Homicide Monitor, the new and to-be-synthesised dataset contained 1271 cases, and information on 1348 victims and 1711 (suspected) perpetrators.

Finally, we determined two logical dependencies in our dataset that are present in the original data and would need to be replicated in a synthetic dataset. First, we determined that any homicide victim involved in a *child homicide* needed to be *younger than 17 years*. Another logical dependency in the original data is between the type of homicide and victim-perpetrator relationship. Specifically, we determined that homicides categorized as intimate partner homicides needed to co-occur with the victim-perpetrator relationship of *partner* or *ex-partner*.

Step 2: Choosing generation method and tool

The choice of generation method and tool was determined by two factors: First, the generation method needed to fit the structure of our data. Second, the generation method and tool specifically needed to be easy to use, as the goal of the project and this guide on synthetic data was to make synthetic data accessible to a broader audience without extensive knowledge of computations or data science. Therefore, we evaluated the extent of helpful and guiding documentation for each generation method and tool and selected a member of the project team with no experience with synthetic data as the lead for the generation process. Following the principals of open science, only open-source tools (thus no commercial options) were included in the decision process.

As mentioned in Section 3.1, three main categories of synthetic data generation methods can be distinguished: rule-based, statistical, and machine learning methods. We opted for a statistical method as rule-based methods are time-consuming to develop and rely heavily on prior domain knowledge, and our dataset is relatively small to justify using advanced machine learning methods like GANs and VAEs. An additional benefit of statistical methods over machine learning methods are increased interpretability of the generation process. To generate high-utility synthetic data, the statistical method should not only capture univariate distributions, but also dependencies between variables. An apt statistical method are thus copula models, which separately model univariate distributions and the dependency structure between variables, after which they are linked together to model the multivariate dataset [55].

Several tools use copulas as a basis for the synthesis of data. Yet, the structure of our data limited the options of tools available to us, given that not all tools are equipped to synthesise data with dependencies across rows and columns. After reviewing several options, we decided that the best approach to keep these dependencies would be through a multi-table synthesis. In this process, several tables linked through a common id are synthesised whilst keeping the relationships between the tables and variables. In order to prepare our data for this process, we divided the main dataset into three separate ones: one that includes four case-level variables, one that includes thirteen variables containing information on the victims and one that includes twelve variables related to information on the perpetrator. Some variables in the victim- and perpetrator-tables overlap, such as age, gender, country of birth or occupation.

| Case variables | Victim variables | Perpetrator variables |
|-----------------------------|---------------------------------|----------------------------------|
| Case number | Serial number | Serial number |
| Description | Gender | Gender |
| Number of victims | Age | Age |
| Number of perpetrators | Profession | Profession |
| Spatial region | Country of birth | Country of birth |
| City | Alcohol (ab)use | Alcohol (ab)use |
| Year homicide was committed | Drug (ab)use | Drug (ab)use |
| Date of homicide | Violent history | Violent history |
| Time of day | Crimescene | Main motive |
| | Modus operandi | Suicide (attempt) by perpetrator |
| | Context of homicide | Stage in judicial system |
| | Victim-perpetrator relationship | |

Table 5.3: DHM variables split by related case-, victim- or perpetrator-level

The Synthetic Data Vault (SDV) is one of the very few open-source tools that allow for multi-table synthesis. In addition, it provides extensive documentation and support, which is why we chose the SDV for the synthesis of our homicide dataset. The SDV runs on Python and includes suggestions and requirements for data preparation for the synthesis of multi table data. Our data fulfilled most of the requirements, with the exception of a document/object that defines the meta data across the tables - in other words a document that provides information about the type of data, the id-variables that

link the tables together and specific connections between the tables. This information can be set-up manually, following instructions in the SDV-documentation, but the SDV-package also comes with functions that can automatically detect this information from your tables, but these require additional verification and checks.

After following the required steps, our homicide dataset was split into one parent table (in our case, the table with case-level variables) and two child tables (one victim table, one perpetrator table). In this meta-data, the case number is identified as an id that connects the inter-dependent rows across those three tables.

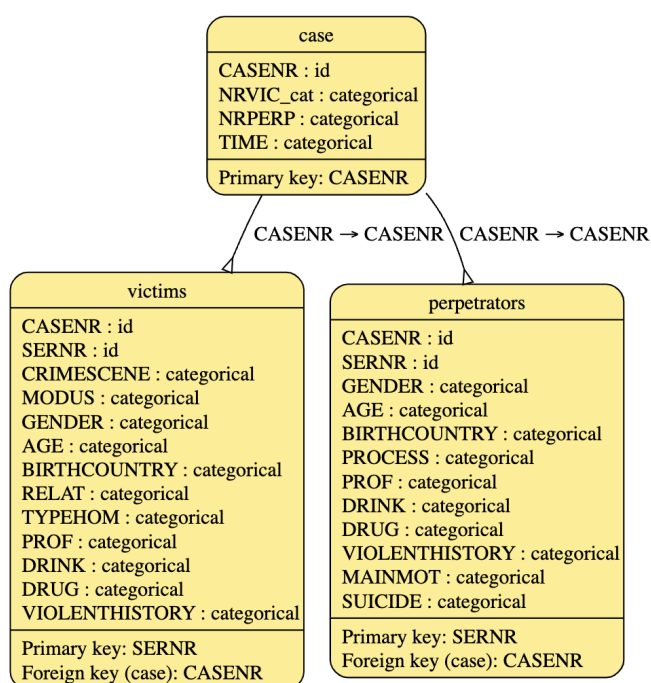


Figure 5.1: Dutch Homicide Monitor split into three separate datasets: cases, victims & perpetrators

In addition, we prepared the logical dependencies as identified in the preparation phase. Therefore, we wrote a code (available on Github) that would force the model during the synthetic data generation to adhere to the rules set out about the dependency of child homicide and victim age, as well as between intimate partner homicide and victim-

perpetrator relationship.

