

Synthetic data in medicine: exploring resilience in emerging human-machine relationships

Paula Ziethmann, Sarah Friedrich, Kerstin Schlögl-Flierl

Angaben zur Veröffentlichung / Publication details:

Ziethmann, Paula, Sarah Friedrich, and Kerstin Schlögl-Flierl. 2024. "Synthetic data in medicine: exploring resilience in emerging human-machine relationships." *Datenschutz und Datensicherheit - DuD* 48 (6): 358–63.
<https://doi.org/10.1007/s11623-024-1926-x>.

Paula Ziethmann, Sarah Friedrich, Kerstin Schlögl-Flierl

Synthetic Data in Medicine

Exploring Resilience in Emerging Human-Machine Relationships

This paper explores the multifaceted implications of synthetic data in AI model development, particularly in medical contexts such as oncology. It examines the benefits of synthetic data, including privacy enhancement and bias reduction, but also highlights the associated risks, such as data loss and bias exacerbation. The paper discusses the ethical considerations and proposes strategies to ensure resilience in human-machine relationships, especially in oncology.



Paula Ziethmann

ist wissenschaftliche Mitarbeiterin am Center for Responsible AI Technologies und Doktorandin an der Universität Augsburg. In ihrer Dissertation erforscht sie unter einer foucaultschen Perspektive den Einsatz von Künstlicher Intelligenz in der Klinik sowie dessen Implikationen für Wissens- und Machtdynamiken.

E-Mail: paula.ziethmann@zig.uni-augsburg.de

Bildrechte: Lux Studio Productions



Sarah Friedrich

ist Professorin für Mathematische Statistik und Künstliche Intelligenz in der Medizin an der Universität Augsburg und befasst sich mit den statistischen Aspekten und der Datengrundlage von maschinellen Lernverfahren.

E-Mail: sarah.friedrich@math.uni-augsburg.de

Bildrechte: Universitätsmedizin Göttingen



Kerstin Schlögl-Flierl

ist Lehrstuhlinhaberin und Professorin für Moralthologie an der Universität Augsburg. Sie ist außerdem Principal Investigator am Center for Responsible AI Technologies und Mitglied im Deutschen Ethikrat.

E-Mail: kerstin.schloegl-flierl@kthf.uni-augsburg.de

Bildrechte: Deutscher Ethikrat

1 Introduction

The integration of artificial intelligence (AI) has ushered in transformative possibilities across diverse aspects of human life and professional domains. A critical aspect of AI model development involves the training process, which traditionally relies on extensive datasets comprising various data types, such as videos, images, texts, and graphs. However, the acquisition of authentic datasets on such a scale is a formidable and resource-intensive undertaking.¹ In response to this challenge, researchers are increasingly turning to synthetic data, artificially generated datasets that closely mirror real-world data.²

Synthetic data can offer several advantages, including the inherent difficulty in re-identifying personal details, thereby exempting synthetic data from the constraints of the General Data Protection Regulation (GDPR) and facilitating more straightforward and secure processing.³ The generation of synthetic data is generally less time- and cost-intensive than collecting real-world data, providing researchers with a more efficient alternative. Additionally, the theoretical possibility to generate an infinite volume of data through synthetic means further accentuates its appeal.⁴ Noteworthy is the assertion that AI models may learn more effectively with synthetic data due to reduced distractions, such as extraneous background objects.⁵ This quality is particularly pertinent in medical applications where precision and focus are para-

© Der/die Autor(en) 2024. Dieser Artikel ist eine Open-Access-Publikation.

1 Zewe, Adam (2022): In machine learning, synthetic data can offer real performance improvements. In: *MIT News Office*. URL: <https://news.mit.edu/2022/synthetic-data-ai-improvements-1103>.

2 UK Statistics Authority (2022): Ethical considerations relating to the creation and use of synthetic data. URL: <https://uksa.statisticsauthority.gov.uk/publication/ethical-considerations-relating-to-the-creation-and-use-of-synthetic-data/pages/2/>.

