frontiers in Big Data

ORIGINAL RESEARCH
published: 29 June 2021
doi: 10.3389/fdata.2021.679939

Holdout-Based Empirical Assessment of Mixed-Type Synthetic Data

Michael Platzer[1]* *and Thomas Reutterer*[2]

[1]*MOSTLY AI, Vienna, Austria,* [2]*Department of Marketing, WU Vienna University of Economics and Business, Vienna, Austria*

AI-based data synthesis has seen rapid progress over the last several years and is increasingly recognized for its promise to enable privacy-respecting high-fidelity data sharing. This is reflected by the growing availability of both commercial and open-sourced software solutions for synthesizing private data. However, despite these recent advances, adequately evaluating the quality of generated synthetic datasets is still an open challenge. We aim to close this gap and introduce a novel holdout-based empirical assessment framework for quantifying the fidelity as well as the privacy risk of synthetic data solutions for mixed-type tabular data. Measuring fidelity is based on statistical distances of lower-dimensional marginal distributions, which provide a model-free and easy-to-communicate empirical metric for the representativeness of a synthetic dataset. Privacy risk is assessed by calculating the individual-level distances to closest record with respect to the training data. By showing that the synthetic samples are just as close to the training as to the holdout data, we yield strong evidence that the synthesizer indeed learned to generalize patterns and is independent of individual training records. We empirically demonstrate the presented framework for seven distinct synthetic data solutions across four mixed-type datasets and compare these then to traditional data perturbation techniques. Both a Python-based implementation of the proposed metrics and the demonstration study setup is made available open-source. The results highlight the need to systematically assess the fidelity just as well as the privacy of these emerging class of synthetic data generators.

Keywords: synthetic data, privacy, fidelity, structured data, anonymization, self-supervised learning, statistical disclosure control, mixed-type data

OPEN ACCESS

Edited by:
Corrado Aaron Visaggio,
University of Sannio, Italy

Reviewed by:
Cuneyt Gurcan Akcora,
University of Manitoba, Canada
Julian Jang-jaccard,
Massey University, New Zealand

****Correspondence:***
Michael Platzer
michael.platzer@mostly.ai

Specialty section:
This article was submitted to
Cybersecurity and Privacy,
a section of the journal
Frontiers in Big Data

Received: *12 March 2021*
Accepted: *04 June 2021*
Published: *29 June 2021*

Citation:
Platzer M and Reutterer T (2021)
Holdout-Based Empirical Assessment
of Mixed-Type Synthetic Data.
Front. Big Data 4:679939.
doi: 10.3389/fdata.2021.679939

1 INTRODUCTION

Self-supervised generative AI has made significant progress over the past years, with algorithms capable of creating "shockingly" realistic synthetic data across a wide range of domains. Illustrations like those presented in **Figures 1**, **2** are particularly impressive within domains of unstructured data, like images (Karras et al., 2017) and text (Brown et al., 2020). These samples demonstrate that it is becoming increasingly difficult for us humans, as well as for machines, to discriminate actual from machine-generated fake data. While less prominent, similar progress is made within structured data domains, such as synthesizing medical health records (Choi et al., 2017; Goncalves et al., 2020; Krauland et al., 2020), census data (Freiman et al., 2017), human genoms (Yelmen et al., 2021), website traffic (Lin et al., 2020) or financial transactions (Assefa, 2020). These advances are particularly remarkable considering that they do not build upon our own human understanding of the world, but "merely" require a flexible, scalable self-supervised learning algorithm that teaches itself to create novel records based on a sufficient amount of training data. These AI-based approaches, with Generative Adversarial Networks (Goodfellow et al., 2014) and Variational

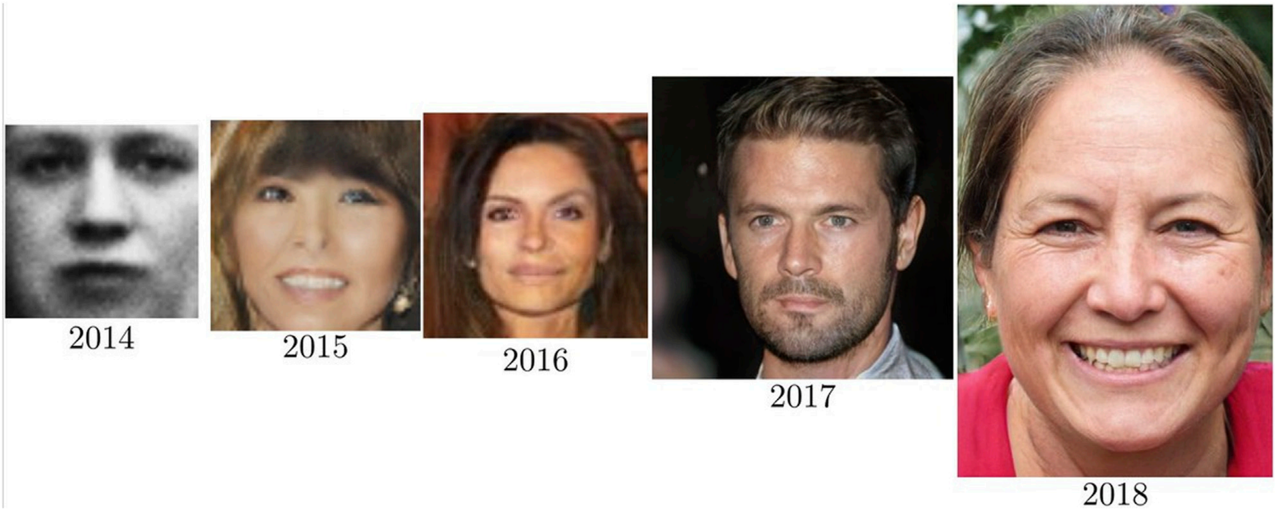
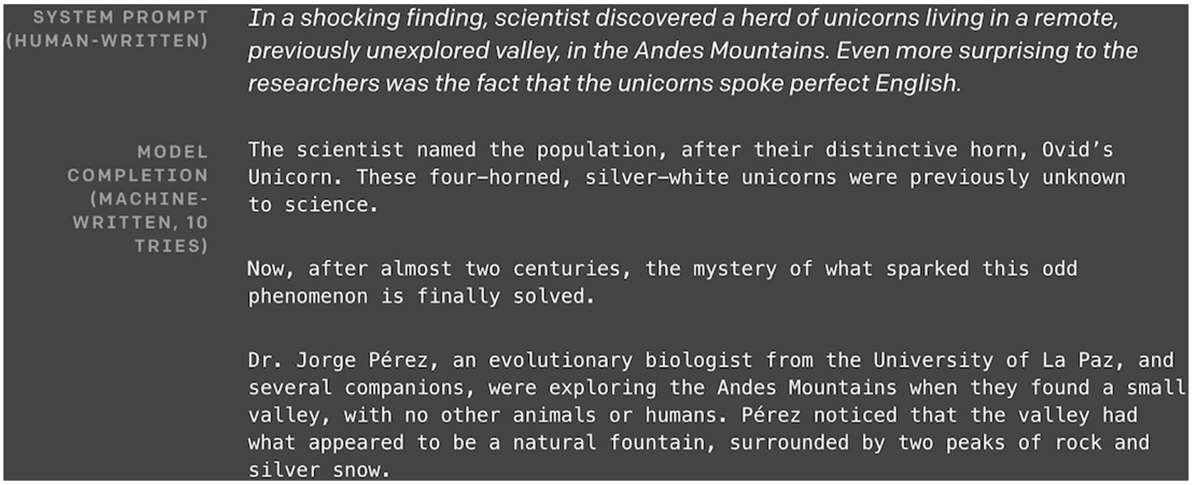


FIGURE 1 | Progress in synthetic face generation due to advances in self-supervised generative AI methods (Source: Tweet by Ian Goodfellow).



SYSTEM PROMPT (HUMAN-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)

```
The scientist named the population, after their distinctive horn, Ovid's
Unicorn. These four-horned, silver-white unicorns were previously unknown
to science.

Now, after almost two centuries, the mystery of what sparked this odd
phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and
several companions, were exploring the Andes Mountains when they found a small
valley, with no other animals or humans. Pérez noticed that the valley had
what appeared to be a natural fountain, surrounded by two peaks of rock and
silver snow.
```

FIGURE 2 | Sample text generated by Generative Pre-trained Transformer 2 (GPT-2), a large-scale open-source generative language model created by OpenAI (Radford et al., 2019).

Autoencoders (Kingma and Welling, 2013) being two prominent representatives, have in common that they fit high-capacity deep neural networks to training data, that can then be leveraged for sampling an unlimited amount of new records. This is in contrast to traditional synthetization techniques, that either rely on expert-engineered generation mechanisms or on the perturbation of existing data (Muralidhar et al., 1999; Reiter, 2010; Wieringa et al., 2021).

Given this growing capability to generate arbitrary amounts of new data, many applications arise and provide rich opportunities. These range from automated content creation (Shu et al., 2020), test data generation (Popić et al., 2019), world simulations for accelerated learning (Ha and Schmidhuber, 2018), to general-purpose privacy-safe data sharing (Howe et al., 2017; Surendra and Mohan, 2017; Bellovin et al., 2019; Hittmeir et al., 2019; Li et al., 2019).

