Synthetic Data, Similarity-based Privacy Metrics, and Regulatory (Non-)Compliance[†]

Georgi Ganev [1 2]

1. Motivation

Synthetic tabular data, or data generated by machine learning generative models, is gaining popularity beyond academia and moving into real-world deployments. Examples include releasing public census data by US (Abowd et al., 2022), UK (UK ONS, 2023), and Israel (Hod & Canetti, 2024), as well as sharing sensitive financial and health data through private synthetic data vendors (UK ICO, 2023; Microsoft, 2024). While these releases satisfy a formal definition of privacy, i.e., Differential Privacy (DP) (Dwork et al., 2006), this is still not the norm in numerous scientific papers (Park et al., 2018; Lu et al., 2019; Zhao et al., 2021; Borisov et al., 2023; Yoon et al., 2023; Kotelnikov et al., 2023; Zhang et al., 2024) and leading synthetic data vendors (Mostly AI, 2020; Syntegra, 2021; Panfilo & Aindo, 2022; Syntho, 2023). Instead, the papers/companies rely entirely on empirical ad-hoc privacy metrics based on the *similarity* between synthetic and real personal datasets.

Main Question. This prompts asking: *“Is using similarity-based privacy metrics enough to consider synthetic data regulatory compliant?”* Due to their fundamental issues and unreliable, inconsistent nature, *we argue that it is not*.

2. Definitions

Synthetic Data. We denote a real personal dataset as $\mathcal{D}$. A generative model, G, is trained on $\mathcal{D}_{train}$ (a subset of $\mathcal{D}$; the remaining data, $\mathcal{D}_{test}$, is set aside for test purposes) to capture a probability representation, and could later be sampled to generate new (synthetic) data $\mathcal{D}_{synth}$ of arbitrary size (see bottom of Fig. 1). Popular generative models include Graphical Models (Zhang et al., 2017; McKenna et al., 2021), GANs (Xie et al., 2018; Jordon et al., 2019; Xu et al., 2019), Diffusion Models (Kotelnikov et al., 2023; Zhang et al., 2024), and Transformers (Borisov et al., 2023).

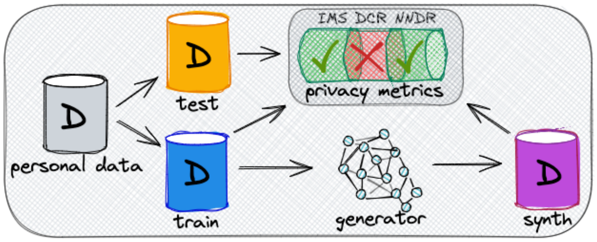


Figure 1. Data flow overview.

Similarity-based Privacy Metrics (SBPMs). The intuition behind SBPMs is that $\mathcal{D}_{synth}$ should be representable and close to $\mathcal{D}_{train}$, but not closer than to $\mathcal{D}_{test}$ (Platzer & Reutterer, 2021; Mobey Forum, 2022). More precisely, the closest pairwise distances $d_{synth} = d(\mathcal{D}_{train}, \mathcal{D}_{synth})$ and $d_{test} = d(\mathcal{D}_{train}, \mathcal{D}_{test})$ are computed and their distributions compared through a statistical test (see Fig. 1). The passing criterion is a comparison between a simple statistic run on each distribution, e.g., average/5th percentile (p5). In this paper, we focus on the three most widely used SBPMs by scientific papers and synthetic data vendors. Finally, $\mathcal{D}_{synth}$ is deemed private if all three privacy tests pass (Mostly AI, 2020; Panfilo & Aindo, 2022).

Identical Match Share (IMS) calculates the proportion of exact copies (statistic: average; test: $d_{synth} \leq d_{test}$).

Distance to Closest Records (DCR) calculates the distances to their nearest neighbor in $\mathcal{D}_{train}$ (statistic: p5; test: $d_{synth} \geq d_{test}$). DCR is meant to protect against scenarios where $\mathcal{D}_{train}$ is slightly perturbed and passed as $\mathcal{D}_{synth}$.