3 Datenethikkommission (DEK) (2019): Gutachten der Datenethikkommission.

4 Jordon, James; Szpruch, Lukasz; Houssiau, Florimond; Bottarelli, Mirko; Cherubin, Giovanni; Maple, Carsten et al. (2022): Synthetic Data -- what, why and how?, arXiv preprint arXiv:2205.03257.

5 Zewe (Fn. 1).

mount. The measurable nature of synthetic data quality could allow researchers to preserve specific characteristics from base data or strategically eliminate biases that could lead to discriminatory outcomes.⁶ The ability of synthetic data to augment the quantity and diversity of datasets concurrently enhances the robustness and adaptability of AI models in medical contexts.⁷

However, the increased reliance on synthetic data introduces a paradigm shift in ethical considerations and guidelines in machine learning.⁸ Trustworthy AI applications are now subject to elevated standards, including requirements for fairness, correctness, and robustness.⁹ Additionally, the quality of synthetic data is intricately tied to the quality of the underlying real data, introducing a critical dimension of dependency.¹⁰ In addition, synthetic data may actually exacerbate an existing bias in the data. It is also debatable whether synthetic data really leads to more data security or even less.

This paper seeks to delve into the multifaceted implications of synthetic data, addressing the risks associated with potential loss of crucial details, the persistence of biases, and challenges in determining the novelty of data generated synthetically.¹¹ First, we will define the characteristics of synthetic data and how it can be generated. We will then look at our application example of oncology and describe how synthetic data is used here. In Chapter 4, we will discuss the dual function of synthetic data and the current state of research: On the one hand, synthetic data are seen as having the potential to increase privacy and reduce bias; on the other hand, the exact opposite is also described. In Chapter 5, we will link the dual role of synthetic data to ethical considerations on how we can ensure the resilience of the human-machine relationship in oncology.

2 Synthetic Data: Characteristics and Generation

Synthetic data is data that has been generated from or according to real-world data, thus resembling the statistical properties of the real data set. The degree to which a synthetic data set is a proxy for the real data varies depending on the method of synthesizing the data.¹² Synthetic data may be generated from a real data set directly or may be based on an existing statistical model and/or background knowledge. Furthermore, it is important to distinguish the type of data to be synthesized: this may include images, tabular data (e.g. electronic health records) or sensor data (usually measured longitudinally over time). Depending on the type of data, different methods for generating synthetic data are available:

Early ideas of generating synthetic data stem from Rubin and date back to 1993. His idea was to use the framework of multiple

imputation, where missing data is imputed based on the characteristics of the existing data. In his 1993 paper, he suggested to use this approach for generating a completely artificial data set.¹³ This approach can be applied to tabular data.

A recent systematic review¹⁴ gives an overview of methods for generating tabular data in healthcare applications. In particular, they differentiate between “baseline methods” (relating to anonymization techniques such as deleting sensitive attributes), “statistical models” (simulating synthetic data based on statistical properties such as correlation structures between variables), “machine learning models” (predicting new data based on supervised learning approaches), “deep learning approaches” (such as Autoencoders, Generative Adversarial Networks (GANs) and Ensembles, see more details below) and “other approaches” (including two-step procedures incorporating expert knowledge and SynSys, which was developed by Dahmen, to generate sensor data based on Hidden Markov Models).¹⁵

One of the most interesting developments in AI, especially with regard to synthetic data generation, are GANs (Generative Adversarial Networks). The idea is to combine two AI systems: one is generating the data, while the other is the so-called “adversary”, which “judges” the performance of the generated data. A special aspect is that the data-generating network has no information of the original data, only the adversary knows these properties and can therefore evaluate the performance of the generator. GANs are also the most popular choice for generating synthetic image data.¹⁶

Similarly to different types of data, there also exist different aims of using synthetic data. An important aspect is the lack of enough (freely available) real-world data, especially in the context of data-hungry AI applications. For the training of these models, large amounts of data are required, which are often not available. This problem is especially prevalent in a medical context, where data sharing is difficult and especially rare diseases lack enough data. Related to this issue is the aim of sharing data without privacy concerns. This is usually not easily possible for real-world data, especially in a medical context, where the data may contain highly sensitive information. Here, synthetic data could offer an alternative. Another aspect, however, is that of generating data with known properties, that is, data that may not exist in this form in the real world. In the statistical literature, this process is referred to as simulation.¹⁷