We focus on the data sharing use cases, where data owners seek to provide highly accurate, yet truly anonymous statistical representations of datasets. AI-based approaches for generating synthetic data provide a promising novel tool box for data stewards in the field of statistical disclosure control (SDC) (Drechsler, 2011), but just as more traditional methodologies also share the fundamental need to balance data utility against disclosure risk. One can maximize utility by releasing the full original dataset, but would thereby expose the privacy of all contained data subjects. On the other hand, one can easily minimize the risk by releasing no data at all, which naturally yields zero utility. It is this privacy-utility trade-off that we seek to quantify for mixed-type synthetic data. To this end, in this paper we introduce and empirically demonstrate a novel, flexible and easy-to-use framework for measuring the fidelity as well as the privacy risk entailed in synthetic data in mixed-type tabular data setting. After briefly discussing the background we present the building blocks of the proposed framework in section *Framework*. This will then allow us to compare the performance of generative models from the rapidly growing field of synthetic data approaches against each other, as well as against alternative SDC techniques in section *Empirical Demonstration*.

2 RELATED WORK

The field of generative AI gained strong momentum ever since the introduction of Generative Adversarial Networks (Goodfellow et al., 2014) and its application to image synthesis. This seminal and widely cited paper assessed synthetic data quality by fitting Gaussian Parzen windows to the generated samples in order to estimate the log-likelihood of holdout samples. At that time the authors already called out for further research to assess synthetic data, as they highlighted the limitations of Parzen window estimates for higher dimensional domains. Theis et al. (2015) further confirmed the fundamental shortcomings of likelihood-based measures as quality metrics, as they were easily able to construct counter examples where these two do not align.

In addition to quantitative assessments, nearly all of the research advances for image synthesis also present non-cherry picked synthetic samples as an indicator for quality (see e.g., Radford et al., 2015; Liu and Tuzel 2016; Karras et al., 2017). While these allow to visually judge plausibility of the generated data, they do not allow to capture a generator's ability to faithfully represent the full variety and richness of a dataset, i.e., its dataset-level statistics. On the contrary, by overly focusing on "realistic" sample records in the assessment, one will potentially favor generators that bias toward conservative, safe-bet samples, at the cost of diversity and representativeness. Note that methods like temperature-based sampling (Ackley et al., 1985), top-k sampling (Fan et al., 2018), and nucleus sampling (Holtzman et al., 2019) are all techniques to make such trade-offs explicitly, and are commonly applied for synthetic text generation.

For mixed-type tabular data a popular and intuitive approach is to visually compare histograms and correlation plots (see e.g., Howe et al., 2017; Beaulieu-Jones et al., 2019; Lu et al., 2019). While this does allow to capture representativeness, it is typically applied to only a small subset of statistics and misses out on systematically quantifying any discrepancies thereof.

A popular assessment technique within structured as well as unstructured domains is known as "Train on Synthetic, Test on Real" (TSTR) method (Esteban et al., 2017). Using this technique, a supervised machine learning task is trained on the generated synthetic data to then see how its predictive accuracy compares against the same model being trained on real data (Jordon et al., 2018; Xu et al., 2019). By validating against an actual holdout dataset, that is not used for the data synthesis itself, one gets an indication for the information loss for a specific relationship within the data incurred due to the synthesis. If the chosen predictive task is difficult enough and a capable downstream machine learning model is used, this can indeed yield a strong measure. However, results will depend on both of these assumptions, and will vary even for the same dataset from predicted target to predicted target, as it tests only for a singular relationship within the high dimensional data distribution. And more importantly, the measure again does not allow statements regarding the overall statistical representativeness. Any accidentally introduced bias, any artifacts, any misrepresentations within the generated data might remain unnoticed. Yet, all of these are of particular importance when a data owner seeks to disseminate granular-level information with highest possible accuracy, without needing to restrict or even to know the downstream application.

No accuracy assessment of a synthetic data solution can be complete, if it does not include some measurement of its ability to produce truly novel samples, rather than merely memorizing and recreating actual data. Closely related, users of synthetic data solutions seek to establish the privacy of a generated dataset. i.e., whether the synthetic dataset is considered to be anonymous, non-personal data in a legal sense. With data protection regulations varying from country to country, and industry to industry, any ultimate assessment requires legal expertise and can only be done with respect to a given regulation. However, there are a growing number of technical definitions and assessments of privacy being introduced, that serve practitioners well to make the legal case. Two commonly used concepts within the context of synthetic data are empirical attribute disclosure assessments (Taub et al., 2018; Hittmeir et al., 2020), and Differential Privacy (Dwork et al., 2006). Both of these have proven to be useful in establishing trust in the safety of synthetic data, yet come with their own challenges in practice. While the former requires computationally intensive, case-specific repeated synthetization re-runs that can become infeasible to perform on a continuous base, the latter requires the inspection of the algorithms as well as their actual implementations for these to be validated.

3 FRAMEWORK

We seek to close these existing gaps for evaluating data synthesizers by offering 1) a flexible, model-free and easy to reason empirical assessment framework of data fidelity, and 2) an easy to compute summary statistic for the empirical assessment of privacy for mixed-type tabular data. Grace to their purely data-driven, non-parametric nature both measure neither require any a priori domain specific knowledge nor assumptions of the investigated synthetization process. This framework allows for a systematic assessment of black-box synthetic data solutions that can be performed on a continuous base and thus shall help to establish transparency and ultimately trust in this new technology.

3.1 Fidelity

We motivate our introduced fidelity measure by visualizing selected distributions and cross-tabulations for the "adult"

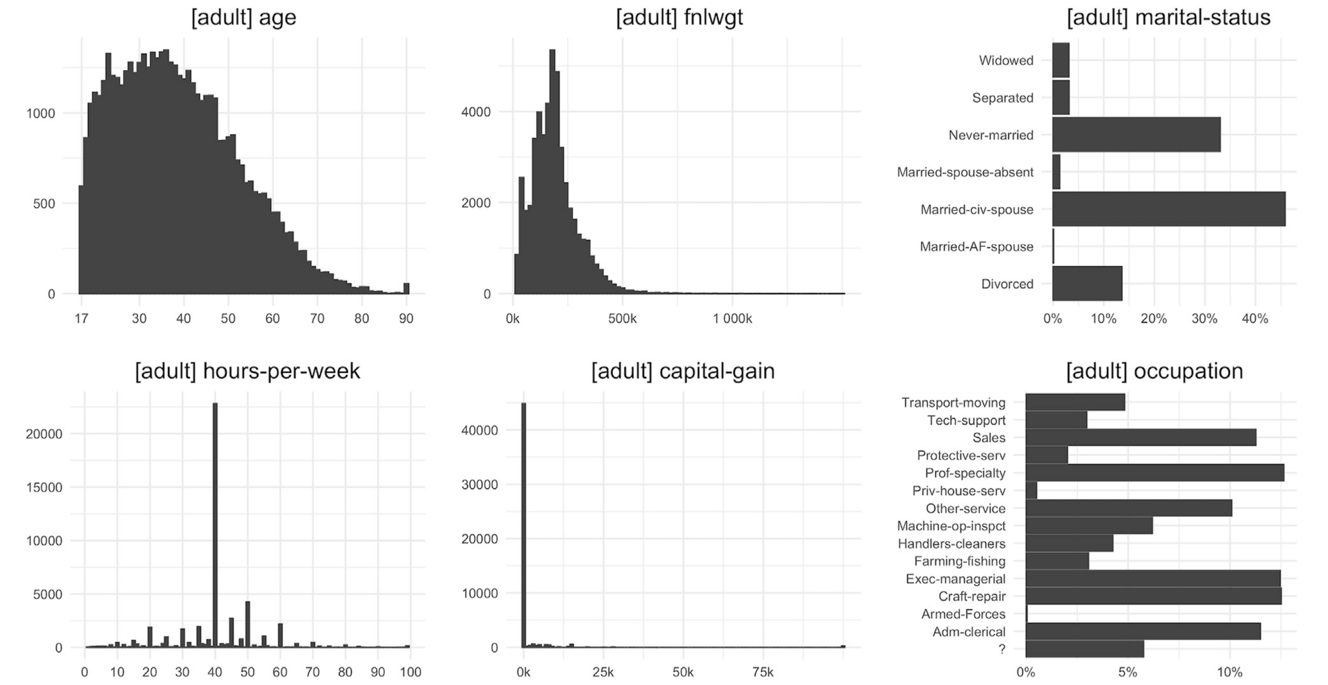


FIGURE 3 | Selected univariate marginal distributions for dataset "adult" demonstrating the broad spectrum of distributional patterns encountered in real-world datasets.

dataset, which we will use later in our empirical demonstration study. **Figure 3** exhibits the distribution of four selected numeric attributes and shows the wide variety of shapes that can occur in real-world datasets. For example, the numeric attribute "age" ranges from 17 to 90, with a small group of subjects that are exactly 90 years old, while hardly any subject is between 85 and 89 years old. The numeric attribute "fnlwgt" spans a much wider range, with nearly all observed values being unique within the dataset. Thus these values need to be binned to adequately visualize the shape of the variable's distribution. Attribute "hours-per-week" is characterized by specific outstanding integer values, while "capital-gain" is dominated by zeros with only a few exceptions that themselves can range up to 100,000. Since we would want to see synthesizers faithfully retaining any of these different types and shapes of univariate distributional patterns we also require an accurate fidelity measure to capture any such discrepancies as well.