Step 3: Generation of synthetic data

The meta-data information prepared in the previous step is then used to create a synthesiser, that is an object that creates synthetic data. The meta-information provides the synthesiser with the relevant information about the structure of the original data, such as the number of columns or categories within a variable. This structure determines the desired structure of the synthetic dataset. After, we loaded the original data into this synthesiser through which it can learn about the patterns and relationships between the variables in our dataset. Finally, the synthesiser generates a synthetic dataset using the structure and identified relationships. We chose to generate as many cases as are included in the original dataset.

```
1 from sdv.multi_table import HMA Synthesizer
2
3 synthesizer = HMASynthesizer(metadata, #this creates a model that
    knows the structure of the original dataset
4                                     locales=['nl_NL'] #this generates
    city names in Dutch
5
6 synthesizer.add_constraints(constraints=[
7     infanticide_constraint, IPH_constraint
8 ])
9
10 synthesizer.fit(DHM) #this tells the model to learn from our
    original dataset
11 synthetic= synthesizer.sample(scale=1) #this tells the model to
    create a synthetic dataset with as many rows as the original
    dataset
```

Listing 5.1: Synthetic Data Generation: First Attempt

The result are three synthetic tables that together form the first synthetic version of the Dutch Homicide Monitor.

| Case number | Number of victims | Number of perpetrators | Time of day |
|---------------|-------------------|------------------------|-------------|
| sdv-id-nFatwl | 1 | 1 | 6am-12am |
| sdv-id-MulxVy | 1 | 1 | 12pm-6pm |
| sdv-id-gzVJSQ | 2 | 2 | 12pm-6pm |
| ... | ... | ... | ... |

Table 5.4: Synthetic case dataset (first three cases)

| Case number | Serial number | Modus operandi | Gender | Age | ... |
|---------------|---------------|-----------------------------|---------|---------|-----|
| sdv-id-nFatwl | sdv-id-fgFyn | Firearm | Missing | 65 | ... |
| sdv-id-MulxVy | sdv-id-MwbyTx | Missing | Missing | Missing | ... |
| sdv-id-gzVJSQ | sdv-id-cQNzXJ | Knife or other sharp object | Male | Missing | ... |
| sdv-id-gzVJSQ | sdv-id-xgDRmZ | Knife or other sharp object | Missing | 24 | ... |
| ... | ... | ... | ... | ... | ... |

Table 5.5: Synthetic victim dataset (associated with first three cases)

| Case number | Serial number | Gender | Age | Country of birth | ... |
|---------------|---------------|--------|-----|------------------|-----|
| sdv-id-nFatwl | sdv-id-Edlwqz | Male | 19 | Missing | ... |
| sdv-id-MulxVy | sdv-id-LWxGTc | Male | 17 | Missing | ... |
| sdv-id-gzVJSQ | sdv-id-imlpGf | Male | 38 | Netherlands | ... |
| sdv-id-gzVJSQ | sdv-id-UINuoi | Male | 24 | Netherlands | ... |
| ... | ... | ... | ... | ... | ... |

Table 5.6: Synthetic perpetrator dataset (first three cases)

Step 4: Privacy and Utility Evaluation

With regards to the utility, the generated synthetic dataset was of moderate quality. The SDV package offers several options within the package to evaluate the quality of the synthetic data. First, one can run diagnostics on the synthetic data structure and validity, which test whether the synthesiser adhered to the categories of the original dataset, that variables classified as id have unique numbers for each row and that all of the variables have been adapted into the synthetic dataset. In our synthetic dataset, both statistics were at 100 percent, meaning that the structure and validity of the original dataset was fully adopted onto the synthetic dataset.

Secondly, SDV can create a quality report of the synthetic dataset that measures the similarity of the synthetic dataset to the original dataset, based on a comparison of the univariate and bivariate distributions in the synthetic and original dataset. In this attempt, the scores for the univariate distribution measure was 79.17%, and the measure for bivariate distributions 57.21%, as shown in the table below. Additional reports show that some variables had a significantly lower score than others.

| Quality Type | Score (%) | Interpretation |
|---|-----------|--|
| Data validity | 100 | The synthetic data has the same internal structure as the original one |
| Data structure | 100 | The synthetic data has the same columns with the same names |
| Relationship validity | 99.94 | Each case in the victim & perpetrator datasets are linked to a case in the case dataset |
| Univariate distribution | 79.17 | A variable in the synthetic dataset is on average 80% similar to the variable in the real dataset |
| Bivariate distribution within one table | 57.21 | The relationship between two variables in the synthetic data is about 57% similar to the relationship in the real dataset |
| Structure across tables | 94.98 | Almost each case in the synthetic dataset has the same number of victims and perpetrators associated as in the real dataset |
| Bivariate distribution across tables | 79.91 | The relationship between two variables across the three synthetic datasets is about 80% similar to the relationship in the real data |
| Overall | 75.82 | Overall, the synthetic dataset compares around 75% to the real data |

Table 5.7: General quality report

| Variable | Score (%) |
|---|-----------|
| Perpetrator's main motive | 90.54 |
| Perpetrator committed or attempted suicide | 89.61 |
| Perpetrator drug (ab)use | 88.8 |
| Perpetrator gender | 84.74 |
| Perpetrator birth country | 83.58 |
| Perpetrator violent history | 82.77 |
| Perpetrator profession | 81.61 |
| Perpetrator age | 81.61 |
| Perpetrator alcohol (ab)use | 30.16 |
| Status of case against perpetrator in judicial system | 30.1 |

Table 5.8: Example: quality of each variable in the synthetic perpetrator dataset, ranked from highest to lowest similarity score

The varying degrees of quality across the synthesised variables is also visible in visualisations that compare the univariate and bivariate distributions of the synthetic and original dataset. Through these visualisations, we determined that about 70 percent of the variables in the synthetic dataset compared well to the original data, such as the time of day the homicide was committed.

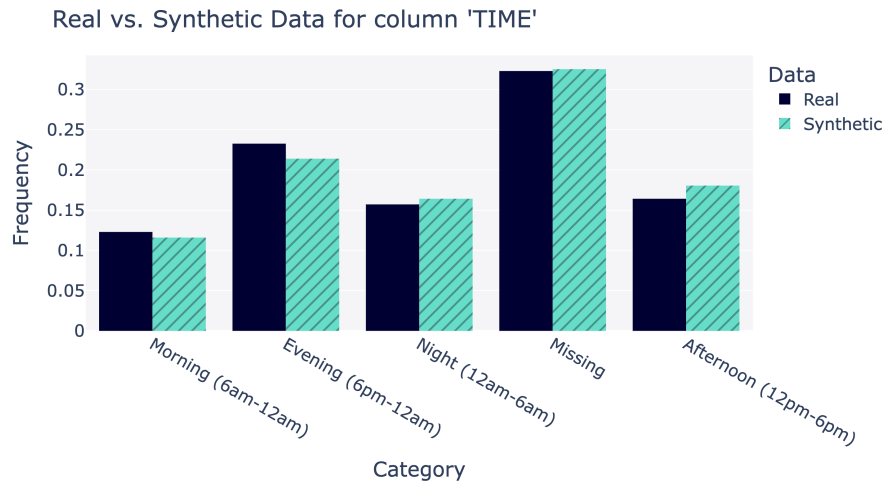


Figure 5.2: Example of a synthesized variable with high quality score

However, the univariate distribution in other variables, such as the judicial process, was significantly worse and below the desired threshold.

