Synthetic Data: Generate Avatar Data on Demand

Thomas Lebrun, Louis Béziaud, Tristan Allard, Tristan Allard, Antoine Boutet, Sébastien Gambs, Mohamed Maouche

Synthetic Data: Generate Avatar Data on Demand

Thomas Lebrun[1 [0000−0002−8463−4793]], Louis Béziaud[2,3 [0000−0002−4974−3492]], Tristan Allard[2 [0000−0002−2777−0027]], Antoine Boutet[1 [0000−0002−4057−416X]], Sébastien Gambs[3 [0000−0002−7326−7377]], and Mohamed Maouche[1 [0000−0001−6473−7173]]

[1] Univ Lyon, INSA Lyon, Inria, CITI, Lyon, France
[2] Univ Rennes, CNRS, IRISA
[3] Université du Québec à Montréal, Montréal, Canada

Abstract. Anonymization is crucial for the sharing of personal data in a privacy-aware manner yet it is a complex task that requires to set up a trade-off between the robustness of anonymization (*i.e.*, the privacy level provided) and the quality of the analysis that can be expected from anonymized data (*i.e.*, the resulting utility). Synthetic data has emerged as a promising solution to overcome the limits of classical anonymization methods while achieving similar statistical properties to the original data. Avatar-based approaches are a specific type of synthetic data generation that rely on local stochastic simulation modeling to generate an avatar for each original record. While these approaches have been used in healthcare, their attack surface is not well documented and understood. In this paper, we provide an extensive assessment of such approaches and comparing them against other data synthesis methods. We also propose an improvement based on conditional sampling in the latent space, which allows synthetic data to be generated on demand (i.e., of arbitrary size). Our empirical analysis shows that avatar-generated data are subject to the same utility and privacy trade-off as other data synthesis methods with a privacy risk more important on the edge data, which correspond to records that have the fewest alter egos in the original data.

Keywords: Synthetic data, Avatar-based generation, Privacy, Re-identification

1 Introduction

The collection of personal data has grown to a tremendous proportion and is done through diverse sources such as credit cards, medical records, digital photographs, emails, websites, social media, Internet of Things (IoT), smartphones, wearable technologies, to name a few. All of this data has enormous value for improving the understanding of human behavior and creating useful societal applications, but it also raises serious privacy concerns. For instance, healthcare generates massive amounts of data whose sharing and re-using is essential for accelerating research and to develop machine learning algorithms methods that

can be deployed in clinical settings. However, this data is very sensitive and must be anonymized before it can be used beyond the purpose of its initial collection.

Anonymization is a complex task that requires calibrating a trade-off between the privacy guarantees and the remaining usefulness of anonymized data, which is difficult to control and depends on the data and the analysis considered. Thus in practice, a high privacy protection often results in a limited utility. To overcome this limitation, the use of synthetic data that resemble the real data (*i.e.*, which preserves global statistical properties and task-specific performance) is increasingly recognized as a promising way to enable such reuse while addressing personal data privacy concerns [5]. For example, some projections predict that synthetic data will completely eclipse real data in AI models by 2030[4]. However, there is still no consensus on a standard approach to systematically and quantitatively assess the privacy gain and residual utility of synthetic data, which slow their adoption. Nonetheless, to shed some light on the real guarantees of synthetic data and help hospitals position themselves on this new technology, some papers have started to assess the privacy [2] and utility [25] of synthetic data for medical data analyses.

Recently, new approaches based on avatar data have attracted for generating synthetic patient-data [14]. For each individual observation, this approach identifies the k nearest neighbors in a latent space and leverages this neighborhood to generate an avatar through a local stochastic modeling. While appealing these avatar-based approaches lack a proper privacy assessment [17]. To overcome this limitation, in this paper we present an extensive utility and privacy assessment of avatar data based on a wide variety of metrics. More precisely, we quantify the privacy of synthetic data through criteria used to evaluate anonymization schemes according to the GDPR, namely singling-out, linkability and inference. In addition, we have also implemented a re-identification attack (*i.e.*, mapping a synthetic data record to a close original data record) and a membership inference attack (*i.e.*, inferring data records leveraged to generate a synthetic dataset), and focus on the most vulnerable data. We evaluate the utility and compared the avatar approach to different synthetic data generation methods as well as anonymization schemes. Our main objective is to provide a comprehensive assessment of the utility and privacy of avatar data to subsequently facilitate their use in the medical field under the best conditions. We also propose an improvement is based on conditional sampling in the latent space, which allows synthetic data to be generated on demand (i.e., of arbitrary size), and which depicts utility and privacy trade-off aligned with the state-of-the-art.