However, we expect from synthetic data that they are not only representative for the distribution of individual attributes, but for all multivariate combinations and relationships among the set of attributes. For example, **Figure 4** displays three selected bivariate distributions for the dataset "adult," each with distinct patterns and insights that are to be retained and assessed. The challenge for deriving a metric that accommodates these empirical interdependencies in an adequate way is that the number of relationships to investigate grows quickly with the number of attributes. More specifically, a dataset with m attributes results in $\binom{m}{k}$ combinations of k-way interactions. For example, for 50 attributes this yields 1,225 two-way, and 19,600 three-way interactions. Ideally, we would want to compare the full joint empirical distributions ($m = k$) between the actual and synthetic data, but that is, except for the most trivial cases, infeasible in practice.

The curse of dimensionality (Bellman, 1966) strikes here again; i.e.,the number ofcross-combinationsof attribute values grows exponentially as more attributes are considered, resulting in the available data becoming too sparse in a high-dimensional data space. While binning and grouping of attribute values mitigates the issue for the lower-level interactions, this fundamental principle cannot be defeated for deeper levels. Thus we propose as a non-parametric, model- and assumption-free approach to empirically measure the fidelity of a synthetic dataset with respect to a target dataset by averaging across the total variation distances.[1] Further, (TVD) of the corresponding discretized, lower-level empirical marginal distributions.

The construction of our proposed fidelity metric is as follows: Let's consider a random split of the available records into a training dataset T and a holdout dataset H. We only expose the training data T to a synthesizer which yields a synthetic dataset S of arbitrary size n_S. Further, let's transform each of the m attributes of these datasets into categorical variables, that have a fixed upper bound c for their cardinality. For those categorical

[1] We explored other distance measures for empirical distributions, like the maximum absolute error, Euclidean distances, the Jensen-Shannon distance or the Hellinger distance, but they yielded practically identical rankings for the empirical benchmarks. However, the TVD is easy to communicate, easy to reason about and has exhibited in our experiments a low sensitivity with respect to sampling noise.

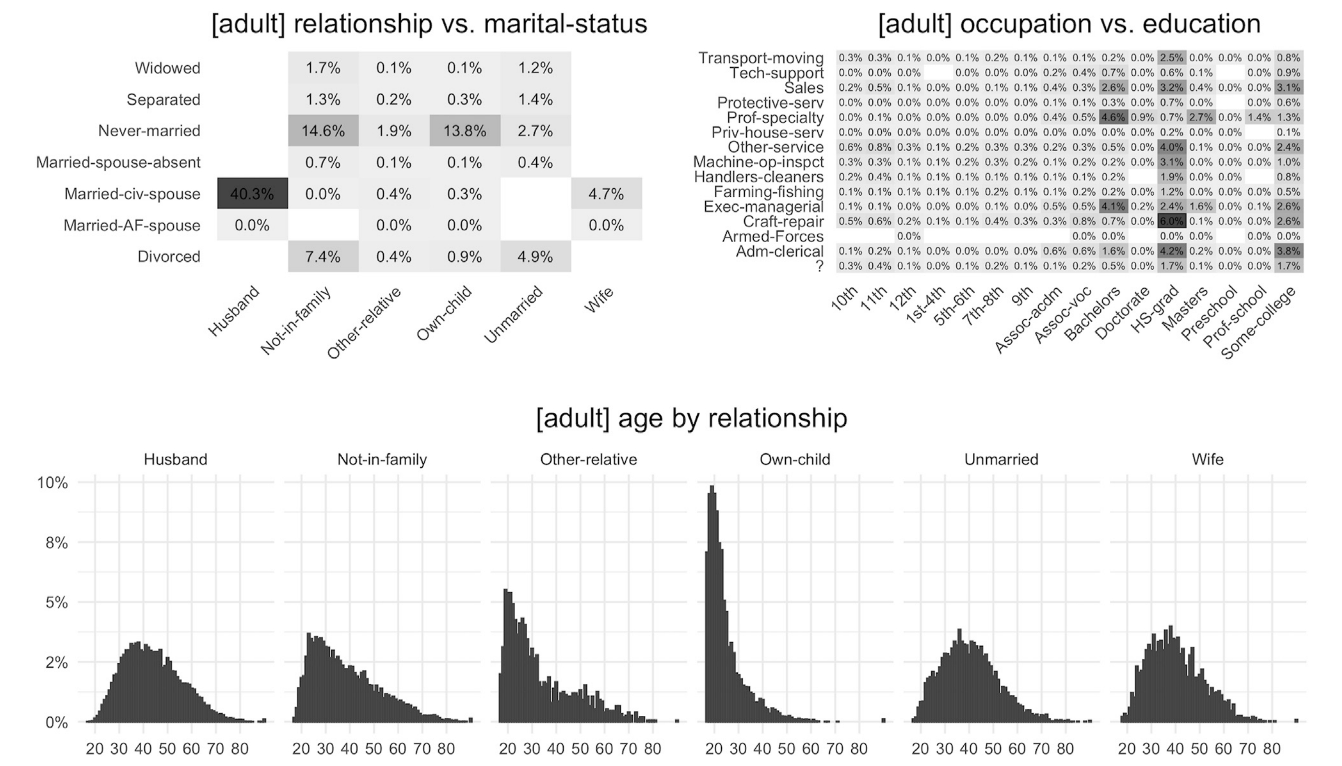


FIGURE 4 | Selected bivariate marginal distributions for dataset "adult"; an accurate measure of data fidelity is expected to be flexible enough to capture discrepancies of such distributional patterns.

variables that have cardinality $c_j > c$, we merge the $(c_j - c + 1)$ least frequent values into a single group. For numeric variables we apply quantile binning, i.e., we cut the range of values into a maximum of c ranges, based on their c quantiles. Any date or datetime variable is to be converted first into a numeric representation before applying the same transformation as suggested above. Any missing values are treated as yet another categorical value, thus can increase cardinality to $c + 1$ for those variables that contain missing values. Note, that the required statistics for the discretization, i.e., the list of least frequent values as well as the quantiles, are to be determined based on the training dataset T alone, and then reused for the discretization of the other datasets.

We then proceed in calculating relative frequencies for all k-way interactions for the discretized m attributes and do so for both the training dataset T and the synthetic dataset S. For each k-way interaction we calculate the TVD between the two corresponding empirical marginal distributions and then average across all $\frac{m}{k}$ combinations. This yields a measure F^k (T,S), which quantifies the fidelity of synthetic dataset S with respect to original training dataset T. Formally, the TVD between a specific k-way combination v for datasets T and S is half the $L1$ distance between the empirical distributions:

$$TVD_v(T, S) = \frac{1}{2} \sum_i \left| f_v^T(X = i) - f_v^S(X = i) \right| \tag{1}$$

with f_v^A denoting the empirical marginal distribution for a dataset A, v being any of the $\binom{m}{k}$ k-element combinations of the set of m attributes and i being any of the occurring attribute values of v. The introduced fidelity metric of dataset S with respect to dataset T is then the average across the TVDs for all possible k-way combinations and can be written as follows:

$$F^k(T, S) := 1/\binom{m}{k} \cdot \sum_v TVD_v(T, S) \tag{2}$$

In order to get a sense of how much information is lost due to the synthetization and how much discrepancies are expected due to sampling noise we need to compare $F^k(T, S)$ with the fidelity measure of the holdout dataset, $F^k(T, H)$. This serves us as a reference for what we aim for when retaining statistics that generalize beyond the individuals. This relationship can be easily quantified as the ratio $F^k_{ratio}(T, H, S) := F^k(T, S)/F^k(T, H)$. Note that a ratio of 1 would be optimal as it indicates that the synthetic dataset S is just as close to the training dataset T as a holdout dataset H is with respect to T due to the sampling noise. On the other hand, a ratio smaller than 1 would indicate that the synthetic dataset is systematically "too close" and contains information that represents training data specific information.

3.2 Privacy

While fidelity is assessed at the dataset-level, we need to look at individual-level distances for making the case that none of the training subjects is exposed by any of the generated synthetic records.

A simplistic approach is to check for identical matches, i.e., records from the training set that are also contained in the synthetic set. However, the occurrence of identical matches is neither a required nor a sufficient condition for detecting a leakage of privacy. Just as any dataset can contain duplicate records, we shall expect a similar relative occurrence within a representative synthetic dataset. Further, and analogous to that metaphorical monkey typing the complete works of William Shakespeare by hitting random keys on a typewriter for an infinite time (also known as the "infinite monkey theorem"), even an uninformed random data generator will eventually end up generating any "real" data record. More importantly, these identical matches must not be removed from the synthetic output, as such a rejection filter actually leaks privacy, since it would reveal the presence of a specific record in the training data by it being absent from a sufficiently large generated synthetic dataset.