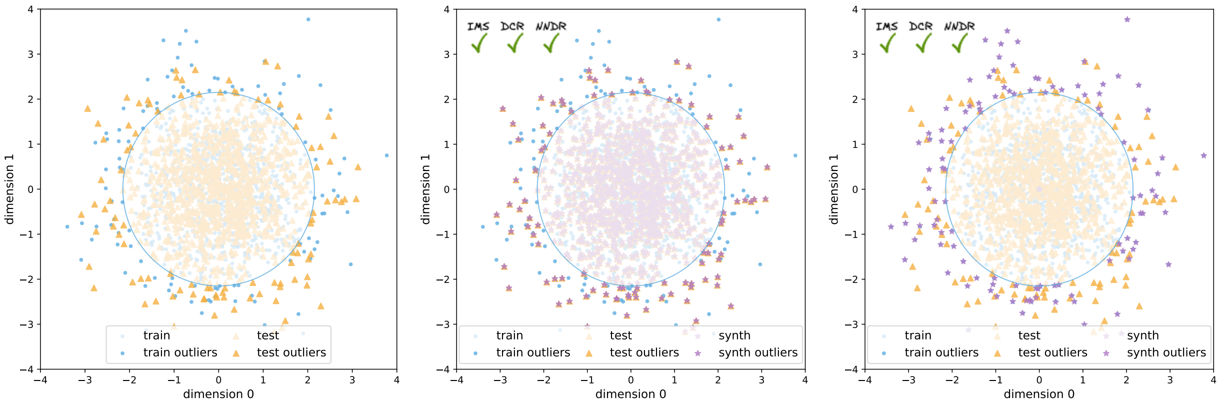
Nearest Neighbor Distance Ratio (NNDR) follows DCR but divides the distances by the distance to their second neighbor (statistic: p5; test: $d_{synth} \geq d_{test}$). The relative computations are supposed to further protect the outliers.

GDPR. EP and Council (2016a) define *personal data* as “any information relating to an identified or identifiable living individual.” Also, EP and Council (2016b) state that effectively anonymized information is not considered personal data and is exempt from data protection regulations. Creating synthetic data from real personal data naturally involves processing it, so whether the result is personal or anonymous depends on the identifiability risk assessment.

[†] For an extended version, refer to (Ganev & De Cristofaro, 2023).

[1] University College London, London, UK [2] Hazy, London, UK. Correspondence to: Georgi Ganev <georgi.ganev.16@ucl.ac.uk>.

Accepted to the 2^{nd} *Workshop on Generative AI and Law*, co-located with the *International Conference on Machine Learning*, Vienna, Austria. 2024.



(a) $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$

(b) $\mathcal{D}_{synth}$ leaking all $\mathcal{D}_{test}$

(c) $\mathcal{D}_{synth}$ leaking $\mathcal{D}_{train}$ outliers

Figure 2. 2d Gauss data counter-examples.

Sufficient Anonymization. "Effective anonymization reduces identifiability risk to a sufficiently remote level" (ICO UK, 2021). Assessing identifiability involves considering practical factors like cost, time, and technology, focusing on what is reasonably likely to be used rather than any theoretical possibility. We mainly focus on two key technical risks for sufficient anonymization (A29WP, 2014; ICO UK, 2021): i) (*singling out*) isolating any individual, and ii) (*linkability*) combining records/datasets with synthetic data to identify an individual. Last, ICO UK (2021) proposes the *motivated intruder* test, suggesting that the risks should be assessed based on whether a competent intruder, with appropriate resources, could achieve identification if motivated enough.

Related Work. For more detailed discussion, refer to: synthetic data (Jordon et al., 2022; De Cristofaro, 2024), SBPMs (Boudewijn et al., 2023; Ganev & De Cristofaro, 2023; Desfontaines, 2024), and regulation (López & Elbi, 2022a;b; Gal & Lynskey, 2023; Ganev, 2023).

