Leveraging synthetic data to improve open science practices in criminology: releasing a synthetic homicide dataset and dashboard

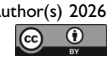
Krusselmann, K.; Achterberg, J.; Haas, M.; Spruit, M.R; Liem, M.C.A.

Special Issue: Open science and open data in European criminological research

European Journal of Criminology

Leveraging synthetic data to improve open science practices in criminology – releasing a synthetic homicide dataset and dashboard

European Journal of Criminology
1–26



DOI: 10.1177/14773708261469812
journals.sagepub.com/home/euc


Katharina Krüsselmann[1,2], Jim Achterberg[3,4], Marcel Haas[3], Marco Spruit[2,3], and Marieke Liem[2]

Abstract

In criminology, the call for transparency and openness in research practices is resounding louder than ever. FAIR (*f*indable, *a*ccessible, *i*nteroperable and *r*eusable) and open data are recognized as a necessity to enable replication of criminological research, enhance future empirical studies, foster collaboration and inform evidence-based policy-making. Yet, criminologists are faced with an intricate web of privacy regulations governing the dissemination of sensitive data about crime victims, perpetrators, witnesses or other stakeholders. In this article, we examine criminology-specific challenges posed by regulations like General Data Protection Regulation, which mandate safeguarding of personally identifiable information, often rendering original data unsuitable for open sharing. To reconcile this dilemma, we discuss the use of synthetic data generation as a tool to overcome (parts of) this challenge. Synthetic data entails creating artificial datasets that mimic the statistical characteristics of real data while obfuscating identifying information. By utilizing synthetic data, criminologists can both adhere to stringent regulations while disseminating data and facilitating open science practices. This article

[1]Erasmus University Rotterdam, the Netherlands
[2]Leiden University, the Netherlands
[3]Leiden University Medical Center, the Netherlands
[4]Statistics Netherlands (CBS), The Netherlands

Corresponding author:
Katharina Krüsselmann, Erasmus School of Social and Behavioural Science, Erasmus University Rotterdam, Rotterdam, the Netherlands.
Email: Krusselmann@essb.eur.nl

discusses benefits, as well as ethical considerations and limitations of synthetic data. As a proof-of-concept, we showcase the synthesis and publicly release a synthetic version of the Dutch Homicide Monitor, using an existing open-source tool for data synthetization.

Keywords
Data privacy, data utility, homicide, open data, open science, synthetic data

Open data in criminological research

Open science represents a transformative shift in how research is conducted, shared and applied. It is 'an umbrella term reflecting the idea that scientific knowledge of all kinds, where appropriate, should be openly accessible, transparent, rigorous, reproducible, replicable, accumulative, and inclusive […]. Open science consists of principles and behaviors that promote transparent, credible, reproducible and accessible science' (Parsons et al., 2022: 312). Much of the institutional attention and literature in criminology and other fields have focused on open access of scholarly literature as an indicator of open science practices (Ashby, 2021; Jacques, 2023; Severin et al., 2020), yet open science encourages openness throughout the *whole* research cycle, from the development of research ideas, to fieldnotes, analysis code and open data. This study discusses criminology-specific obstacles and opportunities related to open data, showing that synthetic data can overcome ethical concerns and privacy regulations that seemingly restrict open data practices.

The Open Knowledge Foundation defines open data as 'data than can be freely used, shared and built-on by anyone, anywhere, for any purpose' (James, 2013). Open data practices promote transparency, reproducibility, and broader collaboration, and increase the visibility and impact of research by enabling secondary analysis and new insights (Molloy, 2011; Umbach, 2024). From an ethical perspective, it has been argued that publicly funded research and data collected with that funding should benefit the general public as much as possible (Borgman, 2012). For these reasons, the sharing of data is encouraged or even legally obligated by many (inter)national funding agencies (European Commission, 2024; Network of the Marie SkŁodowska-Curie Actions National Contact Points, 2023).

However, these benefits are also associated with challenges such as concerns over data privacy, intellectual property, and potential misuse or misinterpretation. Preparing data for sharing can be time-consuming, requiring proper documentation and ethical considerations. While open data fosters progress, it also demands careful balance between openness and responsibility (Borgman, 2012; Tenopir et al., 2011). Recognizing this challenge of fully open data sharing, the concept of FAIR (*f*indable, *a*ccessible, *i*nteroperable, and *r*eusable) data has been introduced which refers to guiding principles on how to increase the reusability of data through practical guidelines on data management (Wilkinson et al., 2016). FAIR principles recognize essential legal and ethical limitations to the openness of data, but encourage data

management within the given boundaries. The phrase 'as open as possible, as restricted as necessary' (European Commission, Directorate-General for Research and Innovation, 2016: 4) has been established to describe the general willingness of sharing data within the legal and ethical limitations imposed by international, national or institutional regulations.

Criminologists have already noted the benefits of open datasets, in particular of data provided by governmental agencies. A few recent studies have provided an extensive overview of accessible international and national data sources that include crime and criminal justice (Buil-Gil et al., 2025; Chermak et al., 2025; Nivette, 2021) as well as terrorism and political violence data (Bowie, 2021). Public sources such as these are used by dozens of criminological studies to test theoretical assumptions (Sigfusdottir et al., 2012; Yodanis, 2004), describe crime trends (Aebi and Linde, 2010; Rosenfeld and Messner, 2009) or to inform evidence-based crime policies (De Bondt, 2014; Hartel et al., 2023). Equally, educational programs may use such datasets to teach future criminologists (Buil-Gil et al., 2025). In return, it might be expected that criminologists are equally eager to share primary data collected for empirical criminological research. Two recent self-report survey-studies indicate that 43 percent of criminologists and about 25 percent of surveyed terrorism researchers regularly share data (Chin et al., 2023). However, an empirical study on open science practices in five leading criminological journals concludes that the vast majority (82.8%) of recently published articles did not include a data availability statement at all, and only about 8 percent of studies stated that their data were publicly available (Greenspan et al., 2024). Focusing on studies that presented primary data or a mix of primary and secondary data, only 1 percent included a data statement. Thus, there seems to be a significant discrepancy between self-reported and observed open data practices.

Challenges: The personal and sensitive nature of criminological data

Some of the literature on open science in criminology or other related fields of study also provide reasons for why data sharing practices may not be adopted. Missing infrastructure, such as discipline-specific data archives, or the fear of 'scooping' or malpractice with open data are some common reasons named (Aleixandre-Benavent et al., 2020; Houtkoop et al., 2018; Schumann et al., 2019). Another reason that is of relevance to criminology is the private and often sensitive nature of the data, the use of which is strictly regulated by data privacy regulations, such as the European General Data Protection Regulation (GDPR) or additional national data protection laws, such as the German Federal Data Protection Act (*BDSG – Bundesdatenschutzgesetz*) or the Dutch Personal Data Protection Act (*Wet bescherming persoonsgegevens*). In general, such privacy protection regulations aim to ensure that personal and other sensitive data are handled in ways that respect individual's privacy rights. These restrictions apply to the collection, storage and sharing of data and are designed to limit data processing to what is necessary and to minimize the risks of misuse. For example, following GDPR regulations, data must be collected for legitimate purposes and on a legal basis and

only the minimum amount of data necessary to fulfill the intended purpose should be collected (European Union, 2016). Many national or organizational privacy regulations, such as the French Data Protection Act (*Loi Informatique et Libertés*) and the German Police Data Protection Act (PDSG – *Polizeiliches Datenschutzgesetz*), entail even stricter rules for the processing of particular types of data, such as crime data.

While these privacy regulations are intended to safeguard individuals from data misuse, they impose a challenge for academic research. One issue may be the restricted access to essential data sources. Criminologists rely on various datasets, including police records, judicial case files and administrative records. Yet, GDPR and national laws impose strict conditions on accessing such data, requiring researchers to demonstrate a legitimate interest or to obtain explicit consent from data subjects (European Union, 2016). Additionally, law enforcement agencies and judicial bodies may be reluctant to share data due to concerns about compliance with their own organizational privacy regulations, leading to inconsistencies in data availability across jurisdictions. Another obstacle relates to cross-border research collaborations: GDPR imposes strict conditions on the transfer of personal data outside the EU, which may impact feasibility of cross-national research projects and the widespread secondary reuse of data. Researchers working with institutions in countries without equivalent privacy protections face additional bureaucratic hurdles, reducing the potential for comparative criminological studies, which have been pivotal for the field (Bennett, 2009; Nivette, 2021).