The concept of identical matches is commonly generalized toward measuring the distance to closest records (DCR) (Park et al., 2018; Lu et al., 2019). These are the individual-level distances of synthetic records with respect to their corresponding nearest neighboring records from the training dataset. The distance measure itself is interchangeable, whereas in line with the discrete perspective we took in our fidelity assessment we opt for the Hamming distance applied to the discretized dataset as an easy-to-compute distance metric that fulfills the conditions of non-negativity and symmetry. However, we note that the very same framework can be just as well applied on top of alternative distance metrics, including ones based on more meaningful learned representations of domain-specific embedding spaces. A DCR of 0 corresponds to an identical match. But as argued above, also that metric in itself does not reveal anything regarding the leakage of individual-level information, but is rather a statistic of the data distribution we seek to retain. Therefore, to provide meaning and to facilitate interpretation the measured DCRs need to be put into the context of their expected value, which can be estimated based on an actual holdout dataset.

As illustrated in **Figure 5**, we therefore propose to calculate for each synthetic record its DCR with respect to the training data T as well as with respect to an equally sized holdout dataset H. The *share of records* that are then closer to a training than to a holdout record serves us as our proposed privacy risk measure. Any ties are to be distributed equally between these two datasets. If that resulting share is then close to 50%, we gain empirical evidence of the training and holdout data being interchangeable with respect to the synthetic data.[2] This in turn allows to make a strong case for plausible deniability for any individual, as the synthetic data records do not allow to conjecture whether an individual was or was not contained in the training dataset. Even for cases of a strong resemblance of a particular record with a real-world subject, it can be argued that such a resemblance can occur for unseen subjects just as well. Translated into the world of motion pictures this idea would correspond to the proverbial disclaimer that "any resemblance to persons living or dead is purely coincidental."

4 EMPIRICAL DEMONSTRATION

To demonstrate the usefulness of the presented framework for assessing fidelity and privacy of synthetic data solutions, we apply it to four publicly available, mixed-type tabular datasets from the UCI Machine Learning repository (Dua and Graff, 2017) and synthesize them using seven publicly available data synthesizers.

The datasets cover a broad range of scenarios and are commonly used in the data synthesis literature (Park et al., 2018; Xu et al., 2019; Zhao et al., 2021) as well as by commercial and open-source software providers[3] to demonstrate the effectiveness of the proposed methods. Each dataset is representative of a common business scenario, where privacy-sensitive data assets are to be shared for analytical tasks. Every record within these datasets corresponds to a single person, whose privacy shall be protected, while the statistical information of the overall dataset shall be retained.

The datasets included for the purpose of demonstration are:

- adult: 48,842 records with 15 attributes (6 numerical, 9 categorical)
- bank-marketing: 45,211 records with 17 attributes (7 numerical, 10 categorical)
- credit-default: 30,000 records with 24 attributes (20 numerical, 4 categorical)
- online-shoppers: 12,330 records with 18 attributes (4 numerical, 14 categorical)

The seven tested generative models include four generators contained as part of MIT's Synthetic Data Vault (SDV) library [i.e., synthesizers CopulaGAN, CTGAN, Gaussian Copula, and TVAE; see Montanez. (2018)], the synthpop R package (Nowok et al., 2016), an open-sourced generator by Gretel[4], and one closed-source solution by MOSTLY AI[5], which is also freely available online community edition.

Each of the four datasets is randomly split into an equally sized training and holdout dataset. The seven generative models are fitted to the training data to then generate 50,000 synthetic records for each dataset. All synthesizers are run with their default settings unchanged, i.e., no parameter tuning is being performed.

[2] Note, that as the holdout records are randomly sampled and never exposed to the synthesizer, the synthesizer can not systematically generate subjects that are closer to these than to the training records. Thus, the presented privacy metric can not be undermined by mixing "too close" with "too far away" records in an attempt to achieve a balanced share.

[3] see https://www.synthesized.io/data-template-pages/bank-marketing, https://mostly.ai/2020/08/07/boost-machine-learning-accuracy-with-synthetic-data/ and https://gretel.ai/blog/machine-learning-accuracy-using-synthetic-data

[4] https://gretel.ai

[5] https://mostly.ai

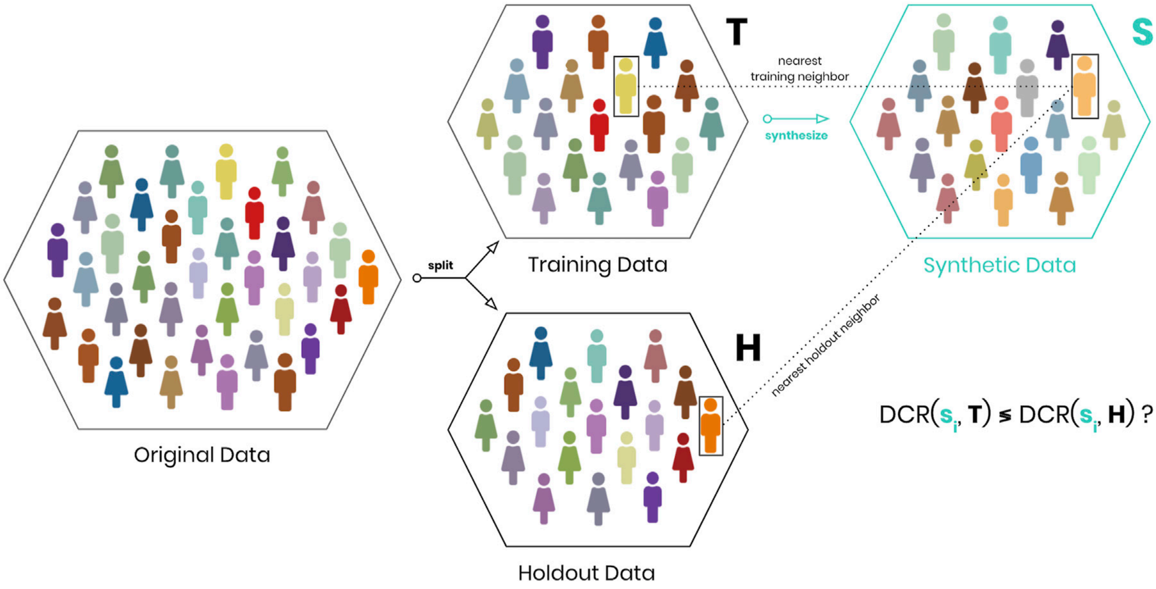


FIGURE 5 | Construction of the holdout-based privacy risk measure. For each synthetic record we determine whether the nearest neighbor within training data T is or is not smaller than the nearest neighbor with respect to holdout data H. The share of records closer to a training than to a holdout data serves for evaluating privacy risk.

To provide further context to the assessment of the synthesized datasets, we also generate additional datasets by simply perturbating the training data with a varying degree of noise. We do so by drawing 50,000 records with replacement from the dataset and then decide for each value of each record with a given probability (ranging from 10% up to 90%) whether to keep it or to replace it with a value from a different record. This approach adds noise to existing records, yet retains the univariate marginal distributions. With more noise being added, one expects privacy to be increasingly protected, while also more statistical relations to be distorted. This allows us to compare the newly emerging class of data synthesizers with a simpler, yet more established method in the field of statistical disclosure control (see also Muralidhar et al. (1999) or Muralidhar and Sarathy (2006) for similar approaches).

Figures 6–8 visualize the resulting distributions of selected univariate, bivariate and three-way attribute interactions for the "adult" dataset across the synthetic datasets generated by the various synthesizers included in our study as well as the two purturbated datasets (i.e., Flip 10% and Flip 90%). While the visual inspection already allows to spot some qualitative differences with respect to the goodness of representativeness of the training data, it is the corresponding fidelity metric F (reported as percentages in brackets) that provides us with a quantitative summary statistic. The reported fidelity measure for the holdout data serves as a reference, as the derived distributions should not be systematically closer to the training data than what is expected from the holdout data. For example, an $F^1(T, S)$ fidelity score coming close to the 2.7% reported for the variable "age" in the holdout (which is due to the sampling noise) can be considered as an accurate representation of the underlying distribution in the training data. However, visually inspecting 15 univariate, 105 bivariate. and 455 three-way interactions for dataset "adult" is prohibitive, but the proposed summary statistics which average across these yield a condensed but informative fidelity assessment of synthetic data.

Figure 9 reports the proposed fidelity measures across all four datasets, the used generative synthetization methods and various degrees of perturbation.[6] It is interesting to note that the rankings with respect to fidelity among synthesizers are relatively consistent across all datasets, showing that these metrics indeed serve as general-purpose measures for the quality of a synthesizer. The reported numbers for the perturbated datasets exhibit the expected relationship between noise level and fidelity. Among the benchmarked synthesizers "synthpop" and "MOSTLY" exhibit the highest fidelity score with the caveat that the former is systematically too close to the training data compared to what is expected based on the holdout.

Figure 10, on the other hand, contains the results for the proposed privacy risk measure. For each dataset and synthesizer the share of synthetic records that is closer to a training record than to a holdout record is being reported. In addition, the average DCRs are displayed, once with respect to the training and once with respect to the holdout data. With the notable exception of "synthpop" all of the presented synthesizers exhibit almost identical DCR distributions for the training as well as for the holdout records. This indicates that no individual-level

[6] For the fidelity assessment we chose $c = 100$ for discretizing the univariate distributions, $c = 10$ for the bivariate combinations, and $c = 5$ for the three-way interactions. Experiments have shown that the obtained rankings among synthesizers remain relatively robust with respect to the cardinality of the categorical variable c.