3. Fundamental Issues of SBPMs

We identify several fundamental issues with using SBPMs to reason about privacy through empirical pass/fail tests.

No Theoretical Guarantees. SBPMs lack a defined threat model or strategic adversary, which ignores essential security (Anderson, 2020) and regulatory principles like the motivated intruder test. Instead, they rely on arbitrarily chosen average-case statistics and held-out datasets, falling into the "Generalization Implies Privacy" fallacy (Del Grosso et al., 2023), where generalization is average-case issue but privacy is a worst-case. Thus, even if a model passes all tests and generalizes, it can still memorize data (Song et al., 2017). Consequently, SBPMs offer no theoretical guarantees and are vulnerable to adversarial attacks. Moreover, it is unclear whether or how the SBPMs correspond to the two technical risks – singling out and linkability.

Privacy as Binary Property. SBPMs treat privacy leakage as binary, assuming one synthetic dataset is as safe as many (if the tests pass) even though $\mathcal{D}_{train}$ needs to be queried at each release. However, the "Fundamental Law of Information Reconstruction" (Dwork & Roth, 2014) warns that too many accurate answers can severely compromise privacy.

Privacy as Data Property. SBPMs see privacy as a property of the data, not of the generative model/process, which requires running the tests for each synthetic dataset. This can lead to inconsistent results across generation runs. Also, this increases privacy leakage as well as singling out and linkability risks.

Non-Contrastive Process. SBPMs do not compare computations with or without an individual. Without noise or randomness, the system becomes vulnerable to attacks like differencing attacks, which, in turn, rules out plausible deniability and increases singling out concerns.

Incorrect Interpretation. The tests can be misinterpreted, as failing to reject the null hypothesis ("privacy is preserved") does not confirm that privacy is indeed preserved.

Practical Issues. Most SBPMs implementations require discretizing the data, leading to imprecise calculations and overstated privacy protections. Also, the typical need for a 50/50% train-test split can hurt the synthetic data quality.

4. SBPMs Counter-Examples

We present three counter-examples showing the unreliability and inconsistency of SBPMs. For all of them, we use a toy dataset, *2d Gauss*, which consists of 2,000 points (split evenly between $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$) drawn from a standard normal 2d distribution with no correlation (see Fig. 2(a)). Approximately 10% of the records, those outside the blue circle, are considered outliers.

Leaking All Test Data. Assume $\mathcal{D}_{synth}$ is an exact replica of $\mathcal{D}_{test}$ (Fig. 2(b)). All privacy tests pass as $d_{synth} = d_{test}$. Naturally, publishing half of the personal records cannot be considered regulatory compliant as this would leak an immense amount of private data and would fail the two technical risks – singling out and linkability.

Leaking Train Data Outliers. Next, assume that $\mathcal{D}_{synth}$ contains all $\mathcal{D}_{train}$ outliers with small perturbations (the purple stars outside the circle in Fig. 2(c)) and numerous copies of the value (0, 0). Again, all tests pass: there are no exact matches, and the (0, 0) values skew the distances enough to trick both DCR and NNDR. Publishing $\mathcal{D}_{synth}$ will leave the outlier individuals unconvinced that their privacy is preserved (UK ONS, 2018; UK ICO, 2022) and would again fail the two technical risks.

SBPMs Inconsistency. We assume access to an oracle with knowledge of the generative process of $\mathcal{D}$. Using the oracle, we sample 1,000 new $\mathcal{D}_{synth}$s. Since no generative model was trained, i.e., $\mathcal{D}_{train}$ was never exposed, its privacy is preserved. However, on only 274 occasions do all privacy tests pass, showing that the SBPMs fail to accurately capture the generating process.

The individual metric pass rates are 1 for IMS, 0.48 for DCR, and 0.38 for NNDR, revealing significant inconsistency. Despite sampling from a fixed distribution, the metrics do not agree with each other while DCR and NNDR behave randomly. Even if G captures the underlying process without overfitting or memorization, the pass/fail outcome is sample-dependent, noisy, and unreliable.