One direct consequence of these challenges is the lack of open data practices in criminology, which, in turn, has serious implications for the field's scientific rigor, transparency and long-term development. When researchers withhold primary data, the field suffers from an overreliance on (often highly aggregated) secondary sources. Previous studies have already noted the necessity of individual- or otherwise disaggregated micro-level data, for example, to uncover underlying trends across homicide types not visible in more aggregated homicide rates (Skott, 2019), or to evaluate the effectiveness of crime prevention measures (Johnson et al., 2001). In addition, the absence of shared primary data undermines replication efforts (Pridemore et al., 2018). Following widely adopted definitions (Crüwell et al., 2019), reproducibility refers to the ability to obtain the same results using the original data and analytical procedures, whereas replications assess whether similar findings can be achieved using new data collected under comparable methodological conditions. Without access to raw, case-level data, it becomes challenging to scrutinize measurement decisions, access sampling biases, explore alternative interpretation, or verify results through reanalysis of the original data. Consequently, even well-documented methodologies cannot be fully reproduced, weaking the empirical foundation of the discipline – an especially serious concern for criminology, where findings inform policy and justice system practices. In addition, withholding primary data limits transparency and accountability more broadly. Stakeholders – including other academics, policymakers and the public – have limited means to assess the reliability and robustness of published results. A cultural and institutional shift toward open sharing of primary data is not just beneficial – it is imperative for the field.

Therefore, it comes to no surprise that researchers have adopted several methods to overcome obstacles and restrictions on data collection, storage and sharing. Table 1 provides a short overview of the most common data sharing practices, ranked by degree of openness of the data. One common privacy preserving method is the practice of anonymization or pseudonymization, which entails the deletion, perturbation or aggregation of identifiable attributes in the dataset, such as names or specific locations. Although fully anonymized data is not considered personal or sensitive and as such not protected by GDPR or other privacy regulations, achieving true anonymization in criminological research is highly challenging. Criminal justice data typically contains indirect identifiers (also called quasi-identifiers, e.g. age, location and offense type) that, when combined, can lead to re-identification of individuals. That is a particular problem for highly publicized and rare crimes, such as homicides. Such scenarios create a dilemma where researchers either risk violating privacy regulations or must significantly alter or aggregate datasets, potentially diminishing their utility for meaningful analysis. Another proposed solution is access control, meaning that the data owner may limit who receives access and in what form. For example, access to the Right-Wing Terrorism and Violence dataset by the Center for Research on Extremism in Oslo is controlled through an application form in which the applicant needs to 'demonstrate a reasonable need […], be associated with a publicly recognized research institution; and sign a written consent to use the data in accordance with GDPR rules and regulations' (C-REX: Center for Research on Extremism, 2025). While effective, access control is often inefficient, as assessing applications for data access can be very laborious. While the adoption of such measures showcases the willingness of researchers to share their data and contribute to the field, they do not fully address the underlying issue and create other ethical problems. In this study, we propose that synthetic data offers an alternative and potentially better approach to overcoming these common hurdles for data sharing in criminology, as synthetic data can be equally privacy-preserving while serving as a valid and useful proxy of real data. It must be mentioned, however, that the absolute suitability of any of these data sharing methods is highly dependent on dataset characteristics, disclosure risk and legal context. Consequently, no method can be considered intrinsically more 'open' than another across contexts.

Synthetic data

There is no agreed-upon definition of synthetic data. For the purpose of this study, we follow Gal and Lynskey (2023: 1090) in defining synthetic data as 'artificial data, generally generated by computer simulations or algorithms, which has analytical value'. Synthetic data can take on many different forms, just like real data. Most commonly, synthetic data takes the form of structured tabular data, such as health records (Walonoski et al., 2018) or longitudinal census data (Nowok et al., 2017). Yet, several studies showcase its use for images or textual data as well, including clinical notes (Melamud and Shivade, 2019) or deep fake imagery (Koetzier et al., 2024). Synthetic data can be created from scratch, meaning that artificial data is created based on a list of requirements, which may or may not mirror reality, based on general expert knowledge, or synthetic data can

Table 1. Common data-sharing approaches in criminological research.

Data examples	Crime rates by area; offence counts by year and type	National crime statistics; anonymized survey micro data	Synthetic incident records simulated offender trajectories	Police case files with unique IDs; offender cohort datasets	Linked police, court and correctional records	Police-university research data; cross-national justice datasets
Disadvantage	No individual-level reanalysis; limited reproducibility	Loss of analytical detail; difficult to achieve	Limited validity for substantive inference	Data remains personal; limits open dissemination	Access barriers; long-term commitment from institution	Restricts openness; administrative burden
Advantage	Low disclosure risk; compatible with open publication	Enables open access; supports transparency and reuse	Allows open sharing; minimal disclosure risk	Preserves analytical detail; allows for longitudinal analysis	Enables use of sensitive data; strong privacy protection	Enables controlled data reuse across institutions
Level of openness	Fully open (public reporting)	Fully open (open access, reuse)	Fully open (open dissemination)	Conditionally open (controlled access)	Conditionally open (approved users only)	Closed but auditable (partner-based)
Definition	Summarization of individual data into group-level statistics	Irreversible removing of identifiability	Artificial data reproducing statistical properties of real data	Replacement of identifiers with pseudonyms allowing re-identification	Access limited to approved users within secure environments	Legal contracts defining purposes, conditions and safeguards
Data-sharing approach	Aggregation	Anonymization	Synthetic Data	Pseudonymization	Restricted Access	Data Sharing Agreements

Note: Based on (OECD, 2023; Templ and Sariyar, 2022).

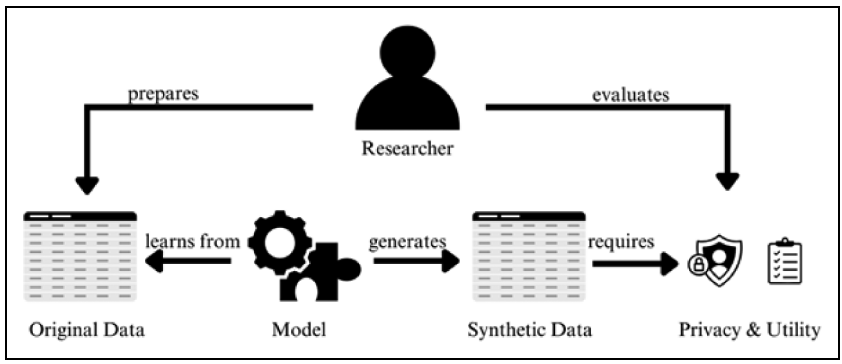


Figure 1. Visual depiction of data synthesis process.

be created with the intention of simulating a specific, already existing dataset. Finally, synthetic data can be fully or partially synthetic, with the latter mixing both real and synthetic data (Emam et al., 2020). Since the idea of synthetic data was proposed in the early 1990s (Rubin, 1993), it has been praised as a tool to address data scarcity or biases in data, but most importantly for this study, for its ability to preserve the analytical value as well as privacy of real data.

Synthetic data is generated using computational algorithms that model the behavior and characteristics of real data. This model could be statistical – like a regression or probabilistic model – or use machine learning techniques such as decision trees or neural networks. A vital distinction between the two is that statistical methods leverage strict assumptions on the data to be able to generate new data points, for example, which family of probability distributions the data originate from, whereas machine learning methods leverage large datasets to learn patterns and generate new data points automatically. Each data generation technique has its own advantages and disadvantages, and the choice of method often depends on the type of data required and the desired level of realism. Either way, the trained model is used to generate new data points that replicate the characteristics of the real data, while ensuring they do not correspond to any actual individual or event. Figure 1 depicts the synthesis process visually.