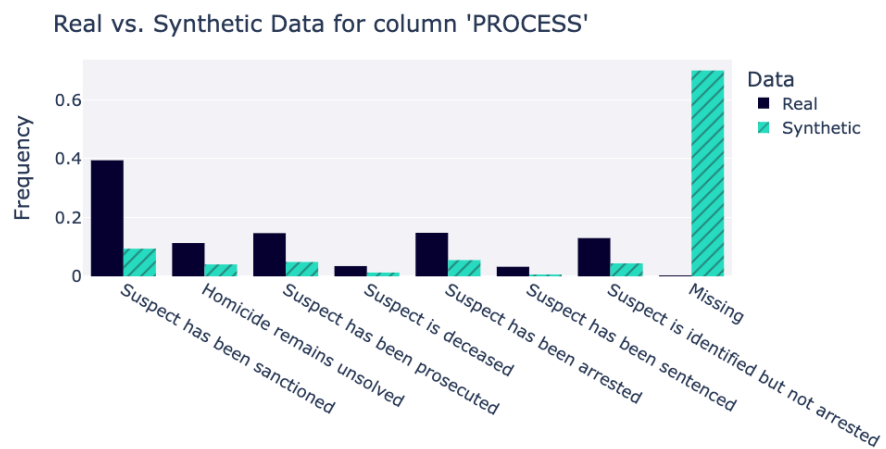


Figure 5.3: Example of a synthesized variable with low quality score

Similarly, significant limitations are visible in the bivariate distributions, that is the relationship between variables.

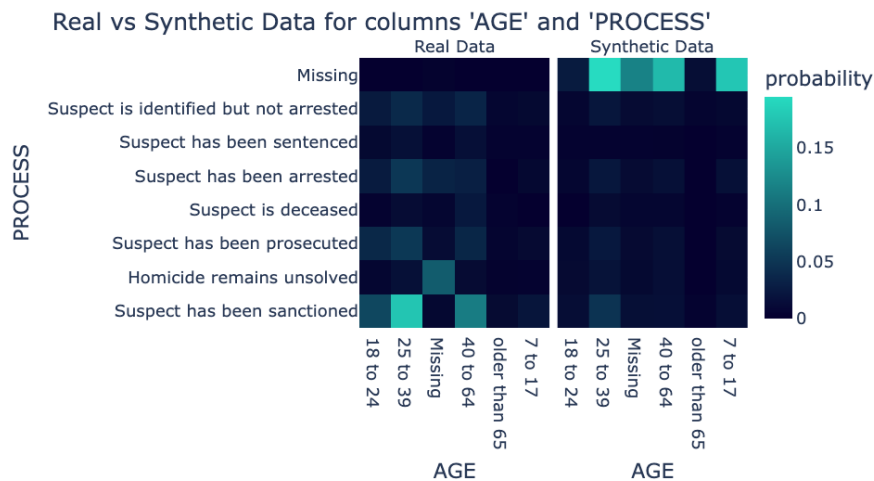


Figure 5.4: Example of bivariate relationship comparison across original and synthetic data

Simply based on these limitations, we concluded that the synthetic dataset did match the requirements set out in the beginning. As such, we did not conduct further analysis, such as replication studies, or an evaluation of the privacy guarantees of the synthetic data.

Reflections

After deliberations, we determined two potential causes for the lower than expected quality of the synthetic dataset. First, our data is relatively complex as it spans over three table, meaning that the synthesiser has to learn not only about the patterns and relationships within *one* table (e.g. the relationship between the victim's gender and age) but also across the three tables (e.g. the relationship between type of crime scene, gender of the victim and gender of the perpetrator). In addition, the synthesiser has to replicate relatively many variables (29 divided over the three tables) and many categories within those variables. As mentioned previously, some of the variables have up to 35 categories, with each of these categories having certain patterns and relationships with all the other categories across all variables. Given that about 70 percent of the variables in the synthetic dataset compared well to the original dataset, but the remaining 30 percent did not, we believe that the synthesiser over-corrected, meaning that it tried to

mimic as many patterns and relationships as best as possible at the costs of the quality of the patterns and relationships that remained. A second problem that might have impacted the results is that our dataset contains relatively little data, with 1271 cases, information on 1348 victims and 1711 perpetrators. Although this amount is technically sufficient to train a model on the patterns and relationships within that dataset, more data provides a better ground for detecting all underlying patterns and relationships, especially when the data contains many variables as in this case.

Therefore, we decided to address these causes in a second round, meaning that we started again at the step of data preparation.

5.2.2 Second attempt

Step 1: Preparations

In order to address the possible causes of the limited data utility in the first attempt, we decided to **(a)** reduce the complexity of the data as much as possible, and **(b)** to increase the amount of data.

To reduce the complexity of the data, we decided to minimise both the number of variables, as well as the number of categories within the variables. In total, we excluded six variables: the profession of the victim or perpetrator, whether the victim or perpetrator was under the influence of drugs or alcohol, whether the victim or perpetrator had a violent history, the main motive of the perpetrator and whether the perpetrator committed or attempted suicide. These variables had a significant amount of missing data (and therefore the lowest quality) and were deemed the least important by the project team members with expertise on homicide research. To reduce the number of categories within each variable, we first detected categories with only few cases, such as sexual homicides. After, we aimed to merge these categories together into categories of 'other', based on the number of cases, as well as on logical reasoning. For example, the Dutch Homicide Monitor differentiates between three different types of child homicides: infanticides - the killing of newborns -, the homicide of children by someone in the child's family and the killing of a child by someone outside of the child's family. For the sake of decreasing categories, these three types were merged into one category: *child homicide*. Yet, in order to keep as much detail as possible in the original and thus also the synthetic data,

we merged as little categories as deemed necessary at this stage in the process.

To increase the amount of data through which the synthesiser can learn about the patterns and relationships, we decided to broaden the scope of homicide data included, from ten years of homicide data to twenty years of homicide data. As a result, the original dataset used to train the synthesiser now contained 3152 cases, and information on 3358 victims and 4394 perpetrators.

Step 2: Choosing generation method and tool

The generation method and chosen tool (Synthetic Data Vault) remained the same. However, given that we added new data and recoded some of our variables, we re-run the steps for data cleaning and the preparation of the meta-data.