2 Related Work

Anonymization and synthetic data generation. Historically, the sharing of personal data was carried out through anonymization, which is the process of transforming data records or datasets to ensure that the individuals to whom the

[4] https://www.gartner.com/en/newsroom/press-releases/2022-06-22-is-synthetic-data-the-future-of-ai

data pertains are not identifiable. This risk of isolating or linking data records comes from the fact that human behavior depicts a strong uniqueness [7]. Common approaches from the family of k-anonymity [24] (*e.g.*, t-closeness and l-diversity) include suppressing highly sensitive records and generalizing data to increase overlap and avoid unique entries. However, these methods, particularly when applied to medical data, often succeed in protecting privacy at the cost of degrading data quality to such an extent that their utility is compromised. Another approach involves perturbation methods, which introduces noise to the data, with Differential Privacy (DP) [8] being a prominent example that offers theoretical guarantees on privacy protection. These methods are particularly successful to protect aggregate queries or statistics but reduces the utility too much when applied individually on data.

More recently, the growing demand for extensive datasets has shifted focus towards generative methods to create large amounts of synthetic data. Synthetic data generation encompasses a wide range techniques used to create artificial datasets that mimic the statistical properties of real-world data while breaking the link to individuals in the real data records. Initially, these techniques were motivated by overcoming data scarcity in some domains in which there may be limited amount of real data. For instance, they are particularly interesting in the context of model training as they permit the construction of additional data points to help address overfitting or data imbalance [11]. In addition, they also represent an opportunity for further testing and validation through various scenarios. These techniques can also be used to address confidentiality requirements that some real-world data have, especially in the context in which the opening of data and the reproducibility of science are essential. In this context, various techniques have been adapted to address confidentiality and privacy, such as with autoencoders and GANs. This includes `CT-GAN` [27], which has extended conditional GANs to tabular data. Some techniques have been purely based on characterizing statistics of various features (e.g., the R package `SynthPop` [21]).

Unfortunately, those techniques opened up the way for privacy attacks. This limitation has motivated the design of methods that are privacy-preserving by design. This include training GANs with DP guarantees such as in DP-CTGAN [10], PATE-Gan [16], and PrivBayes [28]. Finally, another example of DP synthetic generation method is the Maximum Spanning Tree (`MST`) [19] that has won the 2018 NIST differential privacy synthetic data. `MST` is based on a marginal estimation approach that produces differentially private data before using a probabilistic graphical-model [20] to generate synthetic data.

Privacy Assessment. Privacy is a multifaceted concept that depends on the context and the data considered, resulting in multiple existing metrics [26] and similarity tests [12]. Among the most significant guidelines for characterizing anonymization and verifying its success is the opinion from the Article 29 of the GDPR[5], which proposes to quantify the robustness of anonymization tech-

[5] `https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf`

niques by assessing the risks of singling-out, linkability and attribute inference. Singling-out can be seen as a way to indirectly identify a person in a dataset, such as recognizing a unique combination of rare attributes in a medical record. Linkability refers to the ability to map anonymized records to the same individual, for instance, linking a de-identified hospital visit record with an anonymized prescription record. And attribute inference deals with the possibility to learn sensitive information about the individual beyond what is disclosed, such as deducing a patient's HIV status from their medication regimen and frequent visits to a specialist. In this context, `Anonymeter` [13] derived specific attacks for each privacy criterion mentioned in the above. For singling-out, the objective is to assess how well the predicates that isolate a row in the synthetic data also isolate a row in the original data. For linkability, the original dataset is divided horizontally, and the assessment measures how much accessing the synthetic data allows the re-linking of rows. And for attribute inference (AIA), a feature is hidden in the original dataset, and the evaluation considers how well the synthetic data can reveal this hidden attribute. An important safeguard implemented in `Anonymeter` is not to evaluate the success of these attacks in an absolute sense due to the fact that one might perform inferences due to strong correlations between features (*e.g.*, smoking causes cancer) rather than specific data instances. Instead, they use a control dataset not seen during the generation of synthetic data to evaluate how much advantage an adversary gains from accessing the synthetic data (*e.g.*, comparison between identifying the cancer of a patient from its synthetic data versus identifying this information from the control data).