After generation, the synthetic data needs to be evaluated. Evaluating synthetic data involves assessing both its usefulness and its ability to protect privacy. Two key aspects of usefulness are fidelity – also known as general utility – and specific utility (Achterberg et al., 2025). Fidelity refers to how well the synthetic data preserves the overall structure and statistical properties of the original dataset. This might involve checking whether averages, distributions and relationships between variables are similar. Specific utility, on the other hand, measures how well synthetic data supports a particular analysis or task – for example, whether a model trained on synthetic data produces similar results to one trained on real data. Privacy is equally important in synthetic data evaluation. Privacy metrics are used to estimate the risk that synthetic data could accidentally reveal information about real individuals. Common privacy risks include membership

disclosure, where an attacker can guess whether a person was part of the original dataset, and attribute disclosure, where sensitive information about a person might be inferred (Kaabachi et al., 2025). Although synthetic data creates new data points, critical voices have warned that synthetic data is not inherently privacy-preserving (Stadler et al., 2022) – its legal status under GDPR depends on the quality of data generation method used as well as the residual risk that individuals could be re-identified from the synthetic dataset.

In the end, the goal is to balance utility and privacy – creating synthetic data that is both useful for research or analysis and safe to share without compromising anyone's personal information. Yet, researchers generating synthetic data find themselves confronted with the so-called utility-privacy trade-off, a central challenge in synthetic data generation (Emam et al., 2020). Increasing the utility of synthetic data often involves making it more similar to the real data, but this can raise the risk of leaking sensitive information. Conversely, reducing privacy risks by introducing more randomness or noise can lower the data's utility, making it less reliable for research, training models, or running simulations. Differential privacy has emerged as a prominent mechanism to control this trade-off in synthetic data generation through user-specified parameters (Dwork, 2006; Zhang and Kifer, 2017). A debate about the right balance between privacy and utility among researchers, policymakers and legal experts is ongoing and continues, as new methods, such as privacy-preserving machine learning and more nuanced evaluation metrics, seek to optimize both goals (Emam et al., 2020). Ultimately, choosing the right balance depends on the context, the acceptable level of risk and the intended use of the synthetic data.

Regardless of this challenge, the advancements in the synthesis of data have been followed by a wider application of synthetic data in a variety of research fields dealing with personal data, including behavioral educational science (Flanagan et al., 2022) or finance (Assefa et al., 2021). Another field of study in which synthetic data application has gained popularity is public health. Here, synthetic data is used to address data scarcity or to simulate rare scenarios, such as a pandemic, to gain insights and anticipate policy implications (Giuffrè and Shung, 2023). Most relevant to this study though, it is also used as a proxy for confidential patient data. Qian et al. (2024), for example, use sensitive data from the UK Biobank on smokers and demonstrate that the synthetic dataset can be used for prognostic modeling even without access to the original data. In 2020, the US National Center for Health Statistics used synthetic data to substitute sensitive variables in their public use mortality data, such as cause of death (National Center for Health Statistics, 2022). Overall, public health researchers seem to support the use of synthetic data in healthcare, although certain limitations, such as the need for privacy checks and variability in quality of the data, are noted (Appenzeller et al., 2022; Giuffrè and Shung, 2023).

For criminologists, the idea of using computer-generated data is not new either: in a special issue of the *Journal of Experimental Criminology* from 2008, several studies exemplify how data generated from theoretical or computational models can mimic how crime or criminal behavior unfolds under certain conditions (Groff and Mazerolle, 2008). Recently, a series of studies on crime data quality in the United Kingdom generated individual-level synthetic population and crime data based on open-source surveys

and census data (Brunton-Smith et al., 2024; Buil-Gil et al., 2022a, 2022b). With these synthetic datasets, the researchers were able to highlight biases in police data, evaluate the validity of relevant other crime data sources and estimate dark numbers of crime. Other examples include the use of a synthetic dataset to link administrative and survey data on violence (Barbosa et al., 2025), or an ongoing project aimed at creating realistic money laundering synthetic data (The Alan Turing Institute, 2024). Thus, past studies show how simulated data can be used to test theories, explore what-if scenarios, or even evaluate interventions *before* implementing them in the real world (Groff and Mazerolle, 2008). What criminologists seemed to have overlooked so far, though, is the application of computer-generated data to enhance data sharing practices, whereby data is generated as a direct proxy of primary data collected in criminological studies. One specific problem that could be addressed with synthetic data is the lack of reproducibility and replication in the field (Pridemore et al., 2018). Although synthetic data is usually approximate, not exact, it works as a shareable stand-in for restricted data and thereby enables code and workflow verification. Synthetic data can also help to expose data preprocessing decisions, variable constructions and modeling choices that would otherwise remain opaque. As synthetic data cannot replace real data for true empirical replications, its contribution to replication is more indirect, in that it can clarify data-generating processes or support method-focused replication. These potentials of synthetic data as well as relatively recent developments in this field warrant a closer look at the possibilities that synthetic data generation has to offer in fight for better open data and open science practices. The following section exemplifies the use of synthetic data for that purpose, using the individual-level, detailed Dutch Homicide Monitor as a proof-of-concept.

Criminological use case: The synthetic Dutch Homicide Monitor

Data: The Dutch Homicide Monitor

In this study, we use the Dutch Homicide Monitor (DHM) to showcase both abilities and shortcomings of synthetic data. The content of the DHM has been described in detail elsewhere (Krüsselmann et al., 2023). In short, the DHM is a dataset that collects detailed information (with over 85 variables) on homicide cases committed in the Netherlands since 1992. Its nucleus version contains information on 25 characteristics of the homicide case and the involved victim(s) and perpetrator(s). Insights the DHM have been used in a variety of academic studies (Krüsselmann et al., 2024a; Liem et al., 2019) or policy-relevant outputs (Aarten et al., 2020), and served in understanding patterns, risk factors and social dynamics influencing homicide rates and trends in the Netherlands. However, original, disaggregated DHM data has not been published in any form, due to the sensitivity of the information included. Although some data sources of the DHM are public, other information is gathered from nonpublic sources, in particular police data, public prosecution files and forensic reports. The personal and sensitive data included in the

Table 2. Example of Dutch Homicide Monitor dataset structure.

Case number	Serial number	Number of victims	Type	Gender
1	1.1	1	Victim	Female
1	1.2	1	Perpetrator	Male
2	2.1	2	Victim	Male
2	2.2	2	Victim	Male
2	2.3	2	Perpetrator	Male

DHM is protected by GDPR, as well as additional regulations set out by the Dutch National Police and the public prosecution office in the Netherlands.

The DHM is a structured tabular dataset of 22 categorical variables with categories ranging between two and 36 (e.g. victim–perpetrator relationship) and three numerical variables (age, number of victims, and number of perpetrators). Each record in the dataset represents an individual involved as either homicide victim or (suspected) perpetrator. Rows, thus individuals, involved in the same case are linked through a variable assigning a case identification number (see Table 2). Some of the variables, such as the number of victims, are recorded on case-level, meaning that the information is the same across all records that belong to the same homicide case. Other variables are linked to the individual, such as age or gender, and thus differ across the records.

Goals of synthesis

The overarching goal for this synthesis is the creation of a synthetic homicide dataset, which can be published and used for various research purposes. Before preparing the data and deciding on a model and tool for the synthesis, we defined and ranked a number of requirements that the synthetic dataset should fulfill. The synthetic dataset should be (in order of priority):

1. sufficiently high in fidelity, in that the probability distributions in the synthetic dataset closely resemble the ones from the original dataset;
2. sufficiently high in privacy, in that it safeguards any private and/or sensitive information from the original dataset and
3. sufficiently high in specific utility, in that the synthetic data can be used for replication analyses of studies conducted with the original dataset.

Data preparation/preprocessing

Next to general data cleaning processes (e.g. checking for coding errors or erasing duplicates), a number of steps have been taken to provide the best conditions for data synthesis. First, we selected a sample out of the total 5563 homicide cases committed between 1992 and 2023 included in the DHM at the time of data synthesis. For the purpose of this

proof-of-concept synthetic dataset, we selected a sample of 10 years of homicide data with high quality and the lowest fraction of random missing data.[1] The to-be-synthesized dataset then consisted of 1271 cases and included information about 3059 individuals. Second, we identified two logical dependencies in the original data that needed to be represented (and thus manually programed) in the synthetic dataset:

1. in cases classified as child homicides, the associated victim needed to be 17 years or younger and
2. in homicides classified as intimate partner homicides, the victim–perpetrator relationship needed to reflect a current or former intimate relationship (e.g. (ex-)wife, boyfriend etc.).