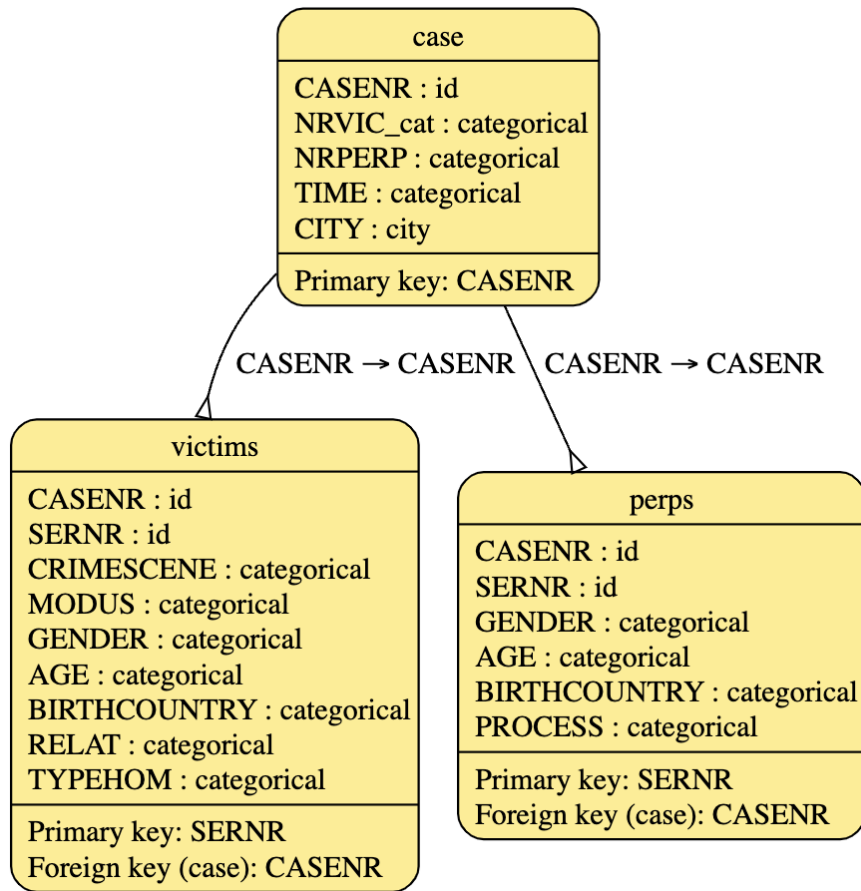


Figure 5.5: Dutch Homicide Monitor split into three separate datasets: cases, victims & perpetrators

Step 3: Generation

As in the previous attempt, we created a synthesiser based on the new meta-data and fed the synthesiser our new dataset to learn the patterns and relationships. Yet, considering that the new adaptations were made in the original dataset, the process and code remained the same.

```

1 from sdv.multi_table import HMA Synthesizer
2
3 synthesizer = HMASynthesizer(metadata, #this creates a model that
    knows the structure of the original dataset
4                                     locales=['nl_NL'] #this generates
    city names in Dutch
5
6 synthesizer.add_constraints(constraints=[
7     infanticide_constraint, IPH_constraint
8 ])
9
10 synthesizer.fit(DHM) #this tells the model to learn from our
    original dataset
11 synthetic= synthesizer.sample(scale=1) #this tells the model to
    create a synthetic dataset with as many rows as the original
    dataset

```

Listing 5.2: Synthetic Data Generation: Second Attempt

Step 4: Privacy and utility evaluation

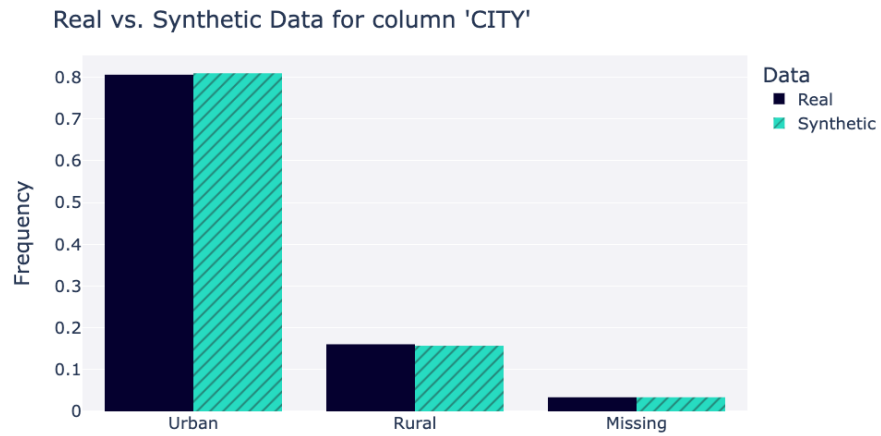
With regards to the utility of the data, we detected similar issues as during the first attempt. Again, most of the variables in the synthesised dataset compared well to the original dataset, yet other variables were synthesised poorly. The bivariate relationships of the variables improved compared to the first attempt, yet were still not at a quality that we were satisfied that this synthetic dataset fulfils the statistical requirements set in the beginning.

| Quality Type | Score (%) | Interpretation |
|---|-----------|--|
| Data validity | 100 | The synthetic data has the same internal structure as the original one |
| Data structure | 100 | The synthetic data has the same columns with the same names |
| Relationship validity | 99.99 | Each case in the victim & perpetrator datasets are linked to a case in the case dataset |
| Univariate distribution | 75.84 | A variable in the synthetic dataset is on average 75% similar to the variable in the real dataset |
| Bivariate distribution within one table | 51.51 | The relationship between two variables in the synthetic data is about 50% similar to the relationship in the real dataset |
| Structure across tables | 90.97 | Around 90% of cases in the synthetic dataset has the same number of victims and perpetrators associated as in the real dataset |
| Bivariate distribution across tables | 63.9 | The relationship between two variables across the three synthetic datasets is about 64% similar to the relationship in the real data |
| Overall | 70.56 | Overall, the synthetic dataset compares around 70% to the real data |

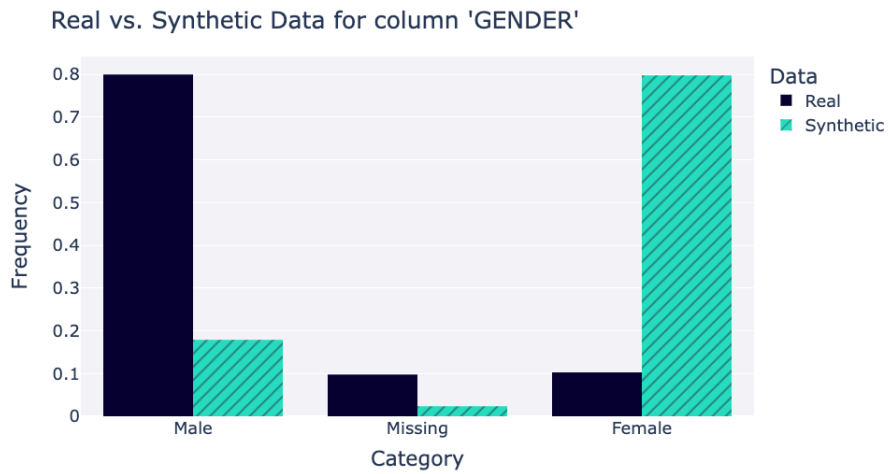
Table 5.9: General quality report

| Variable | Score (%) |
|---|-----------|
| Perpetrator birth country | 89.87 |
| Status of case against perpetrator in judicial system | 82.36 |
| Perpetrator age | 75.78 |
| Perpetrator gender | 30.5 |