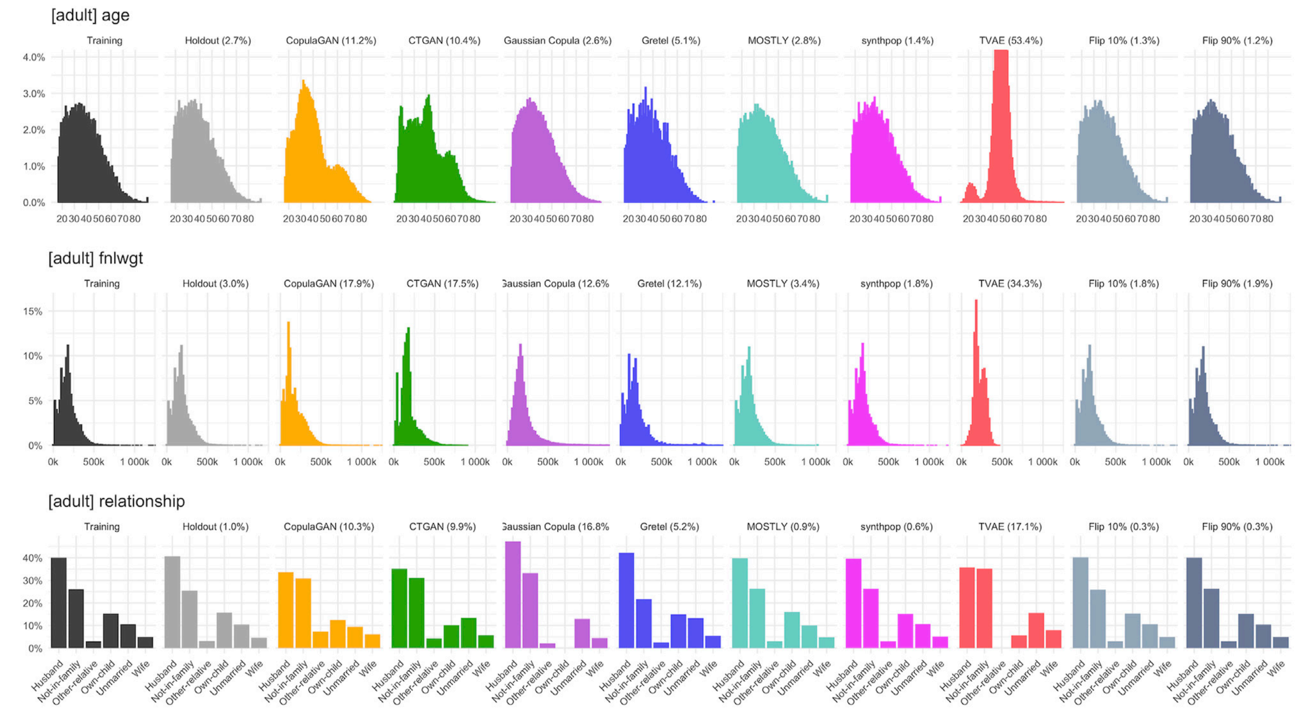


FIGURE 6 | Selected univariate marginal distributions for dataset "adult" across synthesized data and across perturbated data. Their total variation distance (TVD) with respect to the training set, displayed as percentages in brackets, contribute to the respective F^1 fidelity.

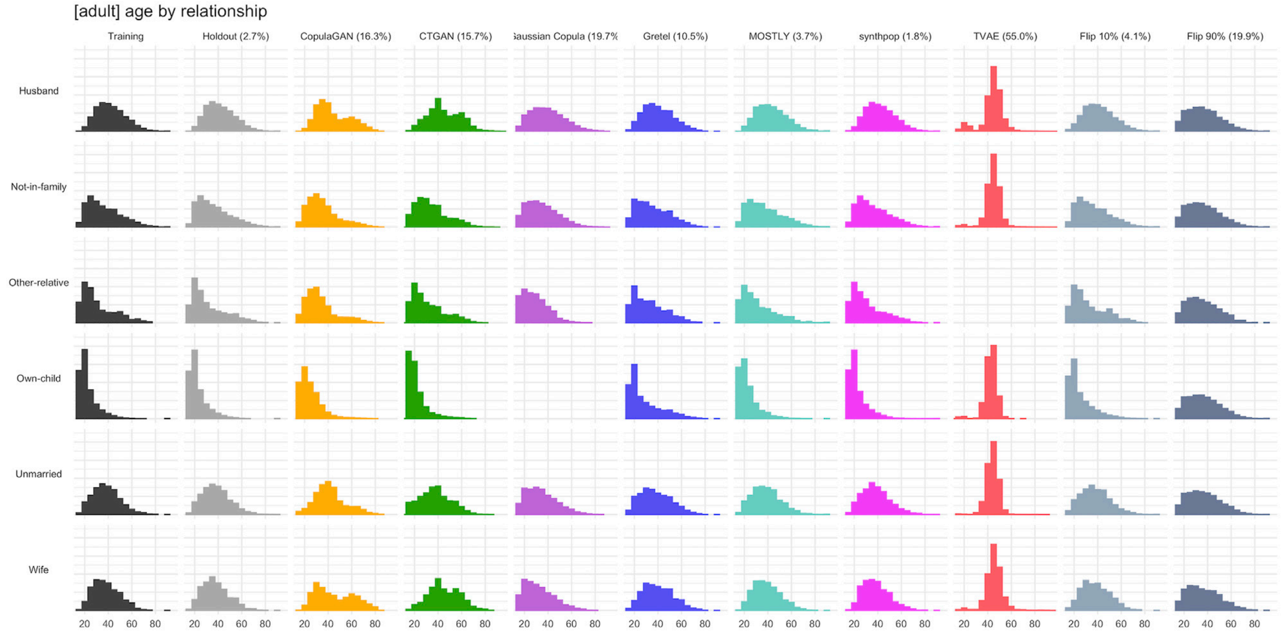


FIGURE 7 | Selected bivariate marginal distribution for dataset "adult" across synthesized data and across perturbated data. Their total variation distance (TVD) with respect to the training set, displayed as percentages in brackets, contribute to the respective F^2 fidelity measure.