Third, we adapted the structure of the dataset before synthesis. Before arriving at the most high-quality synthetic dataset which is presented in the results section, several attempts were made to create a synthetic version that would keep the dataset's original structure, in which (1) the details of each variable is carried over and (2) the dependencies across rows that are connected in the same homicide case would be fully reflected in the synthetic dataset. However, the fidelity of the data was low.[2] Therefore, steps were taken to simplify the structure of the original dataset, by (1) eliminating variables with a large ratio of missings, by (2) merging case-, victim- and perpetrator information onto one row and by (3) merging some categories of variables together. For example, whereas the original variable about the victim–perpetrator relationship contained 36 separate categories, after the merge, we retained nine categories that still reflect most of the original categories.

Model: How does synthesis work?

To select the most suitable synthetic data generating method, we draw on the requirements put forth in the previous section and the general properties of the variety of available synthesization methods. Rule-based methods are time-consuming to set-up and rely mostly on previous domain models. Machine learning methods require vast amounts of training data and are generally not interpretable. The latter causes privacy guarantees from their output to be ambiguous, that is, they are prone to issues such as overfitting which incur various privacy risks (Breugel et al., 2023). Statistical methods strike a balance between the two as they are fairly easy to set up, are more interpretable, and therefore less ambiguous in their privacy guarantees. We conclude that statistical methods fit the structure and goals of the DHM synthesis the best (Krüsselmann et al., 2024b). More specifically, we opt to employ copula models, which generate synthetic data by capturing and simulating the dependency structure between variables separately from their marginal distributions, allowing for flexible and realistic data generation (Patki et al., 2016). They achieve this by fitting a copula function to the empirical dependence structure of real data and then sampling from it to create new data points with the same joint behavior. This process of synthesizing data allowed us to simulate the dependencies of the original

data, to preserve important statistical properties, and to generate a realistic yet privacy-preserving synthetic dataset.

There is a growing variety of open-source and commercial tools for data synthesis (statice, n.d.). The choice of a data synthesis tool is necessarily context-dependent and should be guided by the specific goals of the analysis, the structure and sensitivity of the data, and the intended use of the synthetic outputs. Different tools vary in their ability to preserve statistical properties, handle longitudinal or spatial structure, and limit disclosure risk. Consequently, no single approach can be considered optimal across settings, and tool selection involves trade-offs between utility, privacy and interpretability. We chose the Synthetic Data Vault (SDV) (Patki et al., 2016) due to its ability to create synthetic datasets with complex dependencies across variables, which provided the best fit for our multi-level (case, victim and perpetrator) homicide dataset. SDV implements this through its Hierarchical Modeling Algorithm, which applies a Gaussian Copula recursively across linked tables: it models each table's variable distributions and inter-variable correlations, then aggregates child-table statistics into the parent table before sampling, thereby preserving dependencies both within and across the case, victim and perpetrator levels. This multitable capability is rare among open-source synthesis tools, and was a decisive factor in our selection, as single-table tools would have required flattening the dataset and consequently losing the relational structure that is central to homicide research. In addition, SDV provides extensive support and documentation, and accessible reports that allow a comparison of univariate distributions of the original and synthetic datasets, as well as a comparison of correlations across both datasets.

Results

This section evaluates and showcases the quality of the synthetic dataset based on four factors: (1) univariate distributions, (2) bivariate correlations, (3) utility of data for research purposes and (4) privacy safeguarding.

Univariate distributions

We assess similarity between univariate distributions of synthetic and real data through SDV's TVComplement score. This score relies on the Total Variation Distance (TVD) between real and synthetic variables, which measures discrepancies in category frequencies. More specifically, the score is defined as 1-TVD, such that higher score indicates higher similarity.

Overall, the synthetic dataset had about 98 percent similarity in univariate distributions when compared to the original dataset. The similarity between each of these variables ranged between 99.9 percent (variable: Victim or perpetrator) and 93.5 percent (variable: Context of homicide). The univariate distributions for all included variables are presented in Table 3.

Figure 2 displays three visual distribution plots of exemplary comparisons of the univariate distributions of the original and synthetic dataset. In Figure 2, a variable with one of the highest, an average and the lowest similarity score are presented.

Table 3. Univariate distribution comparison between original and synthetic dataset, ranked from highest to lowest similarity.

Variable	Similarity score (%)
Type: victim or perpetrator	99.93
Urbanization of crimescene	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 country of birth	98.30
Type of crimescene	98.04
Victim gender	97.90
Victim age	97.38
Perpetrator age	97.32
Modus operandi	97.11
Victim–perpetrator relationship	96.95
Victim country of birth	96.89
Year homicide was committed	95.20
Region homicide was committed	94.92
Context of homicide	93.53
Average	97.68

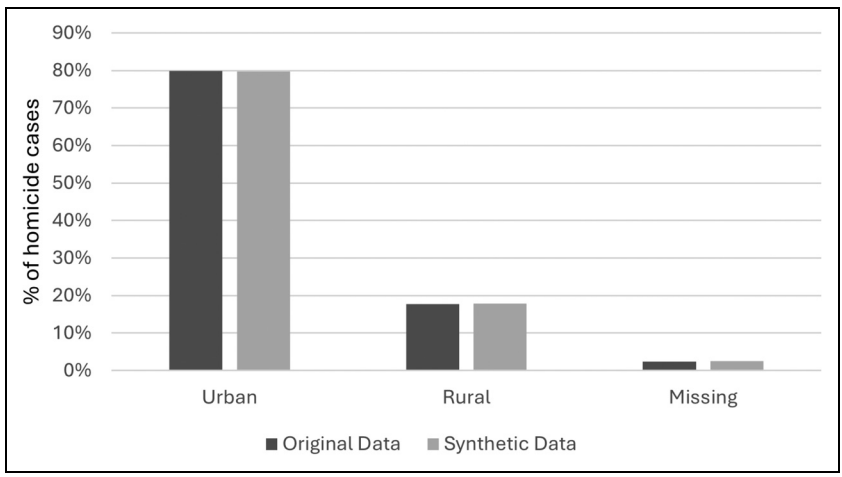


Figure 2. Comparisons of the univariate distributions of the original and synthetic dataset.

(continued)

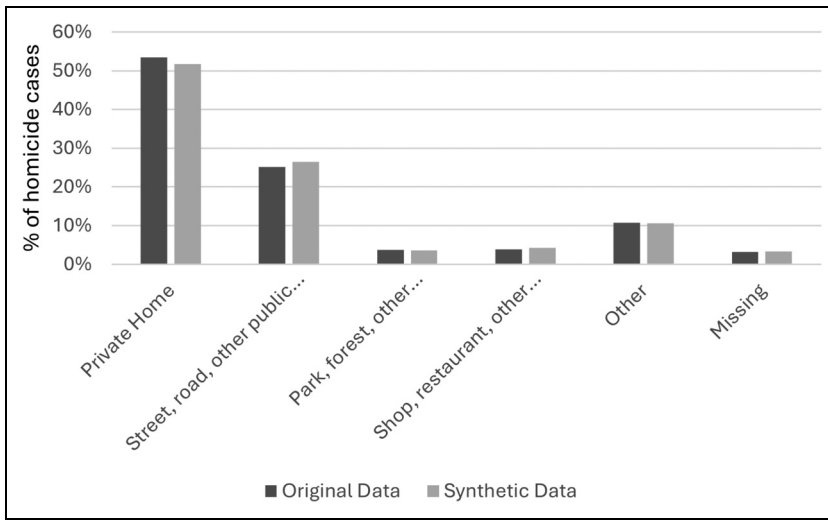


Figure 2. Continued.