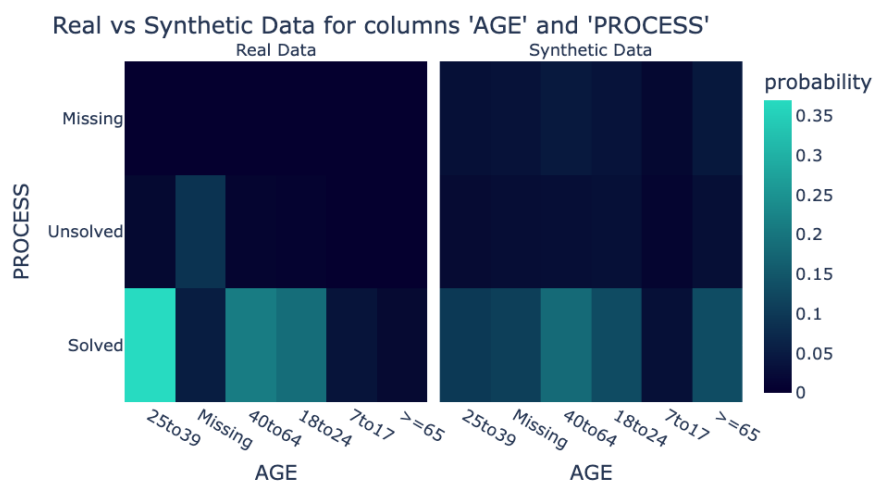
Table 5.10: Example: quality of each variable in the synthetic perpetrator dataset, ranked from highest to lowest similarity score



(a) Example of a synthesised variable with high quality score



(b) Example of a synthesised variable with low quality score



(c) Example of bivariate relationship comparison across original and synthetic data

Figure 5.6: Quality of the second synthetic version of the DHM (continued) 55

Again, we did not conduct any privacy analyses due to the low level of utility.

Reflections

Going back to the potential causes of the limited quality, we concluded that the increased amount of data had little impact on the quality of the synthetic dataset. Instead, we had to again review how to reduce the complexity in our data.

Initially, we followed the same procedure as in our previous attempt, by merging categories within variables wherever logically possible. For example, the variable on victim-perpetrator relationship went from 35 categories to only ten categories.

Unfortunately, although the quality of the synthetic dataset improved slightly with each merging, it did not reach a level of quality that was satisfying. Moreover, at a certain point, even further merging of the variables would have significantly changed the level of detail in the dataset, which is one of the main advantages of the original Dutch Homicide Monitor.

5.2.3 Final attempt

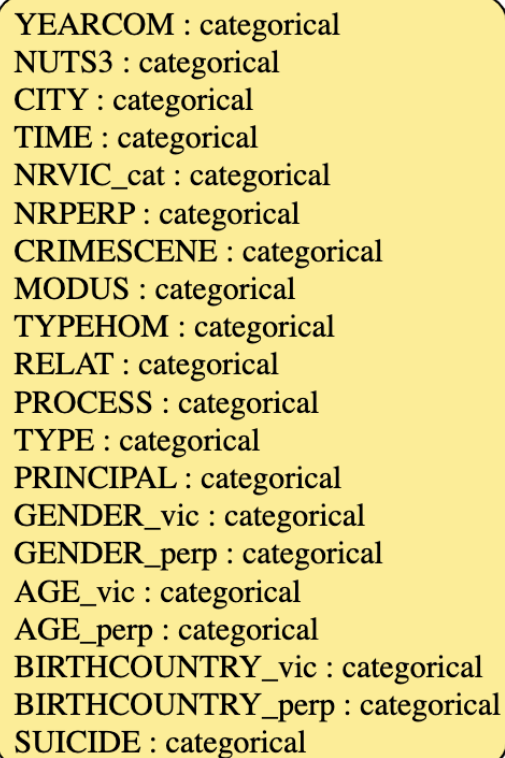
Step 1: Preparations

With this conclusion in mind, we decided to reduce the complexity, by reducing the multi-table approach to a single-table approach in which the synthesiser only has to recognise the patterns and relationships across one table instead of three. This meant that we had to merge the three tables with the original data (case table, victim table and perpetrator table) back into one single table. With the original structure, the synthesiser would be able to learn the dependencies across the rows, thus individuals that are involved as either victim or perpetrator in the same case. As such, we had to concede that we would not be able to create a synthetic dataset that would allow analysis on the level of case, victim **and** perpetrator. Based on domain expertise, we decided to merge the information of the main perpetrator onto the row of each associated victim and add a variable that would define one main victim per case. As a result, all information on each individual case and each individual victim is kept in the dataset, yet information for some perpetrators (specifically in cases with multiple perpetrators) were lost. In addition, to

increase the quality of the remaining data, we decided to again only use homicide data of ten years as this is the most complete data in the original dataset. Finally, we also reverted most of the merging of categories within the variables, in order to keep as much detail as possible.

Step 2: Choosing generation method and tool]

Given the changed structure of the original dataset, we had to revisit the requirements of the Synthetic Data Vault for single table synthesis. In essence, only few changes had to be made, in particular with regards to the meta-data.



YEARCOM : categorical
NUTS3 : categorical
CITY : categorical
TIME : categorical
NRVIC_cat : categorical
NRPERP : categorical
CRIMESCENE : categorical
MODUS : categorical
TYPEHOM : categorical
RELAT : categorical
PROCESS : categorical
TYPE : categorical
PRINCIPAL : categorical
GENDER_vic : categorical
GENDER_perp : categorical
AGE_vic : categorical
AGE_perp : categorical
BIRTHCOUNTRY_vic : categorical
BIRTHCOUNTRY_perp : categorical
SUICIDE : categorical

Figure 5.7: Dutch Homicide Monitor as one dataset, with each row representing one victim

Step 3: Generation

As in the previous steps, we used the meta-data of the single table with homicide information to create a synthesiser, and trained said synthesiser with the original dataset.

```

1 from sdv.single_table import GaussianCopulaSynthesizer
2
3 synthesizer = GaussianCopulaSynthesizer(metadata,
4                                         locales=['nl_NL'])
5
6 synthesizer.add_constraints([IPHConstraint, kindermoordConstraint])
7
8 synthesizer.fit(DHM)
9 synthetic=synthesizer.sample(num_rows=1364)