Once the synthetic data is generated, it can be used for different purposes: to train or test new models, for example, to generate ideas that can then be verified in real-world data or to compare different methods with regard to their performance.¹⁸

The quality of the synthetic data thus strongly depends on its intended use. Here, four aspects are usually distinguished:

6 Datenethikkommission (Fn. 3).

7 Jacobsen, Benjamin N. (2023): Machine learning and the politics of synthetic data. In: *Big Data & Society* 10 (1), 2023. DOI: 10.1177/20539517221145372.

8 Jacobsen (Fn. 7).

9 Hecker, Dirk; Voss, Angi; Paaß, Gerhard; Wirtz, Tim (2023): Big Data 2.0 – mit synthetischen Daten KI-Systeme stärken. In: *Wirtschaftsinformatik und Management* 15 (2), 2022, S. 161–167. DOI: 10.1365/s35764-022-00437-z.

10 UK Statistics Authority (Fn. 2).

11 Raji, Behrang (2021): Rechtliche Bewertung synthetischer Daten für KI-Systeme. In: *Datenschutz Datensicherheit* 45 (5), 2021, S. 303-309. DOI: 10.1007/s11623-021-1439-9.

12 Emam, Khaled; Moaquera, Lucy; Hoptroff, Richard (2020): Practical synthetic data generation. Balancing privacy and the broad availability of data. First edition. Sebastopol, CA: O'Reilly Media, Inc.

13 Rubin, D.B., (1993): Discussion: Statistical Disclosure Limitation. In: *Journal of Official Statistics* 9, S. 461-468.

14 Hernandez, Mikel; Epelde, Gorka; Alberdi, Ane; Cilla, Rodrigo; Rankin, Debbie (2022): Synthetic data generation for tabular health records: A systematic review. In: *Neurocomputing* 493, 2022, S. 28–45. DOI: 10.1016/j.neucom.2022.04.053.

15 Dahmen, Jessamyn; Cook, Diane (2019): SynSys: A Synthetic Data Generation System for Healthcare Applications. In: *Sensors* 19 (5), 2019. DOI: 10.3390/s19051181.

16 Raji (Fn. 11).

17 Friedrich, Sarah; Friede, Tim (2024): On the role of benchmarking data sets and simulations in method comparison studies. In: *Biometrical journal. Biometrische Zeitschrift* 66 (1), 2024. e2200212. DOI: 10.1002/bimj.202200212.

18 Friedrich/Friede (Fn. 17).

(i) resemblance (how well the synthetic data represents the real-world data),

(ii) utility (describes the usability of statistical conclusions drawn from the synthetic data as well as the (predictive) quality of the resulting machine learning models),

(iii) privacy (can personal data be deduced from the synthetic data), and

(iv) fairness¹⁹.

The latter refers to the representation of (minority) sub-groups in the synthetic data and the extent to which certain classes are discriminated (by being under-represented) and is closely linked to bias.

It is also worth noting that not all of these aspects are equally relevant for each application. While a high utility is, for example, highly relevant for pre-training models that will later be applied to real data, a reduced utility might not be an issue in other areas, for example if the synthetic data is only used to generate ideas or compare existing models.

3 Synthetic Data in Oncology

In this chapter, we aim to delve into the utilization of synthetic data in medicine. We will examine two examples: the synthetic cancer registry “Simulacrum” and a study involving the extraction and simulation of survival times, also in oncology.

An example of synthetic data in a medical context is the synthetic cancer registry data “Simulacrum”, which was developed by the National Health Service (NHS) in the UK in cooperation with AstraZeneca and IQVIA. This register, which is publicly available at <https://simulacrum.healthdatainsight.org.uk/>, is based on the cancer registry by the National Disease Registration Service (NDRS).

The Simulacrum contains only synthetic data and therefore enables data sharing without compromising privacy issues. The data structure mimics that of the real cancer registry and data was simulated by capturing the statistical relationships between the variables contained in the real data.²⁰ This is a three-step process: after identification of the statistical properties of the real data, a generative data model is developed. This is then used to produce the synthetic data available for public use.