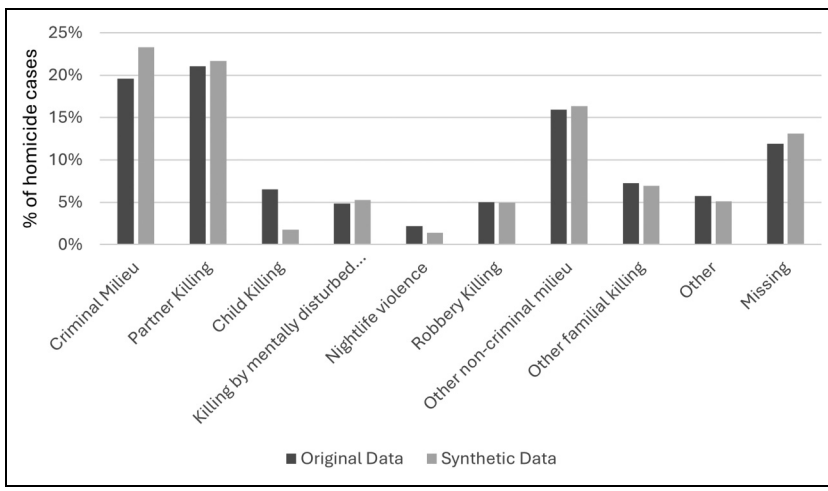


Figure 2. Continued.

These distributions display the differences between the original and synthetic dataset. Slight differences to the original data can be observed, yet those differences are slim and mostly do not diverge from the trends within each of these variables. For example, in both the original and synthetic dataset, about 80 percent of homicides occurred in urban areas. The variable about the homicide context had the lowest similarity score between original and synthetic dataset, which is reflected in several differences across the two datasets.

Table 4. Probability score of urbanization of crime scene and perpetrator country of birth of original and synthetic dataset, in percent.

	Urban		Rural		Missing	
	Original	Synthetic	Original	Synthetic	Original	Synthetic
Netherlands	25.2	25.4	6.8	7.9	1.0	1.3
Other European	7.2	7.5	1.6	1.1	0.3	0.0
Other non-European	12.8	10.4	1.6	2.7	0.3	0.3
Missing	34.7	36.4	7.6	7.0	0.9	0.9

While the original dataset shows that the most significant fraction of homicides occurred in the context of partner killings (21% in original and 22% in the synthetic dataset), in the synthetic dataset, homicides in the criminal milieu made up the largest fraction of homicides (20% in original and 23% in synthetic). Although these differences could impact the fidelity of the synthetic data, the overarching similarity score of this particular variable (93.5%), as well as the synthetic dataset overall (97.8%) are still high, thus indicating a high fidelity based on these univariate distributions

Bivariate correlations

With regard to bivariate correlations, for categorical variables, this corresponds to contingency similarity. To this end, we compute the TVD between contingency tables of a synthetic and real variable. The score is, again, computed as 1-TVD. Averaged over all categorical variables, we find a bivariate similarity score of 89.7 percent, meaning that the vast majority of correlations have been replicated well in the synthetic dataset. Table 4 provides an example of the comparison between the bivariate distribution of variables about the country of birth of the perpetrator and the urbanization of the crime scene, for both the original and synthetic data.

Comparing both figures, slight, but not relevant differences between both original and synthetic data are visible, which in return indicates a high utility of the synthetic data.

Data analysis with synthetic data

To test the analysis-specific utility of the synthetic data, we ran an analysis previously applied to the original data. In particular, we aimed to replicate a previous study on firearm homicides (Krüsselmann et al., 2023). This study found that, in the Netherlands, firearm homicides are significantly more likely to be associated with public crime scenes, urban areas, multiple victims, multiple perpetrators, homicides in the criminal milieu and an unsolved status than nonfirearm homicides. These results were based on chi-square tests.

Initially, we used the synthetic dataset presented above to run the same analysis. Unfortunately, the results show that none of the significant findings from the original study could be replicated (see Table 5).

Table 5. First replication attempt; case characteristics of firearm and nonfirearm homicides in the Netherlands (2001–2016 original dataset); percent of known cases.

	Original study findings		Replication with synthetic dataset	
	Firearm	Nonfirearm	Firearm	Nonfirearm
Location				
Public	66.9	33.8	49.2	37.0
Private	33.1	66.2	50.8	63.0
Significance	***			
Area				
Urban	81.2	69.5	83.1	78.9
Rural	18.8	30.5	16.9	21.1
Significance	***			
Number of victims				
Single	92.9	95.7	90.6	90.0
Multiple	7.1	4.3	9.4	10.0
Significance	**			
Number of perpetrators				
Single	43.5	77.1	77.8	78.6
Multiple	56.5	22.9	22.2	21.4
Significance	***			
Type homicide				
Intimate partner homicide	13.2	27.3	26.5	23.4
Other domestic	6.4	19.7	14.6	15.6
Criminal milieu	43.9	4.5	20.9	21.1
Robbery	8.9	8.6	8.7	4.4
Nightlife	3.8	2.2	1.9	2.6
Sexual	0.3	3.2	1.1	1.3
Dispute	23.5	34.5	26.3	31.6
Significance	***			
Solved				
Solved	75.5	95.4	85.9	14.1
Not solved	24.5	4.6	91.4	8.6
Significance	***		*	

Note: *$p < .05$, **$p < .01$, ***$p < .001$.

We assume that this finding has several reasons: first, the coding of some variables from the original study does not match with the coding of the synthetic dataset. For example, in the original study on firearm homicides, the homicide context was presented in seven categories (intimate partner homicide, other domestic homicide, criminal milieu, robbery, nightlife homicide, sexual homicide and dispute homicide), whereas the synthetic dataset contained nine similar, but not identical, separate categories. Second, the original firearm homicide study contains more and different homicide cases than the synthetic dataset, as we selected a specific sample for the synthesis to increase quality of data. Third, and possibly most importantly, the synthetic dataset was not synthesized for the

purpose of being used for this study. We aimed to generate a synthetic homicide dataset for the general purpose of informing about trends observed. Therefore, we included as many variables as possible with as much detail as possible. Including the chosen level of detail and number of variables may have lowered the quality of the synthetic dataset for specific purposes, as the model had to replicate many uni-, bi- and multivariate relations.

Hence, we assume that a synthetic dataset coded and synthetized for the purpose of replicating the original firearm homicide study would perform better. To test this assumption, we created a synthetic dataset that matched the data from the original firearm homicide study in the number of cases selected from the same years and the selected variables, as coded in the original study. When the same chi-square analyses are performed, we indeed see an improvement, as four out of the six variables show similar significant findings as in the original study. In the analysis with the synthetic data, firearm homicides are significantly more likely to take place in public locations, to remain unsolved, to involve multiple perpetrators and to take place in the context of the criminal milieu. Only the number of victims ($p = .75$) and urbanity of the crime scene ($p = .87$) do not have the expected significance (see Table 6). In short, we can conclude that the dataset synthesized for the specific purpose of replicating the original study performs better than the general-purpose synthetic dataset. Yet, even the former cannot fully replicate all relevant findings from the original study. For maximum use-case specific utility, a generation method that focuses on utility on that specific downstream task, and less focus on general fidelity might be a better approach (Achterberg et al., 2025).

Privacy safeguarding

Finally, we tested the synthetic dataset on privacy measures. SDV – the program we used for the synthesis of our data – has a built-in privacy metric, the disclosure protection metric, which addresses the attribute disclosure risk (DataCebo, Inc., 2023). This metric calculates the risk that real sensitive information can be shared through the synthetic dataset. Specifically, it simulates a situation in which an entity already knows some of the real data (i.e. quasi-identifiers) but is trying to infer knowledge on sensitive attributes they don't have access to through the synthetic dataset. For example, in the context of the DHM, the metric will be able to answer the question 'How high is the risk that someone will learn about the context of the homicide through the synthetic dataset, if they already know the modus operandi and gender of the victim from the real dataset?'.

To measure the data protection metric for our synthetic dataset, we tested all combinations of variables. The results from the protection metric ranged between 6 and 56 percent, with an average of 13 percent. These percentages alone carry little meaning, given the lack of standards on classifications on what constitutes low, high or generally acceptable risks in the field of criminology, or for synthetic data more broadly. Instead, these metrics must be interpreted in the context of the underlying data. Homicides are rare, yet serious violent events, which usually attract public attention. In the Netherlands, an average of 0.7 homicides per 100 000 citizens, or about 110 cases occur annually, which are accompanied by extensive media reporting. These media reports,

Table 6. Second replication attempt; case characteristics of firearm and nonfirearm homicides in the Netherlands (2001–2016); percent of known cases.