Alternatively, if $\mathcal{D}_{synth}$ is fixed and $\mathcal{D}$ is randomly split into $\mathcal{D}_{train}$/$\mathcal{D}_{synth}$, we again run into inconsistencies. Out of 1,000 such repetitions, only 380 cases pass all three tests. This yet again demonstrates the inherent randomness and unreliability of the data-dependent train/test split and, in general, reasoning about privacy through SBPMs.

5. Possible Countermeasures

We discuss and disprove the efficacy of three intuitive solution intended to overcome the limitations of SBPMs.

DP Generative Models. The established framework to limit the ability of an attacker to exploit privacy leakage from trained models is to train them while satisfying DP. Inspired by product deployments by synthetic data providers, we assume access to a single DP trained generative model and unperturbed metrics (per generation run).

In this scenario, although the likelihood of the model memorizing and reproducing real data records would be reduced, leakage can still occur when multiple $\mathcal{D}_{synth}$s are released along with the metrics. The leakage comes from the privacy metrics themselves; since they require access to $\mathcal{D}_{train}$ and are deterministic (ruling out plausible deniability), they compromise the end-to-end DP pipeline. Strategic and motivated adversaries could exploit this vulnerability to reconstruct real data records (Ganev & De Cristofaro, 2023). Adding additional privacy mechanisms on top of the metrics is unlikely to mitigate the problem, as the overall privacy integrity of the entire system needs to be carefully considered.

DP-fying the Metrics. Another possible solution could be to apply DP to the metrics. However, this would not be robust. Implementing DP to the metrics would require additional privacy budget for each generation run, which contradicts one of the main claimed advantages of adopting synthetic data: the ability to generate unlimited data.

Disabling Metrics Access. Finally, not disclosing the privacy metrics while still conducting statistical pass/fail tests would create significant issues. First, users and customers would have to blindly trust the provider that the synthetic data meet a certain threshold. Second, it would undermine a key selling point of providers: providing a tangible measure of compliance, essential for the product's transparency and explainability. Additionally, with statistical pass/fail tests, sensitive information could still be vulnerable to present/future privacy attacks.

6. Conclusion

In this paper, we argue that SBPMs cannot ensure regulatory compliance of synthetic data. SBPMs do not protect against singling out and linkability and, among other fundamental issues, completely ignore the motivated intruder test.

In (Ganev & De Cristofaro, 2023), we discuss further fundamental issues of SBPMs, provide more SBPMs counter-examples, propose a novel reconstruction attack, which is capable of recovering the majority of $\mathcal{D}_{train}$ outliers, and argue that training DP generative models without access to SBPMs addresses their issues (note that DP comes with its own disadvantages, which we also discuss).

Empirical Evaluations. Privacy attacks and empirical evaluations should not be overlooked as they play a crucial role in identifying flaws, errors, and bugs in algorithms and implementations. They contribute significantly to model auditing (Jagielski et al., 2020; Nasr et al., 2023; Annamalai et al., 2024; Ganev et al., 2024) and improve the interpretability of theoretical privacy protections (Houssiau et al., 2022a;b).

Acknowledgements

We would like to thank Emiliano De Cristofaro for reviewing a draft of this paper and providing technical feedback.

References

A29WP. Opinion on anonymisation techniques. https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf, 2014.

Abowd, J. M., Ashmead, R., Cumings-Menon, R., Garfinkel, S., Heineck, M., Heiss, C., Johns, R., Kifer, D., Leclerc, P., Machanavajjhala, A., et al. The 2020 census disclosure avoidance system topdown algorithm. *Harvard Data Science Review*, 2022.

Anderson, R. *Security engineering: a guide to building dependable distributed systems*. John Wiley & Sons, 2020.

Annamalai, M. S. M. S., Ganev, G., and De Cristofaro, E. "What do you want from theory alone?" Experimenting with Tight Auditing of Differentially Private Synthetic Data Generation. *USENIX Security*, 2024.