The idea of the Simulacrum is not to replace real data, but to learn the data structure and generate hypotheses, which can then be verified in real-world data, for example in the NDRS cancer registry.

Since the synthetic data so closely mimics the real data, it is even possible to apply models developed on the Simulacrum to the NDRS cancer registry. This enables answering specific research questions on real-world data without compromising patient privacy, since an external team of researchers can develop and validate code on the Simulacrum, which can then be applied on the real-world data by the NDRS team itself, see e.g. Legg et

al.²¹ and Shaw et al.²² for such analyses. Thus, only the code and the corresponding results are transferred, but the data itself never leaves “its owner”.

Another example of the application of synthetic data in oncology is shown by Thurow et al.²³ Their study involves the extraction and simulation of survival times from published datasets, with a focus on mitigating bias. Unlike the Simulacrum cancer registry, which replicates the entire data structure, this study specifically targets survival times and does not include covariates. Thurow et al.’s approach involves the extraction of survival data from published sources, followed by the development of a simulation methodology to generate synthetic survival times. The primary goal is to create a dataset that reflects the statistical properties of real survival data, allowing researchers to explore hypotheses and validate models without direct access to sensitive patient information.

As we can see from these two examples, the benefits or promises of using synthetic data in concrete applications often lie in ensuring privacy and reducing bias. In the next chapter, we will have a closer look at both of these promises through a review of the state of research in this area.

4 The Dual Role of Synthetic Data

Synthetic data offer a range of promises over real-world data: they facilitate data sharing without privacy concerns, mitigate bias, e.g. by providing additional data when real-world data may be scarce, particularly in the context of rare diseases. In this chapter, we will look at and discuss these promises.

Several studies and papers write about improved privacy through the use of synthetic data. Giuffrè and Shung highlight the utility of synthetic text in protecting sensitive information, particularly in the mental health domain, thereby reducing the risk of compromising individual patient data. In addition, the adoption of Differential Privacy (DP) principles in conjunction with synthetic data is highlighted as a pragmatic approach to strengthening privacy frameworks.²⁴

Gonzales et al. further elucidate the benefits of synthetic data, highlighting its potential to provide developers with realistic datasets while circumventing privacy concerns, thereby accelerating development processes and conserving resources.²⁵ In addition, the compositional flexibility of synthetic data complicates re-identification attempts, thereby enhancing privacy protections.²⁶ Furthermore, the use of synthetic data provides a via-

21 Legg, Alex; Lambova, Alexandrina; Broe, Anne; Levy, Julia; Medalla, Greg (2023): Real-World Experience With CPX-351 Treatment for Acute Myeloid Leukemia in England: An Analysis From the National Cancer Registration and Analysis Service. In: *Clinical lymphoma, myeloma & leukemia* 23 (10). S. 323-330. DOI: 10.1016/j.clml.2023.07.003.

22 Shaw, Clare; Starling, Naureen; Reich, Adam; Wilkes, Emily; White, Rebecca; Shepelev, Julian; Narduzzi, Silvia (2020): Modification of systemic anti-cancer therapies and weight loss, a population-level real-world evidence study. In: *Therapeutic advances in medical oncology* 12. DOI: 10.1177/1758835920982805.

23 Thurow, Maria; Dormuth, Ina; Sauer, Christina; Ditzhaus, Marc; Pauly, Markus (2023): How to Simulate Realistic Survival Data? A Simulation Study to Compare Realistic Simulation Models. In: arXiv:2308.07842 2023.

24 Giuffrè, Mauro; Shung, Dennis L. (2023): Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. In: *NPJ digital medicine* 6 (1), S. 186. DOI: 10.1038/s41746-023-00927-3.

25 Gonzales, Aldren; Guruswamy, Gururabha; Smith, Scott R. (2023): Synthetic data in health care: A narrative review. In: *PLOS digital health* 2 (1). e0000082. DOI: 10.1371/journal.pdig.0000082.