	Original study findings		Replication with synthetic dataset	
	Firearm	Nonfirearm	Firearm	Nonfirearm
Location				
Public	66.9	33.8	48.7	41.0
Private	33.1	66.2	51.3	59.0
Significance	***			
Area				
Urban	81.2	69.5	75.3	73.3
Rural	18.8	30.5	24.7	26.7
Significance	***		***	
Number of victims				
Single	92.9	95.7	93.9	94.3
Multiple	7.1	4.3	6.1	5.7
Significance	**			
Number of perpetrators				
Single	43.5	77.1	42.4	70.1
Multiple	56.5	22.9	57.6	29.9
Significance	***		***	
Type homicide				
Intimate partner homicide	13.2	27.3	25.5	21.6
Other domestic	6.4	19.7	11.3	18.7
Criminal milieu	43.9	4.5	26.4	12.1
Robbery	8.9	8.6	5.7	11.0
Nightlife	3.8	2.2	1.2	3.5
Sexual	0.3	3.2	1.7	2.5
Dispute	23.5	34.5	28.1	30.9
Significance	***		***	
Solved				
Solved	75.5	95.4	84.9	88.4
Not solved	24.5	4.6	15.1	11.6
Significance	***		**	

Note: *$p < .05$, **$p < .01$, ***$p < .001$.

such as the annual overview of homicide cases by the national news magazine EW, usually contain information on victims, perpetrators and the contexts of the homicides. Thus, when evaluating the privacy risks, we evaluate whether the information at risk is both sensitive, and not already considered public knowledge. For example, the highest privacy risk of 56 percent is registered for information on the perpetrator's gender, when the victim's gender and modus operandi is already known. Given our domain knowledge, we consider 56 percent an acceptable risk, given that the perpetrator's gender is commonly reported on in news or other public sources, such as court records. Overall, we concluded

that the risks of our synthetic homicide dataset are either very low, or otherwise acceptable.

In addition to the disclosure protection metric, we manually checked whether we could identify a random sample of 20 cases as well as outliers from the original dataset in the synthetic dataset. Outliers in the original dataset are rare homicide cases, based on characteristics, such as a very high or young age of the victim or offender, or a combination of characteristics, such as sexual homicides involving minors. We defined rare homicides as occurring in less than 10 percent of homicide cases included for the synthesis. Based on the (combination of) rare attribute(s), we attempted to identify the same or similar cases in the synthetic dataset. For the 73 rare homicide cases identified in the original dataset, no similar cases could be found in the synthetic dataset. We also could not identify cases with a high similarity for the random sample of homicide cases.

Based on these factors, we concluded that the synthetic dataset was fulfilling the privacy requirements set out at the start of the synthesis. Yet, it must be noted here that this evaluation is based on our own estimation of high privacy and high utility, an issue further discussed in the following discussion on the advantages and shortcomings of synthetic data for data sharing in the field of criminology.

As a result of the satisfactory fidelity and privacy evaluation, we published the synthetic homicide dataset for public reuse on an open repository (Krüsselmann et al., 2024c), where users can download the dataset and conduct their own analyses. In addition, we created a public website that includes a user-friendly dashboard in which users, such as journalists or policymakers, can visualize specific homicide trends without having to analyze any data themselves (www.dutchhomicide.streamlit.app). The underlying data of this dashboard is the synthetic homicide dataset presented in this study.

Discussion

Open data practices enable scientific progress and increase credibility in research. Yet, the sharing of personal and sensitive data used in criminology and other fields is heavily restricted due to ethical concerns and privacy regulations. This study argues that synthetic data may offer an opportunity to overcome common hurdles to data sharing and thus move the research fields toward improved open data practices, better transparency and, as a result, high credibility. The proof-of-concept of a synthetic homicide dataset highlights several promises of synthetic data for enhanced data sharing practices in the field, as well as some shortcomings that should be considered.

We showed that synthetic data can have a high general and specific utility and therefore be used as a valid proxy for nonpublicly accessible data, even fine-grained, individual-level data. The synthetic dataset we generated had a 98 percent similarity score (based on TVD) with the original data based on univariate distributions, but it also mirrored closely bivariate distributions and internal dependencies in the original data. With such a high general utility, we can be confident that the synthetic homicide dataset can be used for a variety of purposes, such as countering myths about homicide amongst the public, or be used as an empirical basis for relevant policies or as educational material for criminology students. To that end, we publicly released the dataset, both as an

individual-level dataset that can be used for simple analytical purposes, as well as an accessible dashboard that provides accessible uni- and bivariate graphics for all included variables (Krüsselmann, Achterberg, et al., 2024b, www.dutchhomicide.streamlit.app). In addition, we have shown that it is possible to generate synthetic datasets with a high specific utility, as we were able to generate a synthetic version of the DHM that allowed for a replication of a previous study. In our study, not all findings from the original study were fully reproduced; other studies, however, have demonstrated that synthetic data can be generated with a high specific utility for a variety of purposes (Achterberg et al., 2025; Hittmeir et al., 2019). In addition, our study focused on univariate, bivariate and significance testing, but excluded the synthesis of temporal and geographical data in order to decrease potential risks. However, other studies have already demonstrated the use of synthetic data for longitudinal (Dennett et al., 2016; Miletic and Sariyar, 2025) and geographic analyses (Quick and Waller, 2018), and future studies should assess these types of synthetic data as well. These abilities of synthetic data would address several problems facing criminologists, such as the lack of reproducibility in the field. Pridemore (2011) and Pridemore et al. (2018) caution that reproducibility and replication studies are integral to the trustworthiness of criminological research, but that issues such as data availability hinder structural efforts. Another issue that could be addressed with the use of synthetic data is the challenge of international data transfers, which hinders comparative studies (Bennett, 2009; Nivette, 2021). Currently, cross-border data transfers can be a hurdle due to differences in national privacy laws or data safeguarding requirements, between EU member states and beyond. For example, in the case of the European Homicide Monitor – a framework for collecting detailed data on homicide cases which is used by criminologists across several European countries – researchers would have to physically come together to merge their respective national homicide datasets to conduct joined analyses. Such efforts are rarely feasible, be it for financial or practical reasons. As a result, comparative research had to remain relatively superficial, with each country team conducting their own analyses before comparing results (see, e.g. Krüsselmann et al., 2023; Liem et al., 2019). With the use of synthetic data, researchers would be able to join the respective synthetic versions of their datasets into a larger European dataset, enabling new possibilities for research and insights into the phenomenon of homicide.

However, our proof-of-concept also revealed that general and specific utility may not occur at the same time. While the initial synthetic dataset had a high general utility in that it simulated general distributions and correlations from the original dataset well, it did not perform well in the replication of an analysis previously run on the original dataset. This issue has previously been noted and described (Achterberg et al., 2025): when a synthetic dataset is generated for the purpose of being used for several different tasks, the generating model needs to replicate as many variables as closely as possible, which may lead to a high decline of quality of a small number of variables or a limited decline of quality across all variables, making the dataset less useful for very specific tasks. When synthetic datasets are created for specific purposes, the generating model can concentrate the fitting process on the most relevant variables and data points. In practice and as a result, multiple synthetic versions may be generated from the same original dataset,

each with distinct characteristics and degrees of utility depending on their intended use. This proliferation of versions poses a dilemma for data management and research transparency. Without a robust infrastructure to track, store and differentiate these datasets, the research community risks descending into a chaotic data landscape. For (European) criminologists, there are currently no centralized data repositories – hosted by journals, research collectives or other institutions – that would fulfill these needs. In addition, the absence of standardized documentation practices exacerbates this issue, as researchers may fail to adequately record the synthetic data's provenance, intended use or limitations. This not only hampers reproducibility but also undermines trust in research findings. To fully harness the potential of synthetic data in criminology, institutions must prioritize clear governance frameworks, version control mechanisms and research training in data stewardship. These measures are essential to prevent confusion, support ethical data use and maintain scientific rigor in a field where data sensitivity and accuracy are paramount.