Borisov, V., Sessler, K., Leemann, T., Pawelczyk, M., and Kasneci, G. Language Models are Realistic Tabular Data Generators. In *ICLR*, 2023.

Boudewijn, A. T. P., Ferraris, A. F., Panfilo, D., Cocca, V., Zinutti, S., De Schepper, K., and Chauvenet, C. R. Privacy Measurements in Tabular Synthetic Data: State of the Art and Future Research Directions. In *NeurIPS SyntheticData4ML*, 2023.

De Cristofaro, E. Synthetic Data: Methods, Use Cases, and Risks. *IEEE S&P*, 2024.

Del Grosso, G., Pichler, G., Palamidessi, C., and Piantanida, P. Bounding information leakage in machine learning. *Neurocomputing*, 2023.

Desfontaines, D. Empirical privacy metrics: the bad, the ugly… and the good, maybe? https://desfontain.es/blog/bad-ugly-good-maybe.html, 2024.

Dwork, C. and Roth, A. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 2014.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *TCC*, 2006.

EP and Council. Article 4 GDPR Definitions. https://gdpr-info.eu/art-4-gdpr/, 2016a.

EP and Council. Recital 26 EU GDPR. https://www.privacy-regulation.eu/en/recital-26-GDPR.htm, 2016b.

Gal, M. and Lynskey, O. Synthetic Data: Legal Implications of the Data-Generation Revolution. *109 Iowa Law Review*, 2023.

Ganev, G. When synthetic data met regulation. *GenLaw*, 2023.

Ganev, G. and De Cristofaro, E. On the Inadequacy of Similarity-based Privacy Metrics: Reconstruction Attacks agains "Truly Anonymous Synthetic Data". *arXiv:2312.05114*, 2023.

Ganev, G., Annamalai, M. S. M. S., and De Cristofaro, E. The Elusive Pursuit of Replicating PATE-GAN: Benchmarking, Auditing, Debugging. *arXiv:2406.13985*, 2024.

Hod, S. and Canetti, R. Differentially Private Release of Israel's National Registry of Live Births. *arXiv:2405.00267*, 2024.

Houssiau, F., Cohen, S. N., Szpruch, L., Daniel, O., Lawrence, M. G., Mitra, R., Wilde, H., and Mole, C. A Framework for Auditable Synthetic Data Generation. *arXiv:2211.11540*, 2022a.

Houssiau, F., Jordon, J., Cohen, S. N., Daniel, O., Elliott, A., Geddes, J., Mole, C., Rangel-Smith, C., and Szpruch, L. TAPAS: a toolbox for adversarial privacy auditing of synthetic data. In *NeurIPS SyntheticData4ML*, 2022b.

ICO UK. Chapter 2: how do we ensure anonymisation is effective? https://ico.org.uk/media/about-the-ico/documents/4018606/chapter-2-anonymisation-draft.pdf, 2021.

Jagielski, M., Ullman, J., and Oprea, A. Auditing Differentially Private Machine Learning: How Private is Private SGD? *NeurIPS*, 2020.

Jordon, J., Yoon, J., and Van Der Schaar, M. PATE-GAN: generating synthetic data with differential privacy guarantees. In *ICLR*, 2019.

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. N., and Weller, A. Synthetic Data–what, why and how? *arXiv:2205.03257*, 2022.

Kotelnikov, A., Baranchuk, D., Rubachev, I., and Babenko, A. Tabddpm: Modelling tabular data with diffusion models. In *ICML*, 2023.

López, C. A. F. and Elbi, A. On the legal nature of synthetic data. In *NeurIPS SyntheticData4ML*, 2022a.

López, C. A. F. and Elbi, A. On synthetic data: a brief introduction for data protection law dummies. *EU Law Blog*, 2022b.

Lu, P.-H., Wang, P.-C., and Yu, C.-M. Empirical evaluation on synthetic data generation with generative adversarial network. In *WIMS*, 2019.

McKenna, R., Miklau, G., and Sheldon, D. Winning the NIST Contest: A scalable and general approach to differentially private synthetic data. *JPC*, 2021.