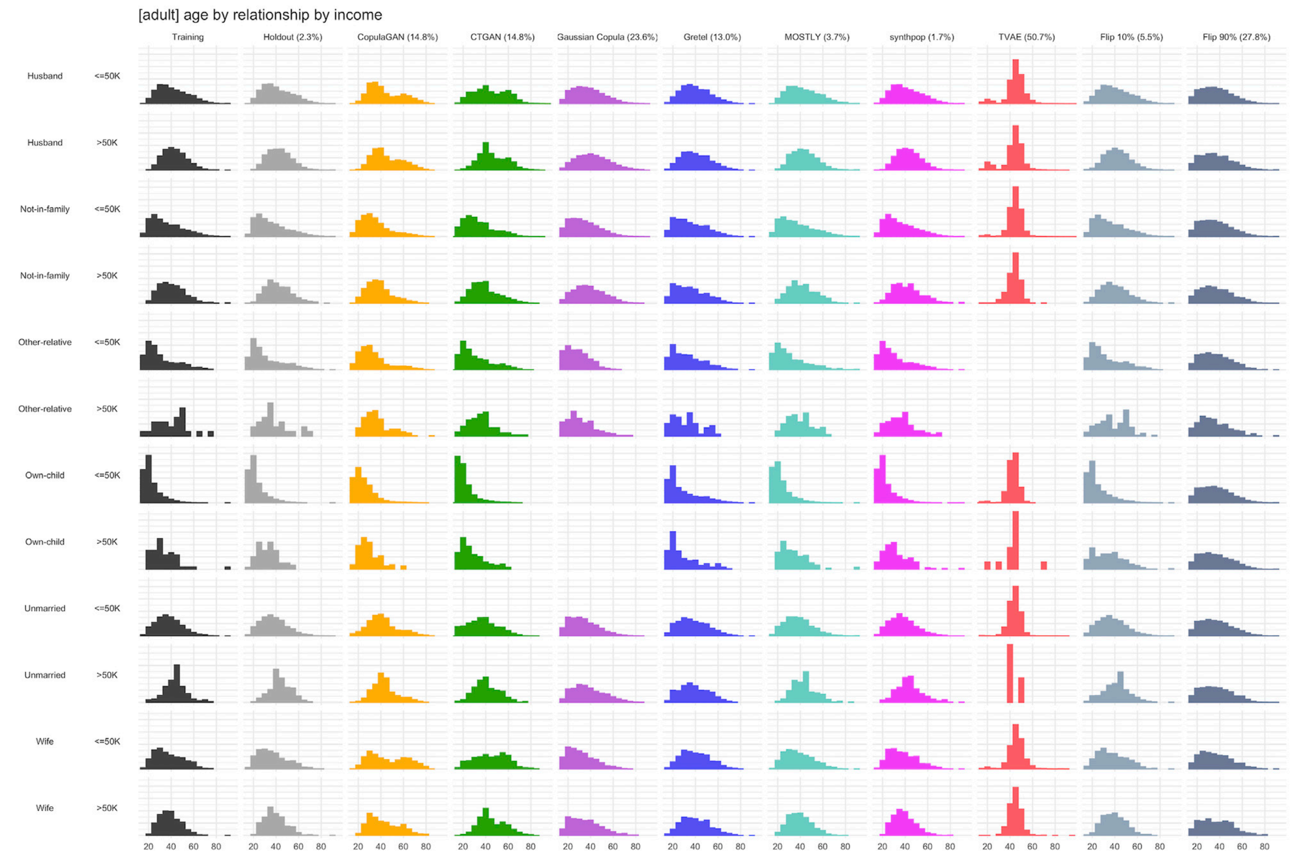
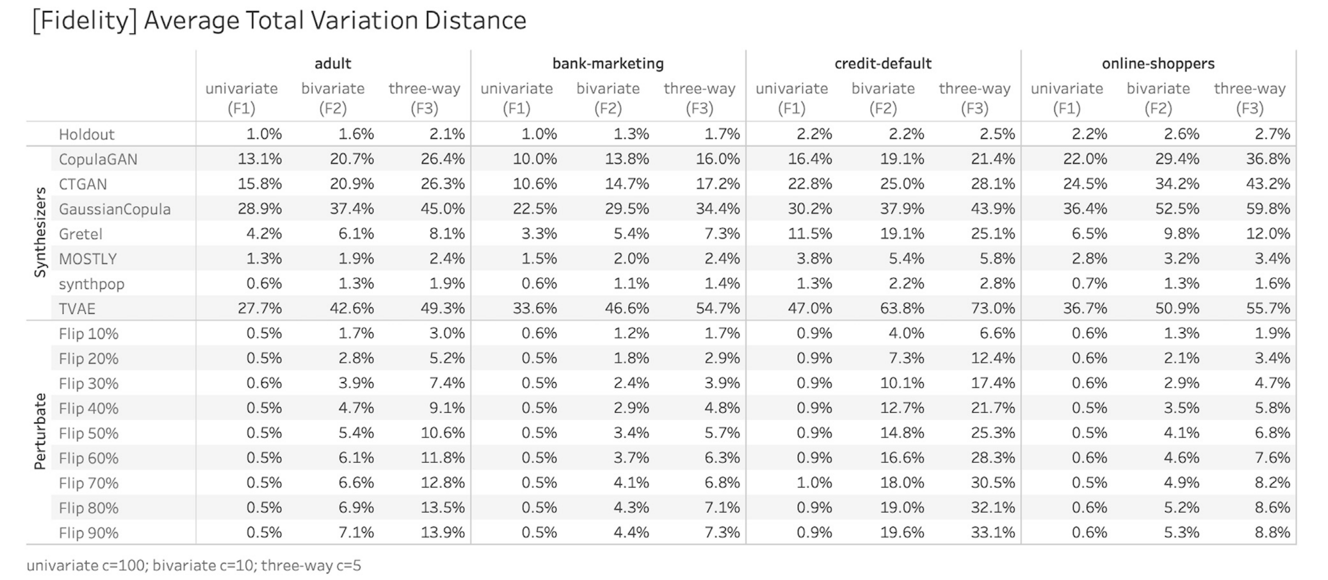


FIGURE 8 | Selected three-way marginal distribution for dataset "adult" across synthesized data and across perturbated data. Their total variation distance (TVD) with respect to the training set, displayed as percentages in brackets, contribute to the respective F^3 fidelity measure indicated as percentages in brackets.

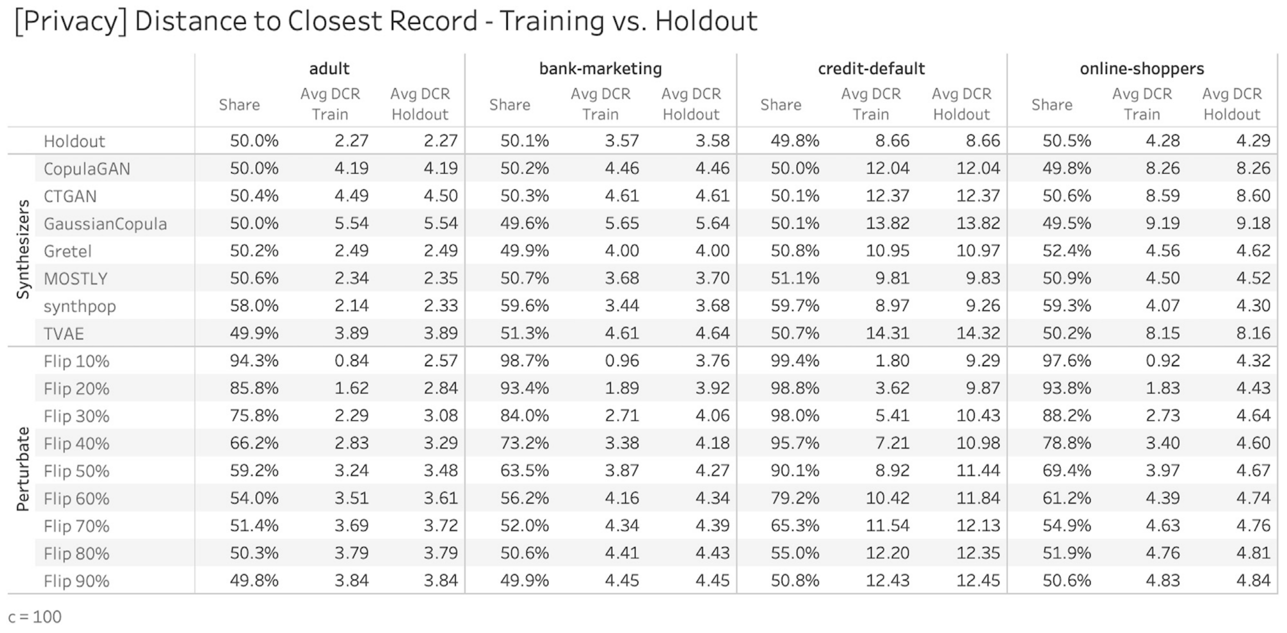


[Fidelity] Average Total Variation Distance

		adult			bank-marketing			credit-default			online-shoppers		
		univariate (F1)	bivariate (F2)	three-way (F3)	univariate (F1)	bivariate (F2)	three-way (F3)	univariate (F1)	bivariate (F2)	three-way (F3)	univariate (F1)	bivariate (F2)	three-way (F3)
	Holdout	1.0%	1.6%	2.1%	1.0%	1.3%	1.7%	2.2%	2.2%	2.5%	2.2%	2.6%	2.7%
Synthesizers	CopulaGAN	13.1%	20.7%	26.4%	10.0%	13.8%	16.0%	16.4%	19.1%	21.4%	22.0%	29.4%	36.8%
	CTGAN	15.8%	20.9%	26.3%	10.6%	14.7%	17.2%	22.8%	25.0%	28.1%	24.5%	34.2%	43.2%
	GaussianCopula	28.9%	37.4%	45.0%	22.5%	29.5%	34.4%	30.2%	37.9%	43.9%	36.4%	52.5%	59.8%
	Gretel	4.2%	6.1%	8.1%	3.3%	5.4%	7.3%	11.5%	19.1%	25.1%	6.5%	9.8%	12.0%
	MOSTLY	1.3%	1.9%	2.4%	1.5%	2.0%	2.4%	3.8%	5.4%	5.8%	2.8%	3.2%	3.4%
	synthpop	0.6%	1.3%	1.9%	0.6%	1.1%	1.4%	1.3%	2.2%	2.8%	0.7%	1.3%	1.6%
	TVAE	27.7%	42.6%	49.3%	33.6%	46.6%	54.7%	47.0%	63.8%	73.0%	36.7%	50.9%	55.7%
Perturbate	Flip 10%	0.5%	1.7%	3.0%	0.6%	1.2%	1.7%	0.9%	4.0%	6.6%	0.6%	1.3%	1.9%
	Flip 20%	0.5%	2.8%	5.2%	0.5%	1.8%	2.9%	0.9%	7.3%	12.4%	0.6%	2.1%	3.4%
	Flip 30%	0.6%	3.9%	7.4%	0.5%	2.4%	3.9%	0.9%	10.1%	17.4%	0.6%	2.9%	4.7%
	Flip 40%	0.5%	4.7%	9.1%	0.5%	2.9%	4.8%	0.9%	12.7%	21.7%	0.5%	3.5%	5.8%
	Flip 50%	0.5%	5.4%	10.6%	0.5%	3.4%	5.7%	0.9%	14.8%	25.3%	0.5%	4.1%	6.8%
	Flip 60%	0.5%	6.1%	11.8%	0.5%	3.7%	6.3%	0.9%	16.6%	28.3%	0.6%	4.6%	7.6%
	Flip 70%	0.5%	6.6%	12.8%	0.5%	4.1%	6.8%	1.0%	18.0%	30.5%	0.5%	4.9%	8.2%
	Flip 80%	0.5%	6.9%	13.5%	0.5%	4.3%	7.1%	0.9%	19.0%	32.1%	0.6%	5.2%	8.6%
	Flip 90%	0.5%	7.1%	13.9%	0.5%	4.4%	7.3%	0.9%	19.6%	33.1%	0.6%	5.3%	8.8%

univariate c=100; bivariate c=10; three-way c=5

FIGURE 9 | Fidelity measures F^1, F^2, and F^3 of the presented empirical study, across all four datasets, seven synthesizers, and in comparison to basic data perturbation techniques. The closer the fidelity scores are to the respective scores of the holdout, the better the synthetic data represent the distributions in the original training data.