In addition, we showcased that this detailed synthetic dataset based on sensitive personal data of homicide victims and perpetrators can be shared whiet adhering to privacy protection regulations and institutional ethical standards. The ability of artificially generating data that adheres to regulations and other standards opens new avenues for data sharing practices for criminological research. Yet, there is still an ongoing legal debate surrounding the ambiguity of privacy risks linked to synthetic data, which centers around two opposing approaches to privacy protection: the zero-risk approach and the acceptable risk approach (López and Elbi, 2022). The zero-risk approach emphasizes complete elimination of any possibility of re-identifying individuals, advocating for data transformations that render the risk of privacy breaches effectively nonexistent. This view tends to favor highly abstracted or heavily perturbed data, which may compromise utility in favor of privacy. In contrast, the acceptable-risk approach acknowledges that absolute privacy is often unattainable and instead focuses on minimizing risk to a legally and ethically acceptable threshold. This perspective is embedded in many data protection frameworks, such as the GDPR, which permit data processing and sharing when appropriate safeguards and risk assessments are in place.

Synthetic data generation sits at the intersection of these views. Supporters argue – mostly from an acceptable-risk point of view – that synthetic data, when properly evaluated and documented, can offer a practical balance between privacy and utility (Appenzeller et al., 2022; Emam et al., 2020). Yet, while it can significantly reduce re-identification risks, it rarely guarantees zero risk (Stadler et al., 2022). Therefore, it is important to critically review data sharing practices and work on shared definitions and standards. Given the unique sensitivities of criminological data, it is crucial to continue developing discipline-specific frameworks that address ethical standards, privacy risks and data utility. Such frameworks must be informed by ongoing collaborative dialogue among legal scholars, criminologists, data stewards and privacy officers, which ensures that privacy protections are not only legally robust but also contextually appropriate, aligning with the specific needs and ethical imperatives of criminological research. This dialogue should be an ongoing endeavor, as advancements in generative models and improved evaluation metrics offer new opportunities to broaden the use and application

of synthetic data in the field. At the same time, however, this growth is met with increasing scrutiny and regulation, such as the recent European Union AI Act, which introduces stricter guidelines for the use of AI-generated data (European Union, 2024). In the end, synthetic data represents a promising advancement for criminology, offering new opportunities for privacy-preserving analysis, broader data access, replication studies and more. However, while it is part of the solution to many of the ethical and legal conundrums faced by criminologists working with sensitive and personal data, it is not the entire answer, and its broader adoption must be guided by clear understanding of its limitations and ethical implications.

ORCID iDs

Katharina Krüsselmann https://orcid.org/0000-0001-6181-425X
Jim Achterberg https://orcid.org/0009-0000-9589-7831
Marcel Haas https://orcid.org/0000-0003-2581-8370

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (grant number Open Science Fund (OSF23.1.006)).

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data and code availability statement

All code used for this study is available from a Github repository (github.com/KKrusselmann/SENSYN) The original Dutch Homicide Monitor is not publicly available due to privacy regulations. The synthetic dataset used in this study is available on Zenodo (zenodo.org/records/13378063).

Notes

1. To decrease the re-identification risk of this study, we opt to not disclose the years selected for the sample. This should have no further consequences on the results.
2. More detailed descriptions of the lower quality synthetic datasets produced in this study can be found in Krüsselmann et al. (2024a).

References

Aarten P, Boelma Robertus C, Alink L, et al. (2020) *Wanneer blaffende honden bijten*. Politie en Wetenschap 115. The Hague: Leiden University. Available at: https://www.politieenwetenschap.nl/publicatie/politiewetenschap/2020/wanneer-blaffende-honden-bijten-344 (accessed 11 May 2025).

Achterberg J, Haas M, van Dijk B, et al. (2025) Fidelity-agnostic synthetic data generation improves utility while retaining privacy. *Patterns* 6(0): 101287.

Aebi MF and Linde A (2010) Is there a crime drop in Western Europe? *European Journal on Criminal Policy and Research* 16(4): 251–277.

The Alan Turing Institute (2024) Synthetic data for anti-money laundering. Available at: https://www.turing.ac.uk/research/research-projects/synthetic-data-anti-money-laundering (accessed 3 February 2026).

Aleixandre-Benavent R, Vidal-Infer A, Alonso-Arroyo A, et al. (2020) Research data sharing in Spain: Exploring determinants, practices, and perceptions. *Data* 5(2): 29.

Appenzeller A, Leitner M, Philipp P, et al. (2022) Privacy and utility of private synthetic data for medical data analyses. *Applied Sciences* 12(23): 12320.

Ashby MPJ (2021) The open-access availability of criminological research to practitioners and policy makers. *Journal of Criminal Justice Education* 32(1): 1–21.

Assefa SA, Dervovic D, Mahfouz M, et al. (2021) Generating synthetic data in finance: Opportunities, challenges and pitfalls. In: Proceedings of the First ACM International Conference on AI in Finance, New York, NY, USA, 7 October 2021, pp.1–8: Association for Computing Machinery.

Barbosa EC, Blom N and Bunce A (2025) Look-alike modelling in violence-related research: A missing data approach. *PLoS ONE* 20(1): e0301155.

Bennett RR (2009) Comparative criminological and criminal justice research and the data that drive them. *International Journal of Comparative and Applied Criminal Justice* 33(2): 171–192.

Borgman CL (2012) The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology* 63(6): 1059–1078.

Bowie NG (2021) 40 Terrorism databases and data sets: A new inventory. *Perspectives on Terrorism* 15(2): 147–161.

Breugel BV, Qian Z and Schaar MVD (2023) Synthetic data, real errors: How (not) to publish and use synthetic data. In: Proceedings of the 40th international conference on machine learning, 3 July 2023, pp.34793–34808: PMLR.

Brunton-Smith I, Buil-Gil D, Pina-Sánchez J, et al. (2024) Using synthetic crime data to understand patterns of police under-counting at the local level. In: Huey L and Buil-Gil D (eds) *The Crime Data Handbook*. Bristol University Press, 60–79.

Buil-Gil D, Brunton-Smith I, Pina-Sánchez J, et al. (2022a) Comparing measurements of violent crime in local communities: A case study in Islington, London. *Police Practice and Research* 23(4): 489–506.

Buil-Gil D, Bui L, Trajtenberg N, et al. (2025) Diversifying crime datasets in introductory statistical courses in criminology. *Journal of Criminal Justice Education* 36(1): 19–45.

Buil-Gil D, Moretti A and Langton SH (2022b) The accuracy of crime statistics: Assessing the impact of police data bias on geographic crime analysis. *Journal of Experimental Criminology* 18(3): 515–541.

Chermak SM, Freilich JD, Greene-Colozzi E, et al. (2025) Open-Source research in criminology and criminal justice. *Annual Review of Criminology* 8: 141–170.

Chin JM, Pickett JT, Vazire S, et al. (2023) Questionable research practices and open science in quantitative criminology. *Journal of Quantitative Criminology* 39(1): 21–51.

C-REX: Center for Research on Extremism (2025) RTV Dataset. Available at: https://www.sv.uio.no/c-rex/english/groups/rtv-dataset/access-to-the-data-set.html (accessed 2 May 2025).

Crüwell S, Van Doorn J, Etz A, et al. (2019) Seven easy steps to open science: An annotated Reading list. *Zeitschrift für Psychologie* 227(4): 237–248.

DataCebo, Inc. (2023) Disclosure protection. Available at: https://docs.sdv.dev/sdmetrics/metrics/privacy-metrics/disclosureprotection (accessed 11 May 2025).

De Bondt W (2014) Evidence based EU criminal policy making: In search of matching data. *European Journal on Criminal Policy and Research* 20(1): 23–49.

Dennett A, Norman P, Shelton N, et al. (2016) A synthetic longitudinal study dataset for England and Wales. *Data in Brief* 9: 85–89.

Dwork C (2006) Differential privacy. In: *Proceedings of the International Colloquium on Automata, Languages, and Programming. ICALP, 2006*.

Emam KE, Mosquera L and Hoptroff R (2020) *Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data*. O'Reilly Media, Inc.

European Commission (2024) *Horizon Europe Programme Guide*. European Commission.

European Commission, Directorate-General for Research & Innovation (2016) *H2020 Programme: Guidelines on FAIR Data Management in Horizon 2020*. Version 3.0. Luxembourg: European Commission, Directorate-General for Research & Innovation.

European Union (2016) General Data Protection Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data. *Official Journal of the European Union* L 119: 1–88.

European Union (2024) Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending certain Union legislative acts (Artificial Intelligence Act). *Official Journal of the European Union* L 212.

Flanagan B, Majumdar R and Ogata H (2022) Fine grain synthetic educational data: Challenges and limitations of collaborative learning analytics. *IEEE Access* 10: 26230–26241.