26 Gonzales et al. (Fn. 25).

19 Bhanot, Karan; Qi, Miao; Ericjson, John S.; Guyon, Isabelle; Bennett, Kristin P. (2021): The Problem of Fairness in Synthetic Healthcare Data. In: *Entropy* 23 (9), 2021. DOI: 10.3390/e23091165.

20 See for more information: ‘Frayling, Lora (2018): Generating the Simulacrum – A methodology overview. URL: <https://simulacrum.healthdatainsight.org.uk/wp/wp-content/uploads/2018/11/Methodology-Overview-Nov18.pdf>.

ble solution to circumvent data availability constraints, offering a plausible alternative to real-world data collection.²⁷

Conversely, privacy and data protection concerns are articulated in the opposing literature, warning of the inherent risks associated with synthetic data. Giuffrè and Shung also caution against overlooking privacy considerations in system design and deployment and advocate a proactive “privacy by design” approach, particularly in clinical contexts.²⁸ Gonzales et al. highlight the susceptibility of synthetic data to re-identification, particularly when rare conditions are replicated within synthetic datasets, thereby exacerbating privacy risks. They also highlight the trade-off between realism and disclosure risk in the construction of synthetic data, raising questions about the adequacy of privacy safeguards.²⁹ In addition, the potential exploitation of synthetic data for nefarious purposes, such as patient impersonation and fraudulent insurance claims, underscores the vulnerability of privacy in healthcare.³⁰

This duality in the effectiveness of synthetic data extends beyond privacy concerns to include biases within datasets. Draghi et al. propose methods to mitigate biases in synthetic data generation through probabilistic approaches, aiming to improve the realism and predictive performance of synthetic datasets³¹. Similarly, Barbierato et al. advocate the integration of probabilistic networks to model biases within synthetic datasets, thereby promoting fairness and mitigating discrimination.³² Wang et al. posit synthetic data as a potential way to expose biases inherent in real-world datasets, facilitating algorithm development and fairness assessments.³³

However, the literature also cautions against the inadvertent perpetuation or exacerbation of bias through synthetic data generation methods. Giuffrè and Shung warn of the amplification of biases present in primary datasets, highlighting the potential for discriminatory outcomes in medical research and practice.³⁴ In addition, challenges related to the interpretability and transparency of synthetic data generation algorithms hinder efforts to mitigate bias, potentially undermining trust in generated datasets. Gonzales et al. highlight the reproduction of biases from primary datasets in synthetic versions, further complicating efforts to ensure fairness.³⁵ These concerns are echoed by Bhanot et al., who highlight the biases embedded in published synthetic

datasets, casting doubt on the fairness and reliability of synthetic data for health care applications.^{36, 37}

The existing literature presents a dichotomous perspective on the efficacy of synthetic data, revealing a dual function. Some sources emphasize its ability to enhance privacy and mitigate bias, while others highlight its increased vulnerability to privacy violations and potential exacerbation of bias. This duality within the technical discourse through narrative explanations allows the following conclusions to be drawn.

5 Ethical Considerations: Resilience in Human-Machine Relationships in Oncology

AI is not used in a vacuum, but in specific contexts – more precisely: in certain relationships. The potential and risks associated with the use of synthetic data depend on the socio-technical context in which the system is applied. In this paper, we have examined the use case in medicine, specifically in oncology, and want to address the ethical considerations involved. In doing so, we advocate an assessment of the resilience of the human-machine relationship in order to be able to evaluate the aforementioned risks and opportunities in a context-specific manner. Although we are referring here to oncology, we believe that it is also worth considering the resilience of human-machine relationships in other domains as a framework for contextually weighing ethical considerations and technical advantages and disadvantages.

What does resilience mean in this medical context? In a general sense, it can be understood as resistance (see in various materials), sometimes it also refers to the ability to adapt. In a broader sense, resilience is also understood as the ability to transform institutions and individuals.³⁸ However, it is not physical or structural resilience that is meant here, but resilience in the human-machine relationship. The human-machine relationship refers to the complex interplay between human individuals and technological systems, in which interactions, dependencies, and mutual influences play a role. This relationship includes physical interactions as well as cognitive, emotional, and social aspects that shape and influence the behavior and interactions of the parties involved.