Microsoft. South Australian Health Partners with Gretel to Pioneer State-Wide Synthetic Data Initiative for Safe EHR Data Sharing. https://startups.microsoft.com/blog/south-australian-health-synthetic-data-safe-ehr-data-sharing/, 2024.

Mobey Forum. Help Me Understand: AI-Generated Synthetic Data. https://mobeyforum.org/download/?file=AI-generated-synthetic-data-report-2.pdf, 2022.

Mostly AI. Truly Anonymous Synthetic Data – Evolving Legal Definitions and Technologies (Part II). https://mostly.ai/blog/truly-anonymous-synthetic-data-legal-definitions-part-ii/, 2020.

Nasr, M., Hayes, J., Steinke, T., Balle, B., Tramèr, F., Jagielski, M., Carlini, N., and Terzis, A. Tight Auditing of Differentially Private Machine Learning. In *USENIX Security*, 2023.

Panfilo, D. and Aindo. *Generating Privacy-Compliant, Utility-Preserving Synthetic Tabular and Relational Datasets Through Deep Learning*. University of Trieste, 2022.

Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., and Kim, Y. Data Synthesis Based on Generative Adversarial Networks. *PVLDB*, 2018.

Platzer, M. and Reutterer, T. Holdout-based empirical assessment of mixed-type synthetic data. *Frontiers in Big Data*, 2021.

Song, C., Ristenpart, T., and Shmatikov, V. Machine learning models that remember too much. In *ACM CCS*, 2017.

Syntegra. Fidelity and privacy of synthetic medical data. *arXiv:2101.08658*, 2021.

Syntho. Syntho's quality assurance report. https://www.syntho.ai/synthos-quality-assurance-report/, 2023.

UK ICO. Chapter 5: privacy-enhancing technologies (PETs). https://ico.org.uk/media/about-the-ico/consultations/4021464/chapter-5-anonymisation-pets.pdf, 2022.

UK ICO. Synthetic data to test the effectiveness of a vulnerable person's detection system in financial services. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/privacy-enhancing-technologies/case-studies/synthetic-data-to-test-the-effectiveness-of-a-vulnerable-persons-detection-system-in-financial-services/, 2023.

UK ONS. Privacy and data confidentiality methods: a data and analysis method review. https://analysisfunction.civilservice.gov.uk/policy-store/privacy-and-data-confidentiality-methods-a-national-statisticians-quality-review-nsqr/, 2018.

UK ONS. Synthesising the linked 2011 Census and deaths dataset while preserving its confidentiality. https://datasciencecampus.ons.gov.uk/synthesising-the-linked-2011-census-and-deaths-dataset-while-preserving-its-confidentiality/, 2023.

Xie, L., Lin, K., Wang, S., Wang, F., and Zhou, J. Differentially private generative adversarial network. *arXiv:1802.06739*, 2018.

Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. Modeling tabular data using conditional gan. *NeurIPS*, 2019.

Yoon, J., Mizrahi, M., Ghalaty, N. F., Jarvinen, T., Ravi, A. S., Brune, P., Kong, F., Anderson, D., Lee, G., Meir, A., Bandukwala, F., Kanal, E., Arik, S. O., and Pfister, T. EHR-Safe: generating high-fidelity and privacy-preserving synthetic electronic health records. *NPJ Digital Medicine*, 2023.

Zhang, H., Zhang, J., Shen, Z., Srinivasan, B., Qin, X., Faloutsos, C., Rangwala, H., and Karypis, G. Mixed-Type Tabular Data Synthesis with Score-based Diffusion in Latent Space. In *ICLR*, 2024.

Zhang, J., Cormode, G., Procopiuc, C. M., Srivastava, D., and Xiao, X. Privbayes: Private data release via bayesian networks. *ACM TODS*, 2017.

Zhao, Z., Kunar, A., Birke, R., and Chen, L. Y. Ctab-gan: effective table data synthesizing. In *ACML*, 2021.