```

Listing 5.3: Synthetic Data Generation: Final Attempt

Step 4: Privacy and Utility Evaluation

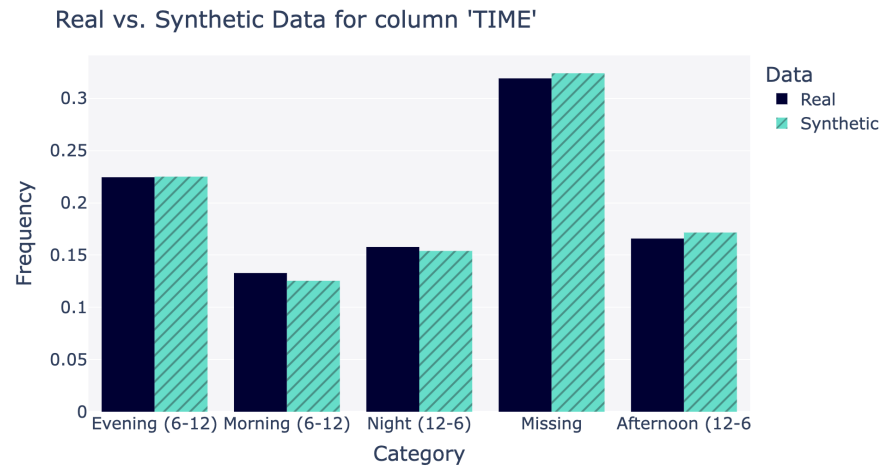
In this attempt, both the comparison of the univariate as well as bivariate patterns between the original and synthetic data looked significantly improved, which is supported by the quality report provided by the Synthetic Data Vault package.

| Quality Type | Score (%) | Interpretation |
|---|-----------|---|
| Data validity | 100 | The synthetic data has the same internal structure as the original one |
| Data structure | 100 | The synthetic data has the same columns with the same names |
| Univariate distribution | 97.8 | A variable in the synthetic dataset is on average 98% similar to the variable in the real dataset |
| Bivariate distribution within one table | 89.67 | The relationship between two variables in the synthetic data is about 90% similar to the relationship in the real dataset |
| Overall | 93.74 | Overall, the synthetic dataset compares around 94% to the real data |

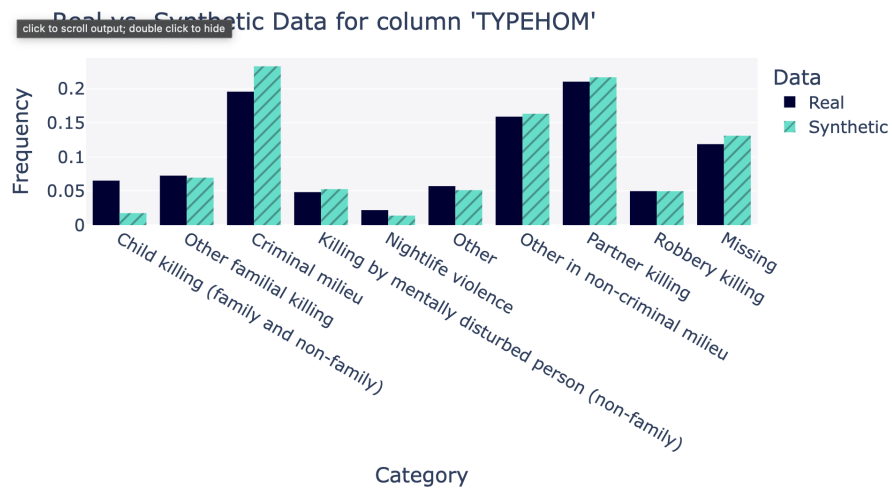
Table 5.11: General quality report

| Variable | Score (%) |
|---|-----------|
| Type: victim or perpetrator? | 99.93 |
| City homicide was committed in | 99.79 |
| Number of perpetrators | 99.21 |
| Status of case against perpetrator in judicial system | 99.13 |
| Number of victims | 98.98 |
| Time of day homicide was committed | 98.87 |
| Perpetrator gender | 98.81 |
| Perpetrator birthcountry | 98.3 |
| Type of crimescene | 98.04 |
| Victim gender | 97.9 |
| Victim age | 97.38 |
| Perpetrator age | 97.32 |
| Modus operandi | 97.11 |
| Victim-perpetrator relationship | 96.95 |
| Victim birthcountry | 96.89 |
| Year homicide was committed | 95.2 |
| Region homicide was committed | 94.92 |
| Context of homicide | 93.53 |

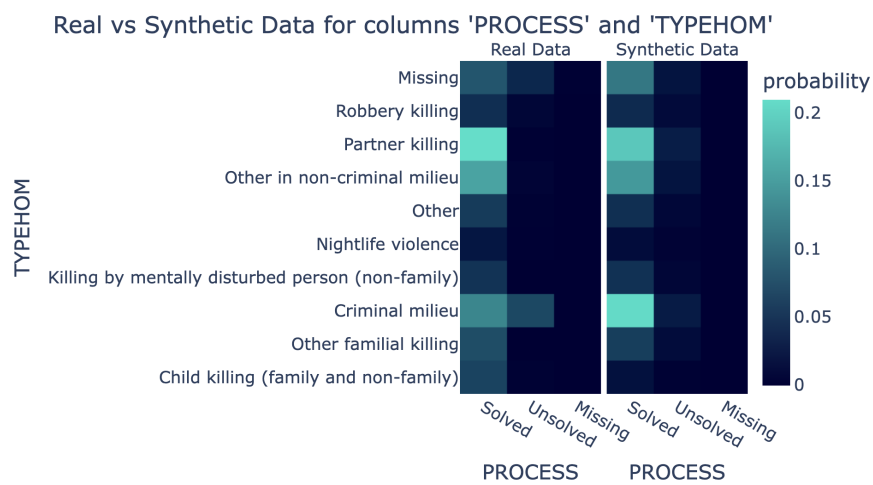
Table 5.12: Quality of each variable in the synthetic dataset, ranked from highest to lowest similarity score



(a) Example of a synthesised variable with high quality score



(b) Example of a synthesised variable with the lowest quality score



(c) Example of bivariate relationship comparison across original and synthetic data

Figure 5.8: Quality of the final synthetic version of the DHM (continued)

To check the extent of the statistical accuracy of the synthetic dataset, we conducted two replication studies, based on studies previously conducted with the original data. Although the previous studies used homicide data for different years, we expected the results to be similar with the ten year of homicide data included in the synthetic data. Our expectations were (mostly) met:

add example study 1

Overall, this synthetic version of the Dutch Homicide Monitor satisfied our requirements with regards to data utility.

With regards to data privacy, we conducted several tests to check whether the safeguarding of sensitive and personal data is equally satisfying. First, we randomly selected about twenty cases from the original dataset and tried to identify cases that match their profiles in the synthetic dataset. In addition, we identified certain outliers in the original dataset, e.g. individuals with unusual combination of attributes, and tested whether we were able to identify similar cases in the synthetic dataset. In both cases, were not able to detect one-on-one matches between the original and synthetic dataset with our chosen sample and outliers. Finally, we used privacy-metrics included in the SDV-package, the so-called *Privacy Against Inference* metrics. These metrics calculate the risk of an attacker learning sensitive information based on the synthetic dataset, assuming that the attacker already has some information based on real data. The metrics allow for a simulation of this risk with all possible variable combinations.

```
1 from sdmetrics.single_table import CategoricalKNN,CategoricalRF
2
3 CategoricalKNN.compute(
4     real_data=case,
5     synthetic_data=synthetic,
6     key_fields=['GENDER_vic','CRIMESCENE'], #these are the
        variables we assume the attacker already knows
7     sensitive_fields=['RELAT'] #this is the variable the attacker
        wants to know
```

Listing 5.4: SDV privacy metrics example

For each variable-combination, the algorithms provides a percentage; 0% means that the data is not safe at all, 100% means that the data is completely safe. In the example in the listing above, the information about the victim-perpetrator relationship was 79.7% safe, given that the attacker already knows the victim's gender and the type of crimescene.