Here, we investigate the resilience of human-machine relationships when using AI in oncology trained on synthesized data. Although the data is synthesized, human authorship must not be compromised.³⁹ Human authorship in this context means that the freedom, autonomy, and responsibility of the people involved is important. From a data ethics standpoint, ensuring transparency, data protection, and maintaining high data quality are imperative measures to uphold these principles.⁴⁰

The demand for high data quality also alludes to the issue of bias. The issue of bias is raised by various stakeholders, including medical professionals and engineers, indicating a growing awareness of bias. Bias awareness requires proactive identifica-

27 Rankin, Debbie; Black, Michaela; Bond, Raymond; Wallace, Jonathan; Mulvenna, Maurice; Epelde, Gorka (2020): Reliability of Supervised Machine Learning Using Synthetic Data in Health Care: Model to Preserve Privacy for Data Sharing. In: *JMIR medical informatics* 8 (7), 2020. e18910. DOI: 10.2196/18910.

28 Giuffrè/Shung (Fn. 24).

29 Gonzales et al. (Fn. 25).

30 Chen, Richard J.; Lu, Ming Y.; Chen, Tiffany Y.; Williamson, Drew F. K.; Mahmood, Faisal (2021): Synthetic data in machine learning for medicine and healthcare. In: *Nature biomedical engineering* 5 (6), S. 493-497. DOI: 10.1038/s41551-021-00751-8.

31 Draghi, Barbara; Wang, Zhenchen; Myles, Puja; Tucker, Allan (2021): Bayesboost: identifying and handling bias using synthetic data generators. In: *Third International Workshop on Learning with Imbalanced Domains: Theory and Applications*, PMLR, S. 49-62.

32 Barbierato, Enrico; Della Vedova, Marco L.; Tessera, Daniele; Toti, Daniele; Vanoli, Nicola (2022): A Methodology for Controlling Bias and Fairness in Synthetic Data Generation. In: *Applied Sciences* 12 (9), S. 4619. DOI: 10.3390/app12094619.

33 Wang, Zhenchen; Myles, Puja; Tucker, Allan (2021): Generating and evaluating cross-sectional synthetic electronic healthcare data: Preserving data utility and patient privacy. In: *Computational Intelligence* 37 (2), S. 819-851. DOI: 10.1111/coi.12427.

34 Giuffrè/Shung (Fn. 24).

35 Gonzales et al. (Fn. 25).

36 Bhanot, Karan; Baldini, Ioana; Wei, Dennis; Zeng, Jiaming; Bennett, Kristin P. (2022): Downstream Fairness Caveats with Synthetic Healthcare Data. arXiv preprint arXiv:2203.04462.

37 Bhanot et al. (Fn. 19).

38 Schlögl-Flierl, Kerstin (2023): Resilience – normatively conceived, transformatively developed. In: *Ethics and Armed Forces* 2023 (1), S. 8-14.

39 Deutscher Ethikrat (2023): Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz, Berlin.

40 Datenethikkommission (Fn. 3).

tion and mitigation strategies throughout the synthetic data generation process, especially in healthcare.⁴¹

Healthcare professionals' understanding of the capabilities and limitations of machines is critical to fostering effective collaboration and a resilient human-machine relationship in the field of oncology. Only when users of AI systems are aware of both the opportunities and risks associated with their use they are enabled to make informed and prudent decisions. This awareness mitigates the potential for automation bias, which is characterized by a tendency to over-rely on AI systems, while avoiding a wholesale rejection of such technologies – and thus ensuring the resilience of the human-machine relationship.