Another approach to evaluate the privacy of trained models is to rely on a Membership Inference Attack (MIA) [3]. When a model is trained on a dataset D, it is possible that an adversary with access to the output of the model (or the model itself) can learn information about the dataset D. More precisely, an MIA aims to determine whether a specific data record was used by the model. If an adversary is not able to detect the usage of a particular data record by the model, then it will likely not be be able to extract more complex information. Thus, preventing this form attack could consequently also prevents stronger attacks. While they are many approaches to perform a MIA, one of the simplest one is to design a threshold-based attack that exploits the prediction vector to distinguish between member and non-members (*i.e.*, the model is more confident on data it has already seen during training).

Although the risk is often reported globally across the entire original dataset, it is important to note that not all profiles have the same level of risk as some are more vulnerable than others. To better reflect the privacy risk, recent works have proposed to focus on the most vulnerable profiles to quantify the upper bound of the risk for privacy [12, 23, 3].

3 Generate Avatar Data on Demand

The `Avatar` method [14] has been designed for biomedical analysis from tabular data. The original dataset is composed of n entries of p variables. Each

entry represents an individual and each variable can be continuous, categorical, Boolean or represent a date. The `Avatar` method aims to create a new dataset of n synthetic observations and p variables with consistent yet different values compared with those of the original dataset. To achieve this, `Avatar` relies on three main steps: 1) the profile of each individual is projected into a latent space using a factor analysis technique (*e.g.*, PCA) ; 2) using the first d dimensions of this space, pairwise distances are calculated between all projections associated with the individuals' data to find the k nearest neighbors ; 3) for each individual, a single avatar is created by pseudo-stochastically weighting the attributes of its k nearest neighbors using all of the dimensions of the latent space. Synthetic data are then shuffled to change the order between the original individuals and the avatars. `Avatar` is not the only method exploiting the neighborhood as for instance the Local Linear Embedding (LLE) [4] first computes the nearest neighbors before doing the projection in an embedding.

Although the `Avatar` method depicts an interesting trade-off between utility and privacy, several issues remain. More specifically, the evaluation of privacy is carried out globally and with ad-hoc metrics, which does not make it possible to properly capture the real risk for certain atypical and vulnerable individuals. To improve the utility and privacy trade-off, the value of k could also be dynamically defined according to the context of each data point to adapt the utility and privacy trade-off for each of them and thus limit the degradation for profiles which are already well protected because they are located in a dense area. The most limiting aspect of `Avatar` is the fact that the input data has the same size as the output data and that each avatar comes from a single original data and its neighborhood. This bijective nature opens up the risk of re-identification (*i.e.*, mapping an avatar to an original data), which is not the case when a model is build and then exploited to generate synthetic data of arbitrary size.

To overcome this limitation, we propose `M-Avatar` an alternative method which builds a global model that makes it possible to generate synthetic data on demand, while removing the constrains of producing one avatar data for each original profile. To achieve this goal, we first construct the data distribution of the projections in the first d dimensions of the latent space. Afterwards to generate synthetic data, we first sample a value in the distribution of the first dimension of the latent space before building the conditional distribution to this sample in the second dimension (we consider a bucket gathering 10% of data around the sample) and sample again a value in this distribution. This operation is then repeated for the first d dimensions. This conditional construction of the distribution makes it possible to preserve the neighborhood information in the first dimensions that contain the most information by sampling in dimension d_i a value consistent with the sample chosen in dimension $d_i - 1$. For dimensions greater than d, the quantity of data respecting the constraints of previous sampling being considerably reduced, sampling from a distribution that is too sparse would reduce the utility too much. To avoid this, we randomly choose a value among the projection values of the original data at the considered dimension. This random choice makes it also possible to mix the influence of different data

while maintaining a good level of utility. The closest state-of-the-art method is Local Resampler [18], which samples locally from the original data distribution (compared to `M-Avatar` which conditionally samples from each dimension of the latent space) to create an avatar data.

4 Evaluation

Dataset. We consider a real-world dataset covering Acquired Immunodeficiency Syndrome (`AIDS`). This dataset gathers 2,139 patients and 26 variables for HIV-infected patients who participated in a clinical trial published in 1996 in the New England Journal of Medicine [15].