5.3 Concluding Reflections

5.3.1 The Process

Taking into account data preparation and necessary adaptations for each cycle, the overall process only took about two full days; the actual generation of the synthetic data between a few seconds for final single table to a few minutes for the most complex multi-table. Moreover, through the detailed documentation provided by the Synthetic Data Vault library, as well as a helpful community of users online, any questions or problems along the process have quickly been answered or solved. In addition, the synthesis of data with the use of the Synthetic Data Vault has proven very accessible, even for someone without extensive knowledge of programming or synthetic data.

5.3.2 The Data

In the beginning of the process, we set high aims and requirements for the synthetic version of the Dutch Homicide Monitor. In simple terms, we wanted to create a synthetic dataset that mimics the original data enough to be used for advanced scientific analysis, to be informative for non-scientific purposes but to safeguard the privacy of all individuals included in the dataset. In the end, not all of the requirements have been met, in particular the requirement that the full original dataset had to be synthesised. Due to the complexity of the original data impacting the quality of the synthetic data, we had to opt to exclude information on certain perpetrators (in cases in which there were more than one perpetrator). Thus, information from the original dataset was lost during the process. The final synthetic dataset, however, provides a good balance between privacy assurance and data quality.

5.3.3 Open & FAIR data

add With regards to open science principles, the synthetic dataset has been developed with the FAIR principles in mind. Therefore, the full disaggregated data can be found, used and downloaded at the projects webpage **add link** **add screenshot webpage**

Chapter 6

Conclusion: Enhancing Open Science Practices through Synthetic Data

Strict regulations surrounding personal and other sensitive data can seem at odds with the implementation of Open Science principles, creating hinderances for researchers, research support and other stakeholders. Synthetic data carries a lot of promises for enhancing data sharing practices, open datasets and the implementation of FAIR principles, in particular for sensitive data. It allows for the creation of artificial datasets that mimic real-world data whilst protecting private and personal information. These datasets can be shared, registered, fit to other open science standards and be used without many restrictions. So far, synthetic data is mostly discussed amongst data scientists, who develop new measures of utility and privacy, new computations for syntheses processes and advance already existing approaches. However, synthetic data is not applied much in research that does not focus on synthetic data as the topic of research yet. With fast developments in artificial intelligence and data science, synthetic data promises to become more and more useful in the future.

We hope this guidebook provides a short introduction to the ideas and implementation of synthetic data, as well as practical tips on how synthetic data can be applied in everyday research to make it more open, findable, interoperable, accessible, and reusable.



Chapter 7

Appendix: Resources

7.1 Resources on Open Science

7.2 Resources on Open Data and Personal Data

7.3 Resources on synthetic data

Bibliography

- [1] UNESCO. *UNESCO Recommendation on Open Science*. SC-PCB-SPP/2021/OS/UROS. United Nations Educational, Scientific and Cultural Organization, 2021.
- [2] Mark D Wilkinson et al. “The FAIR Guiding Principles for scientific data management and stewardship”. In: *Scientific data* 3.1 (2016), pp. 1–9.
- [3] General Data Protection Regulation GDPR. *General data protection regulation*. 2016.
- [4] William Alex Pridemore, Matthew C Makel, and Jonathan A Plucker. “Replication in criminology and the social sciences”. In: *Annual Review of Criminology* 1.1 (2018), pp. 19–38.
- [5] Jeremy Freese, Tamkinat Rauf, and Jan Gerrit Voelkel. “Advances in transparency and reproducibility in the social sciences”. In: *Social Science Research* 107 (2022), p. 102770.
- [6] Stephen M Powers and Stephanie E Hampton. “Open science, reproducibility, and transparency in ecology”. In: *Ecological applications* 29.1 (2019), e01822.
- [7] Khaled El Emam, Lucy Mosquera, and Richard Hoptroff. *Practical synthetic data generation: balancing privacy and the broad availability of data*. O’Reilly Media, 2020.
- [8] Jiri Hradec et al. *Multipurpose synthetic population for policy applications*. Joint Research Centre (JRC), 2022.
- [9] James Jordon et al. “Synthetic Data—what, why and how?” In: *arXiv preprint arXiv:2205.03257* (2022).
- [10] Chao Yan et al. “A multifaceted benchmarking of synthetic electronic health record generation models”. In: *Nature communications* 13.1 (2022), p. 7609.

- [11] Yalin E Sagduyu, Alexander Grushin, and Yi Shi. “Synthetic social media data generation”. In: *IEEE Transactions on Computational Social Systems* 5.3 (2018), pp. 605–620.
- [12] Jianfu Li et al. “Are synthetic clinical notes useful for real natural language processing tasks: A case study on clinical entity recognition”. In: *Journal of the American Medical Informatics Association* 28.10 (2021), pp. 2193–2201.
- [13] Yogesh Patel et al. “Deepfake generation and detection: Case study and challenges”. In: *IEEE Access* (2023).
- [14] Anthony Paproki, Olivier Salvado, and Clinton Fookes. “Synthetic Data for Deep Learning in Computer Vision & Medical Imaging: A Means to Reduce Data Bias”. In: *ACM Computing Surveys* (2024).
- [15] Ghanem Soltana, Mehrdad Sabetzadeh, and Lionel C Briand. “Synthetic data generation for statistical testing”. In: *2017 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE. 2017, pp. 872–882.
- [16] Christian Arnold and Marcel Neunhoeffler. “Really Useful Synthetic Data—A Framework to Evaluate the Quality of Differentially Private Synthetic Data”. In: *arXiv preprint arXiv:2004.07740* (2020).
- [17] Aldren Gonzales, Guruprabha Guruswamy, and Scott R Smith. “Synthetic data in health care: A narrative review”. In: *PLOS Digital Health* 2.1 (2023), e0000082.
- [18] Anmol Arora and Ananya Arora. “Generative adversarial networks and synthetic patient data: current challenges and future perspectives”. In: *Future Healthcare Journal* 9.2 (2022), pp. 190–193.
- [19] Amy Elise Braddon et al. “Exploring the utility of synthetic data to extract more value from sensitive health data assets: A focused example in perinatal epidemiology”. In: *Paediatric and Perinatal Epidemiology* 37.4 (2023), pp. 292–300.
- [20] Samuel A Assefa et al. “Generating synthetic data in finance: opportunities, challenges and pitfalls”. In: *Proceedings of the First ACM International Conference on AI in Finance*. 2020, pp. 1–8.
- [21] Florian Eckerli and Joerg Osterrieder. “Generative adversarial networks in finance: an overview”. In: *arXiv preprint arXiv:2106.06364* (2021).
- [22] Jeremy Kravitz et al. “Potential for high fidelity global mapping of common inland water quality products at high spatial and temporal resolutions based on