5.1 Privacy

Robust human-machine relationships in the use of synthetic data in oncology require specific data types and use strategies. Synthetic data can enhance trust, digital sovereignty, and patient well-being in healthcare, where privacy is paramount.⁴² The promise of synthetic data lies in its potential to address privacy concerns through the statistical replication of real-world data, as demonstrated by studies such as that of Hernandez et al.⁴³

In our oncology use case, we considered two examples of synthetic data: Thurow et al.'s⁴⁴ approach is solely based on published summary statistics, thus individual patient data is not involved in any step of the data-generating process, resulting in synthetic data with a very high level of privacy. The Simulacrum, on the other hand, is based on a real cancer registry. Here, the degree of privacy depends on the data-generating process. Since the Simulacrum is generated based on statistical models, using anonymized patient data and statistical correlations between variables, the degree of privacy preservation is expected to be very high. In particular, the maintainers of the Simulacrum state:

“The data in the Simulacrum is made up entirely of artificial patient records. It does not contain any real patient information, so it cannot be used to identify a real person.”⁴⁵

However, to the best of our knowledge, a formal check using methods as described by Hernandez et al.⁴⁶ has not been performed.⁴⁷ This is, however, necessary to ensure data privacy, see also Jordon et al. for a discussion of different aspects of privacy, including DP.⁴⁸

Concerning the promise of reduced issues with privacy regulation, the extent to which this can be fulfilled by synthetic data depends on the context: As mentioned above, different types of data exist and different approaches of synthesizing them can be applied. Synthetic tabular data, which is created based on a statistical model informed by background knowledge and/or real-world data, does indeed protect the privacy of patients in the real-world data, since only statistical properties of the data set are used. An

example is the Simulacrum mentioned above, which does not contain any data stemming from real patients.

In their systematic review, Hernandez et al. also investigated the extent to which the different methods for generating synthetic data preserve privacy and found that also deep-learning based approaches and GANs work reasonably well in this context.⁴⁹ However, their review showed that the privacy dimension is often not addressed in publications and even if it is, care should be taken since the respective evaluations have been carried out by the authors' themselves and might thus be somewhat over-optimistic.

A related issue is that there exist no standardized metrics to evaluate and compare different approaches. It is not clear how exactly the privacy-preserving nature of a method can be evaluated. Hernandez et al. discuss a range of metrics, including simulated attack scenarios, to see whether real patient data can be disclosed.⁵⁰

5.2 Bias

The second aspect of synthetic data is the claim that it might help mitigate bias found in real-world data. Here, however, one should keep in mind that the synthetic data is generated based on real-world data and, depending on the synthesizing mechanism, can only be as unbiased as the real-world data itself.

Baumann et al. and Bhanot et al. propose frameworks and metrics for assessing fairness and mitigating bias in synthetic data. While synthetic data can mitigate the biases found in real data, they are only as unbiased as the underlying data and the synthesis mechanism used. There is also a trade-off between similarity, utility, privacy, and fairness that must be carefully considered to avoid overfitting and privacy violations.^{51, 52}

An example: if a GAN is used to create images of humans, but the database does not contain images of Asian people, the synthetic data will also not contain such images. This issue is also referred to as the “synthetic data fairness problem”⁵³. If the synthetic data are generated based on a statistical model, it is possible to address issues of bias that exist in the real-world data through adapting the model. Therefore, however, the kind of bias present in the data *needs to be known*, such that it can be explicitly addressed. For example, an existing gender bias in a data set, reflected by a negative correlation between female sex and income, say, can be addressed by a model where this correlation is removed or reduced.

However, in order to do this, the researcher must be aware of the bias present in the real-world data. In the same manner, bias may also be introduced into synthetic data. Baumann et al. develop a framework for generating synthetic data that contains a certain type of bias, with the aim of demonstrating how various biases occur in data. Bhanot et al. also suggest some measures for evaluating fairness in synthetic data.⁵⁴

These considerations demonstrate that there is a trade-off between the four quality aspects of synthetic data mentioned above. While it might at first glance look like a good idea to have a very

41 Giuffrè/Shung (Fn. 24).

42 Ibd.

43 Hernandez et al. (Fn. 14).

44 Thurow et al. (Fn. 23).

45 Simulacrum (2024, March 06): simulacrum.healthdatainsight.org.uk
<https://simulacrum.healthdatainsight.org.uk/>.

46 Ibd.

47 See also Yale, Andrew; Dash, Saloni; Dutta, Ritik; Guyon, Isabelle; Pavao, Adrien; Bennett, Kristin P. (2020): Generation and evaluation of privacy preserving synthetic health data. In: *Neurocomputing* 416, S. 244-255. DOI: 10.1016/j.neucom.2019.12.136.