Gal MS and Lynskey O (2023) Synthetic data: Legal implications of the data-generation revolution. *Iowa Law Review* 109(3): 1087–1156.

Giuffrè M and Shung DL (2023) Harnessing the power of synthetic data in healthcare: Innovation, application, and privacy. *NPJ Digital Medicine* 6(1): 1–8.

Greenspan RL, Baggett L and Boutwell B (2024) Open science practices in criminology and criminal justice journals. *Journal of Experimental Criminology*: 87–107. doi:10.1007/s11292-024-09640-x

Groff E and Mazerolle L (2008) Simulated experiments and their potential role in criminology and criminal justice. *Journal of Experimental Criminology* 4(3): 187–193.

Hartel P, Wegberg R and van Staalduinen M (2023) Investigating sentence severity with judicial open data. *European Journal on Criminal Policy and Research* 29(4): 579–599.

Hittmeir M, Ekelhart A and Mayer R (2019) On the utility of synthetic data: An empirical evaluation on machine learning tasks. In: Proceedings of the 14th international conference on availability, reliability and security, Canterbury CA United Kingdom, 26 August 2019, pp.1–6: ACM.

Houtkoop BL, Chambers C, Macleod M, et al. (2018) Data sharing in psychology: A survey on barriers and preconditions. *Advances in Methods and Practices in Psychological Science* 1(1): 70–85.

Jacques S (2023) Ranking the openness of criminology units: An attempt to incentivize the use of librarians, institutional repositories, and unit-dedicated collections to increase scholarly impact and justice. *Journal of Contemporary Criminal Justice* 39(3): 371–386.

James L (2013) *Defining open data*. Open Knowledge Foundation Blog. Available at: https://blog.okfn.org/2013/10/03/defining-open-data/.

Johnson SD, Bowers KJ, Young C, et al. (2001) Uncovering the true picture: Evaluating crime reduction initiatives using disaggregate crime data. *Crime Prevention and Community Safety* 3(4): 7–24.

Kaabachi B, Despraz J, Meurers T, et al. (2025) A scoping review of privacy and utility metrics in medical synthetic data. *NPJ Digital Medicine* 8(1): 60.

Koetzier LR, Wu J, Mastrodicasa D, et al. (2024) Generating synthetic data for medical imaging. *Radiology* 312(3): e232471.

Krüsselmann K, Aarten P, Granath S, et al. (2023) Firearm homicides in Europe: A comparison with non-firearm homicides in five European countries. *Global Crime* 24(2): 145–167.

Krüsselmann K, Aarten P and Liem M (2024a) Missing the mark? A typology of lethal and non-lethal firearm violence in the Netherlands. *Crime & Delinquency* 72(6-7): 1358–1380.

Krüsselmann K, Achterberg J, Haas M, et al. (2024b) *Making sensitive data open and fair through synthetic data generation – A guidebook*. 12 September. Zenodo. Available at: https://zenodo.org/records/13752141 (accessed 7 May 2025).

Krüsselmann K, Achterberg J, Haas M, et al. (2024c) *Synthetic Dutch Homicide Monitor*. Leiden University/Leiden University Medical Center. Zenodo.

Liem M, Suonpää K, Lehti M, et al. (2019) Homicide clearance in Western Europe. *European Journal of Criminology* 16(1): 81–101.

López CAF and Elbi A (2022) On the legal nature of synthetic data. In: NeurIPS 2022 workshop on synthetic data for empowering ML research, 20 October 2022.

Melamud O and Shivade C (2019) Towards automatic generation of shareable synthetic clinical notes using neural language models. arXiv:1905.07002. arXiv

Miletic M, Sariyar M (2025) Synthetic data generation methods for longitudinal and time series health data. In: Mantas J, Hasman A, Gallos P, et al. (eds) *Studies in Health Technology and Informatics*. IOS Press, 367–371.

Molloy JC (2011) The open knowledge foundation: Open data means better science. *PLOS Biology* 9(12): e1001195.

National Center for Health Statistics (2022) Public-use linked mortality files. Available at: https://www.cdc.gov/nchs/data-linkage/mortality-public.htm (accessed 5 May 2025).

Network of the Marie SkŁodowska-Curie Actions National Contact Points (2023) *Policy-brief: Open science*. 30 March. MSCA-NET. Available at: https://horizoneuropencpportal.eu/sites/default/files/2025-04/task-3.6-open_science_brief.pdf.

Nivette AE (2021) Exploring the availability and potential of international data for criminological study. *International Criminology* 1(1): 70–77.

Nowok B, Raab GM and Dibben C (2017) Providing bespoke synthetic data for the UK longitudinal studies and other sensitive data with the synthpop package for R1. *Statistical Journal of the IAOS* 33(3): 785–796.

OECD (2023) Emerging privacy-enhancing technologies: Current regulatory and policy approaches. *OECD Digital Economic Papers 351*.

Parsons S, Azevedo F, Elsherif MM, et al. (2022) A community-sourced glossary of open scholarship terms. *Nature Human Behaviour* 6(3): 312–318.

Patki N, Wedge R and Veeramachaneni K (2016) The synthetic data vault. In: 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), October 2016, pp.399–410.

Pridemore WA (2011) Poverty matters: A reassessment of the inequality–homicide relationship in cross-national studies. *The British Journal of Criminology* 51(5): 739–772.

Pridemore WA, Makel MC and Plucker JA (2018) Replication in criminology and the social sciences. *Annual Review of Criminology* 1: 19–38.

Qian Z, Callender T, Cebere B, et al. (2024) Synthetic data for privacy-preserving clinical risk prediction. *Scientific Reports* 14(1): 25676.

Quick H and Waller LA (2018) Using spatiotemporal models to generate synthetic data for public use. *Spatial and Spatio-Temporal Epidemiology* 27: 37–45.

Rosenfeld R and Messner SF (2009) The crime drop in comparative perspective: The impact of the economy and imprisonment on American and European burglary rates. *The British Journal of Sociology* 60(3): 445–471.

Rubin DB (1993) Statistical disclosure limitation. *Journal of Official Statistics* 9(2): 461–468.

Schumann S, van der Vegt I, Gill P, et al. (2019) Towards open and reproducible terrorism studies: Current trends and next steps. *Perspectives on Terrorism* 13(5): 61–73.

Severin A, Egger M, Eve MP, et al. (2020) Discipline-specific open access publishing practices and barriers to change: An evidence-based review. *F1000Research* 7: 1925.

Sigfusdottir ID, Kristjansson AL and Agnew R (2012) A comparative analysis of general strain theory. *Journal of Criminal Justice* 40(2): 117–127.

Skott S (2019) Disaggregating homicide: Changing trends in subtypes over time. *Criminal Justice and Behavior* 46(11): 1650–1668.

Stadler T, Oprisanu B and Troncoso C (2022) Synthetic data – anonymisation groundhog day. In: 31st USENIX Security Symposium (USENIX Security 22), 2022, pp.1451–1468.

statice (n.d.) Awesome-synthetic-data: A curated list of awesome synthetic data tools (open source and commercial). Available at: https://github.com/statice/awesome-synthetic-data/?tab=readme-ov-file (accessed 3 February 2026).

Templ M and Sariyar M (2022) A systematic overview on methods to protect sensitive data provided for various analyses. *International Journal of Information Studies* 21(6): 123–123.

Tenopir C, Allard S, Douglass K, et al. (2011) Data sharing by scientists: Practices and perceptions. *PLoS ONE* 6(6): e21101.

Umbach G (2024) Open science and the impact of open access, open data, and FAIR publishing principles on data-driven academic research: Towards ever more transparent, accessible, and reproducible academic output? *Statistical Journal of the IAOS* 40(1): 59–70.

Walonoski J, Kramer M, Nichols J, et al. (2018) Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association* 25(3): 230–238.

Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. (2016) The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* 3(1): 160018.

Yodanis CL (2004) Gender inequality, violence against women, and fear: A cross-national test of the feminist theory of violence against women. *Journal of Interpersonal Violence* 19(6): 655–675.

Zhang D and Kifer D (2017) LightDP: Towards automating differential privacy proofs. In: *Proceedings of the 44th ACM SIGPLAN Symposium on Principles of Programming Languages*, 888–901.