- a synthetic data and machine learning approach”. In: *Frontiers in Environmental Science* 9 (2021), p. 587660.
- [23] Fernando-Juan Pérez-Porras et al. “Machine learning methods and synthetic data generation to predict large wildfires”. In: *Sensors* 21.11 (2021), p. 3694.
 - [24] Ian Brunton-Smith et al. “Using Synthetic Crime Data to Understand Patterns of Police Undercounting at the Local Level”. In: *The Crime Data Handbook*. Bristol University Press, 2024, pp. 60–79.
 - [25] Theodora Kokosi et al. “An overview of synthetic administrative data for research”. In: *International Journal of Population Data Science* 7.1 (2022).
 - [26] Abdul Majeed and Sungchang Lee. “Anonymization techniques for privacy preserving data publishing: A comprehensive survey”. In: *IEEE access* 9 (2020), pp. 8512–8545.
 - [27] Cynthia Dwork. “Differential privacy”. In: *International colloquium on automata, languages, and programming*. Springer. 2006, pp. 1–12.
 - [28] Larry Wasserman and Shuheng Zhou. “A statistical framework for differential privacy”. In: *Journal of the American Statistical Association* 105.489 (2010), pp. 375–389.
 - [29] Alexandra Wood et al. “Differential privacy: A primer for a non-technical audience”. In: *Vand. J. Ent. & Tech. L.* 21 (2018), p. 209.
 - [30] James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. “PATE-GAN: Generating synthetic data with differential privacy guarantees”. In: *International conference on learning representations*. 2018.
 - [31] Bangzhou Xin et al. “Federated synthetic data generation with differential privacy”. In: *Neurocomputing* 468 (2022), pp. 1–10.
 - [32] Chang Sun, Johan van Soest, and Michel Dumontier. “Generating synthetic personal health data using conditional generative adversarial networks combining with differential privacy”. In: *Journal of Biomedical Informatics* 143 (2023), p. 104404.
 - [33] Marco Ajelli et al. “The RAPIDD Ebola forecasting challenge: Model description and synthetic data generation”. In: *Epidemics* 22 (2018), pp. 3–12.
 - [34] Sanket Kamthe, Samuel Assefa, and Marc Deisenroth. “Copula flows for synthetic data generation”. In: *arXiv preprint arXiv:2101.00598* (2021).

- [35] Vasileios C Pezoulas et al. “Synthetic data generation methods in healthcare: a review on open-source tools and methods”. In: *Computational and Structural Biotechnology Journal* (2024).
- [36] Kunfeng Wang et al. “Generative adversarial networks: introduction and outlook”. In: *IEEE/CAA Journal of Automatica Sinica* 4.4 (2017), pp. 588–598.
- [37] Carl Doersch. “Tutorial on variational autoencoders”. In: *arXiv preprint arXiv:1606.05908* (2016).
- [38] Zhiqiang Wan, Yazhou Zhang, and Haibo He. “Variational autoencoder based synthetic data generation for imbalanced learning”. In: *2017 IEEE symposium series on computational intelligence (SSCI)*. IEEE. 2017, pp. 1–7.
- [39] Zhuoyan Li et al. “Synthetic data generation with large language models for text classification: Potential and limitations”. In: *arXiv preprint arXiv:2310.07849* (2023).
- [40] Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. “The synthetic data vault”. In: *2016 IEEE international conference on data science and advanced analytics (DSAA)*. IEEE. 2016, pp. 399–410.
- [41] Zhaozhi Qian, Rob Davis, and Mihaela van der Schaar. “Synthcity: a benchmark framework for diverse use cases of tabular synthetic data”. In: *Advances in Neural Information Processing Systems* 36 (2024).
- [42] YData. *YData Synthetic Data Generation - Documentation*. <https://docs.synthetic.ydata.ai/1.4/>. Accessed: 11 July 2024. 2024.
- [43] Daniele Faraglia. *Faker’s documentation*. <https://faker.readthedocs.io/en/master/>. Accessed: 11 July 2024. 2024.
- [44] Beata Nowok, Gillian M Raab, and Chris Dibben. “synthpop: Bespoke creation of synthetic data in R”. In: *Journal of statistical software* 74 (2016), pp. 1–26.
- [45] Salvador Ochoa et al. “Reidentification of individuals in Chicago’s homicide database: A technical and legal study”. In: *Massachusetts Institute of Technology* (2001).
- [46] Khaled El Emam and Fida Kamal Dankar. “Protecting privacy using k-anonymity”. In: *Journal of the American Medical Informatics Association* 15.5 (2008), pp. 627–637.
- [47] Jihyeon Hyeong et al. “An empirical study on the membership inference attack against tabular data synthesis models”. In: *Proceedings of the 31st ACM Interna-*

- tional Conference on Information & Knowledge Management*. 2022, pp. 4064–4068.
- [48] Markus Hittmeir, Rudolf Mayer, and Andreas Ekelhart. “A baseline for attribute disclosure risk in synthetic data”. In: *Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy*. 2020, pp. 133–143.
 - [49] Markus Hittmeir, Andreas Ekelhart, and Rudolf Mayer. “Utility and privacy assessments of synthetic data for regression tasks”. In: *2019 IEEE International Conference on Big Data (Big Data)*. IEEE. 2019, pp. 5763–5772.
 - [50] Jingchen Hu. “Bayesian estimation of attribute and identification disclosure risks in synthetic data”. In: *arXiv preprint arXiv:1804.02784* (2018).
 - [51] Alexander Theodorus Petrus Boudewijn et al. “Privacy measurements in tabular synthetic data: State of the art and future research directions”. In: *NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI*. 2023.
 - [52] Boris Van Breugel et al. “Membership inference attacks against synthetic data through overfitting detection”. In: *arXiv preprint arXiv:2302.12580* (2023).
 - [53] Fida K Dankar and Mahmoud Ibrahim. “Fake it till you make it: Guidelines for effective synthetic data generation”. In: *Applied Sciences* 11.5 (2021), p. 2158.
 - [54] Khaled El Emam. “Seven ways to evaluate the utility of synthetic data”. In: *IEEE Security & Privacy* 18.4 (2020), pp. 56–59.
 - [55] Roger B Nelsen. *An introduction to copulas*. Springer, 2006.