48 Jordon et al. (Fn. 4).

49 Hernandez et al. (Fn. 14).

50 Ibd.

51 Baumann, Joachim; Castelnovo, Alessandro; Crupi, Riccardo; Inverardi, Nicole; Regoli, Daniele (2023): Bias on Demand: A Modelling Framework That Generates Synthetic Data With Bias. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, S. 1002–1013.

52 Bhanot et al. (Fn. 19)

53 Ibd.

54 Bhanot et al. (Fn. 36)

close resemblance between synthetic and real data, this might lead to overfitting and violate privacy preservation, since synthetic patients may too closely resemble real patients.⁵⁵

6 Conclusion

Synthetic data represent a potential solution to mitigate bias independently; however, synthetic data are derived from real-world data, necessitating the identification and mitigation of bias inherent in the underlying data. Bias detection becomes an integral part of the synthetic data generation process. However, such awareness of various biases is not only a prerequisite during development, but also critical for users, especially in our use case, medical practitioners.

This understanding is critical to fostering resilience in the human-machine relationship. Medical professionals must recognize the capabilities and limitations of the machine or system in order to establish effective collaboration.

Further investigation reveals that resilient datasets require specific types of data and appropriate use strategies, including coping strategies. Synthetic data can contribute to resilient relationships by enhancing trustworthiness, increasing digital data sovereignty, and supporting patient welfare, particularly in medical contexts where patient privacy is a paramount concern.

Ensuring the responsible use of synthetic data involves disclosing synthesis processes, tailoring datasets to specific needs, and acknowledging limitations in extrapolating findings to other domains. It is important to be cautious in expressing hope for the promise of AI, recognizing that resilience may be limited to certain domains and data types. In summary, the question is: Is synthetic data proving to be more resilient? Resilience, particularly with respect to privacy and bias awareness, depends on the field and type of data involved.

Open Access

Dieser Artikel wird unter der Creative Commons Namensnennung 4.0 (CC BY) International Lizenz veröffentlicht, welche die Nutzung, Vervielfältigung, Bearbeitung, Verbreitung und Wiedergabe in jeglichem Medium und Format erlaubt, sofern Sie den/ die ursprünglichen Autor(en) und die Quelle ordnungsgemäß nennen, einen Link zur Creative Commons Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden.

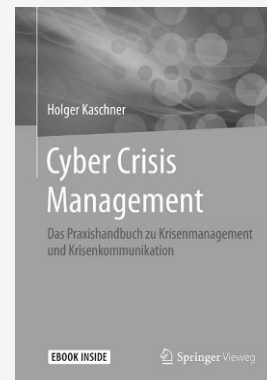
Weitere Details zur Lizenz entnehmen Sie bitte der Lizenzinformation auf <http://creativecommons.org/licenses/by/4.0/deed.de>.

Funding

Open Access funding enabled and organized by Projekt DEAL.

⁵⁵ Ibd.

System- und Datensicherheit



H. Kaschner

Cyber Crisis Management

Das Praxishandbuch zu Krisenmanagement und Krisenkommunikation

2020, XII, 223 S. 10 Abb. Book + eBook. Brosch.

€ (D) 34,99 | € (A) 35,97 | *CHF 39,00

ISBN 978-3-658-27913-4

€ 26,99 | *CHF 31,00

ISBN 978-3-658-27914-1 (eBook)

- Das Praxishandbuch in deutscher Sprache zu Krisenmanagement und Krisenkommunikation
- Hilft auch zur Vorbereitung auf und Prävention von Cyber-Krisen
- Mit zahlreichen Abbildungen und Checklisten

Ihre Vorteile in unserem Online Shop:

Über 280.000 Titel aus allen Fachgebieten | eBooks sind auf allen Endgeräten nutzbar | Kostenloser Versand für Printbücher weltweit.

€ (D): gebundener Ladenpreis in Deutschland, € (A): in Österreich. *: unverbindliche Preisempfehlung. Alle Preise inkl. MwSt.

Part of **SPRINGER NATURE**

springer.com/informatik