Evaluation metrics. There are numerous ways to evaluate synthetic data [9, 6, 1] such as utility metrics that measure the quality of synthetic data and its ability to faithfully reproduce the original data as well as privacy measures, which quantifies the leakage of personal information. More precisely, to evaluate the utility, we considered the SDV quality score [22], which captures the overall assessment of synthetic data's quality, combining various aspects like statistical similarity, data characteristics, and correlations between pairs of attributes. We also considered the prediction accuracy of the synthetic data by examining the performance of a learning model trained with original data or trained with the synthetic data (i.e., the task accuracy to predict if HIV has progressed to AIDS.). In addition, we considered the survival curve, a healthcare metric adapted for this dataset. To assess the privacy guarantees, we leverage on `Anonymeter` [13] (*cf.* Section 2) for singlingo out, linkability and inference. For all these three attacks, the risk assessment quantifies whether an adversary has an advantage in attacking a person that participated in the construction of the synthetic data (*i.e.*, leads to a leak of personal information) compared to attacking a person from the general population (*i.e.*, control dataset). Finally, we have also implemented a re-identification and a membership inference attack. For each avatar data, a re-identification is inferred with the original data whose projection in the latent space is closest to the avatar's projection. Similarly, the c original data whose projections are closest to an avatar's projection are inferred as members. The value of c varies depending on the data density from 1 (*i.e.*, as for re-identification) for dense data, to 20 for edge data. As for MIA, we perform a PCA on the synthetic data, reducing it to five dimensions, and project both the real data (members and non-members) and the synthetic data. For each synthetic data point, we identify its k nearest real data points (k varying from 1 to 20 based on distance quantiles to the barycenter) and increment their MIA risk scores by +1. Finally, the attack predicts the top 50% as members.

Comparative baselines. We evaluate the avatar data against the following alternatives. `SAIPH`[6]: First, we consider a solution that projects the original point into a low-dimensional latent space (like the one used by `Avatar`, with a dimension limited to 20) and reconstruct the data point in the original space from

[6] `https://github.com/octopize/saiph`

this projection. Indeed, passing through this subspace compresses the information and induces a loss of utility. Unlike `Avatar`, it does not exploit the nearest neighbors nor the full latent dimensions to generate synthetic points.

The `MST` [19] algorithm came first in the 2018 NIST Differential Privacy Synthetic Data Competition[7]. It consists of three steps: (1) select a collection of low-dimensional marginals, (2) measure these marginals with an additional noise (we considered for our experiments $\epsilon = 3$) and (3) generate synthetic data that preserve well the noisy marginals. `SynthPop` [21] synthetizes data from the conditional distributions. Variables are synthesised one-by-one using sequential regression modelling and are conditioned on the original variables that are earlier in the synthesis sequence. `CT-GAN` [27] uses a conditional generative adversarial network to generate synthetic tabular data that contains a mix of discrete and continuous columns. `K-anonymity` [24] is not a data generation scheme but rather a data anonymization technique that is used to protect individuals' privacy in a dataset. A dataset is considered k-anonymous when, for every combination of identifying attributes in a dataset, there are at least $k-1$ other people with the same attributes (i.e., $k = 20$ here).

Methodology. To conduct the experiments, we followed the protocol outlined below. First, we have split the data into two equal-sized sets (50-50). The first set, referred to as "original data", is used to generate a synthetic dataset of the same size while the second set, the "control data" is kept aside to compute the baseline metrics. Thus, the creation of synthetic data is not influenced by the control data. For both utility and privacy, the control data is used to ensure that we are exclusively evaluating the impact of the synthetic data generation method. The cross-validation has been performed over 25 iterations. Thus, each metric result represents the average of 25 evaluations on different original/control splits for a given generation method of synthetic data. We used the API of Octopize to generate avatar data[8] with $k = 20$ and $d = 5$.

4.1 Understanding the data topology

In this section, we first aim at analyzing the topology and the relationship between both the original and the avatar data. To achieve this, we measure the distance of each original and avatar data to the barycenter of the data, focusing in particular our attention on the edge data. We consider the Gower and the Euclidean distance for the original and the latent space, respectively. Not reported due to space limitation, we observe that the distributions of the data centroid (*i.e.*, the barycenter) of the original data as well as the avatar data are similar and contain a long tail showing that only a few data that are far from the barycenter. By analysing these distributions, we also observe that the data at the edge in the original data tends to remain at the edge in the avatar data, and vice versa. As these edge data are easily distinguishable and in small numbers, that makes them more vulnerable to re-identification (Section 4.3).