[Privacy] Distance to Closest Record - Training vs. Holdout

		adult			bank-marketing			credit-default			online-shoppers		
		Share	Avg DCR Train	Avg DCR Holdout	Share	Avg DCR Train	Avg DCR Holdout	Share	Avg DCR Train	Avg DCR Holdout	Share	Avg DCR Train	Avg DCR Holdout
	Holdout	50.0%	2.27	2.27	50.1%	3.57	3.58	49.8%	8.66	8.66	50.5%	4.28	4.29
Synthesizers	CopulaGAN	50.0%	4.19	4.19	50.2%	4.46	4.46	50.0%	12.04	12.04	49.8%	8.26	8.26
	CTGAN	50.4%	4.49	4.50	50.3%	4.61	4.61	50.1%	12.37	12.37	50.6%	8.59	8.60
	GaussianCopula	50.0%	5.54	5.54	49.6%	5.65	5.64	50.1%	13.82	13.82	49.5%	9.19	9.18
	Gretel	50.2%	2.49	2.49	49.9%	4.00	4.00	50.8%	10.95	10.97	52.4%	4.56	4.62
	MOSTLY	50.6%	2.34	2.35	50.7%	3.68	3.70	51.1%	9.81	9.83	50.9%	4.50	4.52
	synthpop	58.0%	2.14	2.33	59.6%	3.44	3.68	59.7%	8.97	9.26	59.3%	4.07	4.30
	TVAE	49.9%	3.89	3.89	51.3%	4.61	4.64	50.7%	14.31	14.32	50.2%	8.15	8.16
Perturbate	Flip 10%	94.3%	0.84	2.57	98.7%	0.96	3.76	99.4%	1.80	9.29	97.6%	0.92	4.32
	Flip 20%	85.8%	1.62	2.84	93.4%	1.89	3.92	98.8%	3.62	9.87	93.8%	1.83	4.43
	Flip 30%	75.8%	2.29	3.08	84.0%	2.71	4.06	98.0%	5.41	10.43	88.2%	2.73	4.64
	Flip 40%	66.2%	2.83	3.29	73.2%	3.38	4.18	95.7%	7.21	10.98	78.8%	3.40	4.60
	Flip 50%	59.2%	3.24	3.48	63.5%	3.87	4.27	90.1%	8.92	11.44	69.4%	3.97	4.67
	Flip 60%	54.0%	3.51	3.61	56.2%	4.16	4.34	79.2%	10.42	11.84	61.2%	4.39	4.74
	Flip 70%	51.4%	3.69	3.72	52.0%	4.34	4.39	65.3%	11.54	12.13	54.9%	4.63	4.76
	Flip 80%	50.3%	3.79	3.79	50.6%	4.41	4.43	55.0%	12.20	12.35	51.9%	4.76	4.81
	Flip 90%	49.8%	3.84	3.84	49.9%	4.45	4.45	50.8%	12.43	12.45	50.6%	4.83	4.84

c = 100

FIGURE 10 | Privacy measures of the presented empirical study, across four datasets, seven synthesizers, and in comparison to basic data perturbation techniques. A share close to 50% indicates empirical evidence of privacy preservation for the synthesized data which is the case for most of the data synthesizers under study.

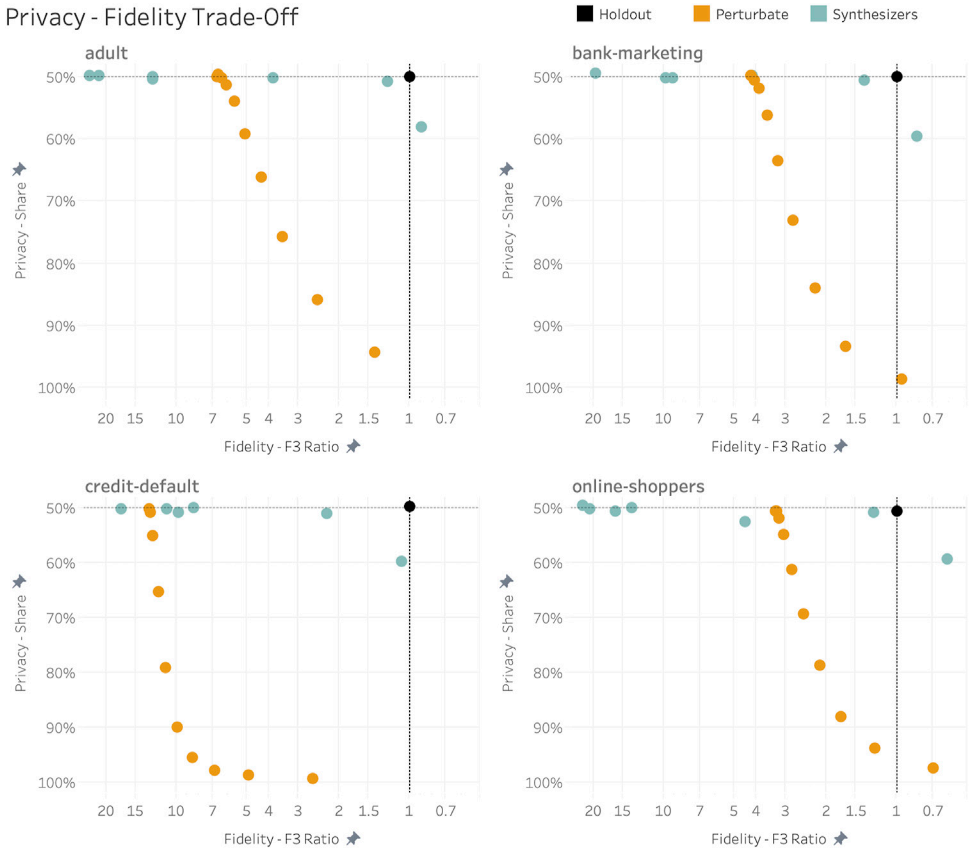


FIGURE 11 | Empirically established trade-off between privacy and fidelity across synthetization and perturbation approaches. Fidelity is displayed as the ratio F^3 (T,S)/F^3 (T,H), as the holdout dataset serves as maximum attainable reference point. In contrast to synthesized data, perturbation techniques fail to protect privacy without sacrificing fidelity.

information of the training subjects has been exposed beyond what is attainable from the underlying distribution and thus makes a strong case for the generated data preserving the privacy of the training subjects. In contrast, the reported numbers for the perturbated datasets reveal a severe exposure of the training subjects, even if a high amount of noise is being added. Only at a level where most of the utility of these datasets is being destroyed, the privacy measures start to align with the holdout dataset.

Based on these results we can further visualize the uncovered empirical relationship between privacy and fidelity. The *x*-axes in **Figure 11** represent the three-way fidelity measure in relation to its corresponding value for the holdout dataset, i.e., the proposed fidelity ratio F^3 (T,S)/F^3 (T,H). The *y*-axes represent the reported share of records that are closer to training than to the holdout. Presented this way, the holdout dataset serves us as a “north star” for truly privacy-respecting data synthesizers in the upper right corner. The orange dots represent the range of perturbated datasets and reveal the difficulties of basic obfuscation techniques to protect privacy without sacrificing fidelity, particularly for higher-dimensional datasets. The turquoise marks on the other hand represent the performance metrics for a broad range of emerging synthesizers, whereas all except one exhibit DCR shares close to 50%, and with some getting already very close to representing the characteristics of a true holdout dataset.

5 DISCUSSION AND FUTURE RESEARCH

The field of supervised machine learning benefited from having commonly used benchmark datasets and metrics in place to measure performance across methods as well as progress over time. The emerging field of privacy-preserving structured synthetic data is still to converge onto commonly agreed fidelity and privacy measures, as well as to a set of canonical datasets to benchmark on. This research aims at complementing already existing methods by introducing a practical, assumption-free and easy to reason empirical assessment framework that can be applied for any black-box synthetization method and thus shall help to objectively capture and measure the progress in the field. In addition, the reported findings from the empirical benchmark experiments demonstrate the promise of AI-based data synthesis when compared to simpler data perturbation techniques. However, they also highlight the need to not only assess fidelity but just as well the privacy risk of these newly emerging, powerful data generators.

We hope this contribution fosters further empirical benchmarks across a broad range of datasets and thus helps to establish transparency as well as facilitate comparability across the various synthetization methods. To the best of our knowledge this paper is the first model-free, non-parametric approach to quantify the fidelity and privacy risk of synthetic data by using two easy-to-compute summary statistics. We acknowledge the fact that TVD is sensitive to the cardinality and the actual shape of the underlying distributions. Furthermore, in our approach we need to discretize the data by specifying an upper bound *c* on the cardinality of all variables. While we find that our results and substantive findings are robust to the choice of *c* and the distance measures, future research could further explore the relationship between these aspects and the power of our proposed fidelity and privacy metrics.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found at https://github.com/mostly-ai/paper-fidelity-accuracy.