[7] `https://www.nist.gov/ctl/pscr/open-innovation-prize-challenges/past-prize-challenges/2018-differential-privacy-synthetic`

[8] `https://www.octopize.io/`

	Avatar	SAIPH	M-Avatar	CT-GAN	SynthPop	MST	K-anonymity	Orig. Data
SDV Score	0.917	0.789	0.850	0.830	**0.935**	0.836	0.584	1.000
Task Acc.	0.820	0.640	0.734	0.501	0.637	**0.930**	0.533	0.997
Linkability	0.306	**0.036**	**0.018**	**0.010**	**0.027**	**0.024**	**0.015**	0.987
Singling-out	0.020	**0.007**	**0.009**	0.017	0.032	**0.009**	**0.006**	0.992
MIA	0.593	0.551	**0.492**	**0.492**	**0.498**	**0.502**	**0.501**	0.751
AIA	0.295	0.075	0.050	**0.023**	0.074	0.042	0.043	0.957

Table 1. Utility and privacy metrics comparison between the different baselines.

4.2 Quantifying the utility loss

First, we assess the impact of the size of the latent space for SAIPH on the utility. As described in Section 4, SAIPH only involves projecting the original data point into a latent space and then projecting back to its original space. The survival curve (not reported here for space reason) according to a growing size of the latent space, from 2 to 20 over the 26 dimensions of the original data. As expected, the larger the size of the latent space, the closer the survival curve is to the one from the original data. We notice that to reconstruct the curve properly, we need at least 20 out 26 dimensions. Meaning, that the AIDS dataset does not contain too many redundant dimensions.

We have also compared the survival curve from the avatar data against the data from other comparative baselines (Figure 1). The results obtained show that both SynthPop and M-Avatar produce a survival curve that closely matches the one from the original data. Conversely, the results show that the data from CT-GAN and K-anonymity provide survival indicators that are not usable. Similarly, MST also strongly deteriorates the survival rate. The survival curve from the avatar data and from the SAIPH latent space is slightly impacted with the difference between these two curves coming from the exploitation of local neighbors for the generation of avatar data. Taking advantage of this neighborhood improves the fidelity of the survival curve compared to the original data.

To evaluate the impact on statistical properties (*e.g.*, statistical similarity, data characteristics and correlations between attributes), we then compute the SDV average quality score. Table 1 reports this quality score for the avatar data and for the other comparative baselines. The results show that apart from K-anonymity, which clearly deteriorates the statistical properties of the data, all other approaches maintain an SDV quality score close to 0.7, in which a score of 0.95 is achieved with data close to the original data.

Finally, to evaluate the impact on predictive tasks, we compare the accuracy of the classification of a Random Forests model trained from original data compared to one trained on synthetic data. Table 1 displays the balanced accuracy of the classifier trained from all the considered synthetic data generation schemes. Results show that the balanced accuracy provided by MST is close to the accuracy from the original data. The Avatar approach is just behind with a balanced accuracy slightly higher than 0.8, followed by the other methods.

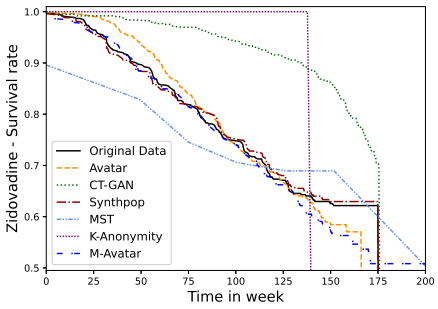

Fig. 1. The survival curve provided by `SynthPop` and `M-Avatar` are very close to the one obtained with the original data.

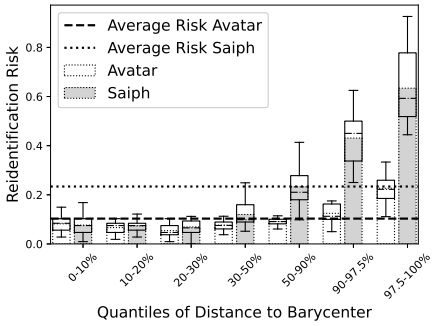

Fig. 2. The risk of re-identification is much more important for avatar data at the edge.

4.3 Measuring the privacy gain

In this section, we evaluate the privacy gain brought by synthetic data methods. Specifically, we quantify the privacy risk associated with the disclosure of synthetic data against a singling-out, linkage, attribute inference, re-identification and membership inference attack. Table 1 depicts the risk of inference for `AIDS`. We notice that the solution that displays the highest risk is the `Avatar` approach (privacy risk around 0.3), followed by the `K-anonymity` solution (privacy risk close to 0.2). We believe that the high risk for the avatar data comes from the fact that both `Avatar` and the implementation of the attack exploit neighborhood information. The other baselines and `M-Avatar` display an inference risk level below 0.1.

Table 1 displays the risk of singling-out and the risk of linkability. Results show that the risk of singling-out remains very low for all baselines and `M-Avatar`, which means that all these baselines significantly reduce the uniqueness of synthetic data compared to the original data which are highly unique. The results also show that the risk of linkability remains very limited for all baselines except for `Avatar`. The high risk for this approach comes from the fact that this attack (as `Avatar`) leverages the closest neighbors to infer the linkability.

Then, we evaluate the risk of re-identification according to the distance of the avatar data to the barycenter (Figure 2). As explained in Section 4.1, the original data which is at the edge tends to remain at the edge also in the avatar data. The results show that the avatar's edge data is more likely to be re-identified than the data in the center of the point cloud. More precisely, the edge data in the last quantile (*i.e.*, more distinguishable) exhibits a risk close to 30% while data belonging to the densest part (*i.e.*, less distinguishable) displays a re-identification risk of 8%. As it is more easy to identify edge data, an average risk of re-identification (here the dotted line close to 10%) does not sufficiently reflect the real risk of re-identification. However, it should be noted that this risk is an average, a data re-identified on a run, will not necessarily be re-identified on another run due to the stochasticity nature of the neighborhood. For instance, in

the last quantile, a data is re-identified 1 time out of 10 (results not depicted for space reason). Finally, Tabular 1 reports the risk of membership inference for all synthetic data generation methods. The results demonstrate that only `Avatar` and `SAIPH` introduce a risk, while the others including `M-Avatar` significantly reduce this risk.

5 Conclusion

In this paper, we have conducted an in-depth utility and privacy assessment of the avatar based approaches. We have found that edge data in the original data tends to remain at the edge in the avatar data, which favors the probability of being re-identified compared to data that is less distinguishable. We also propose an alternative method (called `M-Avatar`) based on conditional sampling in the latent space, which allows synthetic data to be generated on demand. Specifically, by removing the bijective nature of avatar data (i.e., a raw data produces an avatar, a constraint that only concerns certain use cases), `M-Avatar` can generated synthetic data of arbitrary size. Finally, in terms of utility and privacy compromise, `MST`, `SynthPop`, and the proposed `M-Avatar` solution comes out on top in our comparison.

Acknowledgment: We would like to thank Octopize for access to their API. This work has been supported by the ANR 22-PECY-0002 IPOP (Interdisciplinary Project on Privacy) project of the Cybersecurity PEPR, the Trusty-IA project supported by the Auvergne Rhône-Alpes region, and the Canada Research Chair program through a Discovery Grant from the NSERC and the DEEL Project CRDPJ 537462-18 funded by the NSERC and the CRIAQ.

References

1. Alaa, A., Van Breugel, B., Saveliev, E.S., van der Schaar, M.: How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In: ICML. pp. 290–306 (2022)
2. Appenzeller, A., Leitner, M., Philipp, P., Krempel, E., Beyerer, J.: Privacy and utility of private synthetic data for medical data analyses. Applied Sciences **12**(23) (2022)
3. Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., Tramer, F.: Membership inference attacks from first principles. In: Security & Privacy. pp. 1897–1914 (2022)
4. Chen, J., Liu, Y.: Locally linear embedding: a survey. Artificial Intelligence Review **36**, 29–48 (2011)
5. Chen, R.J., Lu, M.Y., Chen, T.Y., Williamson, D.F., Mahmood, F.: Synthetic data in machine learning for medicine and healthcare. Nature Biomedical Engineering **5**(6), 493–497 (2021)
6. Dankar, F.K., Ibrahim, M.K., Ismail, L.: A multi-dimensional evaluation of synthetic data generators. Access **10**, 11147–11158 (2022)
7. De Montjoye, Y.A., Hidalgo, C.A., Verleysen, M., Blondel, V.D.: Unique in the crowd: The privacy bounds of human mobility. Scientific reports **3**(1), 1–5 (2013)