AUTHOR CONTRIBUTIONS

MP concepted, implemented, and empirically applied assessment framework. TR contributed, challenged, and edited during the whole process.

FUNDING

The research is supported by the “ICT of the Future” funding program of the Austrian Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology.

REFERENCES

Ackley, D. H., Hinton, G. E., and Sejnowski, T. J. (1985). A Learning Algorithm for Boltzmann Machines*. *Cogn. Sci.* 9, 147–169. doi:10.1207/s15516709cog0901_7

Assefa, S. (2020). *Generating Synthetic Data in Finance: Opportunities, Challenges and Pitfalls.* Challenges And Pitfalls. SSRN. doi:10.2139/ssrn.3634235

Beaulieu-Jones, B. K., Wu, Z. S., Williams, C., Lee, R., Bhavnani, S. P., Byrd, J. B., et al. (2019). Privacy-preserving Generative Deep Neural Networks Support Clinical Data Sharing. *Circ. Cardiovasc. Qual. Outcomes* 12, e005122. doi:10.1161/circoutcomes.118.005122

Bellman, R. (1966). Dynamic Programming. *Science* 153, 34–37. doi:10.1126/science.153.3731.34

Bellovin, S. M., Dutta, P. K., and Reitinger, N. (2019). Privacy and Synthetic Datasets. *Stan. Tech. L. Rev.* 22, 1.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). *Language Models Are Few-Shot Learners.* arXiv preprint arXiv:2005.14165.

Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., and Sun, J. (2017). “Generating Multi-Label Discrete Patient Records Using Generative Adversarial Networks,” in Machine learning for healthcare conference (PMLR), 286–305.

Drechsler, J. (2011). *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation, Vol. 201.* Springer Science & Business Media.

Dua, D., and Graff, C. (2017). *UCI Machine Learning Repository.*

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). “Calibrating Noise to Sensitivity in Private Data Analysis,” in Theory of cryptography conference (Springer), 265–284. doi:10.1007/11681878_14

Esteban, C., Hyland, S. L., and Rätsch, G. (2017). arXiv preprint arXiv: 1706.02633.Real-valued (Medical) Time Series Generation with Recurrent Conditional gans.

Fan, A., Lewis, M., and Dauphin, Y. (2018). *Hierarchical Neural story Generation.* arXiv preprint arXiv:1805.04833.

Freiman, M., Lauger, A., and Reiter, J. (2017). *Data Synthesis and Perturbation for the American Community Survey at the Us Census bureau.* US Census Bureau.

Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L., and Sales, A. P. (2020). Generation and Evaluation of Synthetic Patient Data. *BMC Med. Res. Methodol.* 20, 108–140. doi:10.1186/s12874-020-00977-1

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). *Generative Adversarial Networks*. arXiv preprint arXiv:1406.2661.

Ha, D., and Schmidhuber, J. (2018). *World Models*. arXiv preprint arXiv: 1803.10122.

Hittmeir, M., Ekelhart, A., and Mayer, R. (2019). “Utility and Privacy Assessments of Synthetic Data for Regression Tasks,” in 2019 IEEE International Conference on Big Data (Big Data) (IEEE), 5763–5772. doi:10.1109/BigData47090.2019.9005476

Hittmeir, M., Mayer, R., and Ekelhart, A. (2020). “A Baseline for Attribute Disclosure Risk in Synthetic Data,” in Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy, 133–143.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2019). *The Curious Case of Neural Text Degeneration*. arXiv preprint arXiv:1904.09751.

Howe, B., Stoyanovich, J., Ping, H., Herman, B., and Gee, M. (2017). *Synthetic Data for Social Good*. arXiv preprint arXiv:1710.08874.

Jordon, J., Yoon, J., and van der Schaar, M. (2018). *Measuring the Quality of Synthetic Data for Use in Competitions*. arXiv preprint arXiv:1806.11345.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2017). *Progressive Growing of gans for Improved Quality, Stability, and Variation*. arXiv preprint arXiv:1710.10196.

Kingma, D. P., and Welling, M. (2013). *Auto-encoding Variational Bayes*. arXiv preprint arXiv:1312.6114.

Krauland, M. G., Frankeny, R. J., Lewis, J., Brink, L., Hulsey, E. G., Roberts, M. S., et al. (2020). Development of a Synthetic Population Model for Assessing Excess Risk for Cardiovascular Disease Death. *JAMA Netw. Open* 3, e2015047. doi:10.1001/jamanetworkopen.2020.15047

Li, S.-C., Tai, B.-C., and Huang, Y. (2019). “Evaluating Variational Autoencoder as a Private Data Release Mechanism for Tabular Data,” in 2019 IEEE 24th Pacific Rim International Symposium on Dependable Computing (PRDC), Kyoto, Japan (IEEE), 198–1988. doi:10.1109/PRDC47002.2019.00050

Lin, Z., Jain, A., Wang, C., Fanti, G., and Sekar, V. (2020). “Using gans for Sharing Networked Time Series Data: Challenges, Initial Promise, and Open Questions,” in Proceedings of the ACM Internet Measurement Conference, 464–483.

Liu, M.-Y., and Tuzel, O. (2016). *Coupled Generative Adversarial Networks*. arXiv preprint arXiv:1606.07536.

Lu, P.-H., Wang, P.-C., and Yu, C.-M. (2019). “Empirical Evaluation on Synthetic Data Generation with Generative Adversarial Network,” in Proceedings of the 9th International Conference on Web Intelligence, Mining and Semantics, 1–6. doi:10.1145/3326467.3326474

Montanez, A. (2018). SDV: an Open Source Library for Synthetic Data Generation. Ph.D. thesis. Massachusetts Institute of Technology.

Muralidhar, K., Parsa, R., and Sarathy, R. (1999). A General Additive Data Perturbation Method for Database Security. *Manag. Sci.* 45, 1399–1415. doi:10.1287/mnsc.45.10.1399

Muralidhar, K., and Sarathy, R. (2006). Data Shuffling-A New Masking Approach for Numerical Data. *Manag. Sci.* 52, 658–670. doi:10.1287/mnsc.1050.0503

Nowok, B., Raab, G. M., Dibben, C., et al. (2016). Synthpop: Bespoke Creation of Synthetic Data in R. *J. Stat. Softw.* 74, 1–26. doi:10.18637/jss.v074.i11

Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., and Kim, Y. (2018). *Data Synthesis Based on Generative Adversarial Networks*. arXiv preprint arXiv: 1806.03384.

Popić, S., Pavković, B., Velikić, I., and Teslić, N. (2019). “Data Generators: a Short Survey of Techniques and Use Cases with Focus on Testing,” in 2019 IEEE 9th International Conference on Consumer Electronics (ICCE-Berlin) (IEEE), 189–194.

Radford, A., Metz, L., and Chintala, S. (2015). *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks*. arXiv preprint arXiv:1511.06434.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language Models Are Unsupervised Multitask Learners. *OpenAI blog* 1, 9.

Reiter, J. P. (2010). Multiple Imputation for Disclosure Limitation: Future Research Challenges. *J. Privacy Confidentiality* 1 (2), 223–233. doi:10.29012/jpc.v1i2.575

Shu, K., Li, Y., Ding, K., and Liu, H. (2020). *Fact-enhanced Synthetic News Generation*. arXiv preprint arXiv:2012.04778.

Surendra, H., and Mohan, H. (2017). A Review of Synthetic Data Generation Methods for Privacy Preserving Data Publishing. *Int. J. Scientific Tech. Res.* 6, 95–101.

Taub, J., Elliot, M., Pampaka, M., and Smith, D. (2018). “Differential Correct Attribution Probability for Synthetic Data: an Exploration,” in International Conference on Privacy in Statistical Databases (Springer), 122–137. doi:10.1007/978-3-319-99771-1_9

Theis, L., Oord, A. v. d., and Bethge, M. (2015). *A Note on the Evaluation of Generative Models*. arXiv preprint arXiv:1511.01844.

Wieringa, J., Kannan, P. K., Ma, X., Reutterer, T., Risselada, H., and Skiera, B. (2021). Data Analytics in a Privacy-Concerned World. *J. Business Res.* 122, 915–925. doi:10.1016/j.jbusres.2019.05.005

Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. (2019). *Modeling Tabular Data Using Conditional gan*. arXiv preprint arXiv: 1907.00503.

Yelmen, B., Decelle, A., Ongaro, L., Marnetto, D., Tallec, C., Montinaro, F., et al. (2021). Creating Artificial Human Genomes Using Generative Neural Networks. *Plos Genet.* 17, e1009303. doi:10.1371/journal.pgen.1009303

Zhao, Z., Kunar, A., Van der Scheer, H., Birke, R., and Chen, L. Y. (2021). *Ctab-gan: Effective Table Data Synthesizing*.arXiv preprint arXiv:2102.08369.

Conflict of Interest: MP is co-founder of MOSTLY AI, a provider of synthetic data solutions.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Platzer and Reutterer. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.