8. Dwork, C., Roth, A., et al.: The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science **9**(3–4), 211–407 (2014)
9. El Emam, K.: Seven ways to evaluate the utility of synthetic data. Security & Privacy **18**(4), 56–59 (2020)
10. Fang, M.L., Dhami, D.S., Kersting, K.: Dp-ctgan: Differentially private medical data generation using ctgans. In: AIME. pp. 178–188 (2022)
11. Fonseca, J., Bacao, F.: Tabular and latent space synthetic data generation: a literature review. Journal of Big Data **10**(1), 115 (2023)
12. Ganev, G., Cristofaro, E.D.: On the inadequacy of similarity-based privacy metrics: Reconstruction attacks against "truly anonymous synthetic data" (2023)
13. Giomi, M., Boenisch, F., Wehmeyer, C., Tasnádi, B.: A unified framework for quantifying privacy risk in synthetic data. PETS (2023)
14. Guillaudeux, M., Rousseau, O., Petot, J., Bennis, Z., Dein, C.A., Goronflot, T., Vince, N., Limou, S., Karakachoff, M., Wargny, M., Gourraud, P.A.: Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis. npj Digital Medicine **6**(1), 37 (2023)
15. Hammer, S.M., Katzenstein, D.A., Hughes, M.D., Gundacker, H., Schooley, R.T., Haubrich, R.H., Henry, W.K., Lederman, M.M., Phair, J.P., Niu, M., et al.: A trial comparing nucleoside monotherapy with combination therapy in hiv-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. New England Journal of Medicine **335**(15), 1081–1090 (1996)
16. Jordon, J., Yoon, J., van der Schaar, M.: Pate-gan: Generating synthetic data with differential privacy guarantees. In: ICLR (2018)
17. Kaabachi, B., Despraz, J., Meurers, T., Otte, K., Halilovic, M., Prasser, F., Raisaro, J.L.: Can we trust synthetic data in medicine? a scoping review of privacy and utility metrics (2023)
18. Kalay, A.F.: Generating synthetic data with the nearest neighbors algorithm (2022)
19. McKenna, R., Miklau, G., Sheldon, D.: Winning the nist contest: A scalable and general approach to differentially private synthetic data (2021)
20. McKenna, R., Sheldon, D., Miklau, G.: Graphical-model based estimation and inference for differential privacy. In: International Conference on Machine Learning. pp. 4435–4444. PMLR (2019)
21. Nowok, B., Raab, G.M., Dibben, C.: synthpop: Bespoke creation of synthetic data in r. Journal of Statistical Software **74**(11), 1–26 (2016)
22. Patki, N., Wedge, R., Veeramachaneni, K.: The synthetic data vault. In: DSAA. pp. 399–410 (2016)
23. Stadler, T., Oprisanu, B., Troncoso, C.: Synthetic data - anonymisation groundhog day. In: USENIX Security Symposium (2022)
24. Sweeney, L.: k-anonymity: A model for protecting privacy. International journal of uncertainty, fuzziness and knowledge-based systems **10**(05), 557–570 (2002)
25. Vallevik, V.B., Babic, A., Marshall, S.E., Elvatun, S., Brøgger, H.M., Alagaratnam, S., Edwin, B., Veeraragavan, N.R., Befring, A.K., Nygård, J.F.: Can i trust my fake data – a comprehensive quality assessment framework for synthetic tabular data in healthcare. International Journal of Medical Informatics **185**, 105413 (2024)
26. Wagner, I., Eckhoff, D.: Technical privacy metrics: A systematic survey. Computing Surveys **51**(3) (2018)
27. Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling tabular data using conditional gan. In: NeurIPS (2019)
28. Zhang, J., Cormode, G., Procopiuc, C.M., Srivastava, D., Xiao, X.: Privbayes: Private data release via bayesian networks. TODS **42**(4), 1–41 (2017)