# Utility and Privacy Assessment of Synthetic Microbiome Data

Markus Hittmeir[1][0000−0002−3363−6270], Rudolf Mayer[1,2][0000−0003−0424−5999], and Andreas Ekelhart[1][0000−0003−3682−1364]

[1] SBA Research gGmbH, Floragasse 7, Vienna, Austria
{mhittmeir, rmayer, aekelhart}@sba-research.org
[2] Vienna University of Technology, Favoritenstraße 9-11, Vienna, Austria

**Abstract.** The microbial communities of the human body are subject to extensive research efforts. The individual variations in the human microbiome reveal information about our diet, exercise habits and general well-being, and are useful for investigations on the prediction and therapy of diseases. On the other hand, these variations allow for microbiome-based identification of individuals, thus posing privacy risks in microbiome studies. Synthetic microbiome datasets hold the promise of reducing said risks while simultaneously keeping the utility of the data for research as high as possible. In this paper, we conduct an empirical evaluation of two open-source data synthetization tools on several publicly available microbiome datasets. In particular, we generate synthetic training data and investigate its performance for a variety of machine learning tasks on microbiome samples. Our findings indicate the suitability of synthetic microbiome data for analysis and for privacy protection.

## 1 Introduction

The composition of bacteria, viruses, fungi and protists on different sites of the human body appear to have a great influence on our health. Research on the human microbiome and its potential for prediction, diagnosis and therapy of diseases has been flourishing for several years. In this context, a particularly well-studied body site is the gastrointestinal tract. Distinct changes in the gut microbiome are related to gastrointestinal diseases [2], obesity [8], diabetes [12], and depression [17]. A most recent study appears to have found a link to the risk for developing post-acute COVID-19 syndrome [10]. While there is a huge potential for analysis in clinical settings, previous research studies have demonstrated that it is important to regard the human microbiome as personal and sensitive medical data. The genetic sequence data obtained from the samples allows for microbiome-based forensic identification of individuals ([3], [23]). However, even the processed, tabular microbiome profiles obtained via the extraction of metagenomic features such as operational taxonomic units (OTUs) show a high individual and temporal stability. In 2015, Franzosa et al. ([4]) revealed that individual variations in such metagenomic features extracted from microbiome

readings allow for the re-identification of individuals among populations of hundreds. Comparing initial microbiome samples with follow-up samples collected 30-300 days later, up to 80% of individuals could still be matched correctly. The authors concluded that the demonstrated possibility of Personal Microbiome Identification (PMI) poses privacy threats to individuals participating in microbiome studies. The findings of Franzosa et al.'s original study have been further aggravated in a recent analysis ([7]), which uses a distance-based approach for the comparison of microbiome samples. Reaching a true-positive identification rate of up to 94%, this technique underlines the need for solutions to protect the privacy of individuals in microbiome databases.

Methods for enhancing the privacy in genomic datasets ([1]) and specific approaches related to human DNA sequence data ([9], [11]) have been studied and improved in the last years. In 2016, Wagner et al. used secure two-party computation, which protects the inputs to the computation, for the analysis of microbiome data ([22]). Besides that, research on privacy-preserving techniques for microbiome data is less advanced. In this paper, we discuss the generation of synthetic data as a solution for privacy-preserving data publishing and for preventing PMI on microbiome reports. The idea of data synthetization is to learn the global properties of an original dataset (that cannot easily be shared), and generate a synthetic dataset, with the aim of preserving these properties and relations between the attributes and without actually revealing the individuals described by the data. The synthetic data can then be shared much more easily and holds the promise to allow data analysis with similar results. For example, we hope to train machine learning models on the synthetic data that ultimately achieve effectiveness comparable to models trained on the real data. Previous publications have shown that the approach of generating synthetic data reduces privacy risks on microdata by simultaneously preserving the utility for performing machine learning tasks (e.g., [5]). We will demonstrate that the same conclusion can be drawn for tasks on metagenomic microbiome profiles.

The remainder of the paper is structured as follows: In Section 2, we will discuss microbiome data, the mentioned techniques for PMI and the data synthetization tools in greater detail. In Section 3, we consider the privacy threats due to PMI and how synthetization can help with reducing these threats. In Section 4, the reader can find the results of our experiments. Section 5 contains concluding remarks, and the appendix provides some additional analysis results.

## 2 Preliminaries & Related Work

We start by discussing the contents of metagenomic microbiome profiles. The datasets in our experiments are tables containing individuals (i.e., sample vectors) described by hundreds to thousands of columns (attributes) that contain the National Center for Biotechnology Information (NCBI) genome identifiers of RefSeq-based OTUs ([14]), which classify groups of closely related organisms found in the microbiome. ,After preprocessing, the values in the tables are the relative abundances of the respective OTU in the sample vector, i.e. they are

normalized row-wise to a unit-vector length. Hence, the attributes contain decimal numbers between 0 and 1 that represent the relative proportion of the OTU in the complete sample of the individual. As a result, the values in each sample vector sum up to 1.

Next, let us briefly discuss the two methods for PMI on metagenomic profiles mentioned in the introduction. In Franzosa et al.'s approach ([4]), a feature is understood as either present or absent, depending on feature detection limits for the relative abundance count (e.g., one limit used is 0.001). Their PMI method then aims to find so-called metagenomic codes that are unique for each individual. These codes are a small subset of features that are present in the individual sample but, in this combination, are not present in any other sample of the dataset. It is then demonstrated in experiments that these codes are stable enough over time such that their comparison may be used to find pairs of samples that belong to the same individual. The second PMI approach ([7]), a distance-based extension of Franzosa et al.'s technique, consists of three phases. In the first phase, the abundance values in the sample vectors are transformed to integral values by using feature abundance limits. The second phase concerns the determination of possible matches. In order to check a specific sample $s$ against a dataset $D$, one computes the most similar (the "nearest-neighbor") sample $\bar{s}$ of $s$ in $D$. By doing so, one obtains $(s, \bar{s})$ as candidate for a pair of samples belonging to the same individual. In the third phase, a threshold criteria is introduced to decrease the false positive count. The candidate pair is accepted only if $\bar{s}$ is a "reverse nearest-neighbor" of $s$, i.e., if there is no sample in $D$ that is closer to $\bar{s}$ than $s$. The method in [7] shows an increased identification rate over the method in [4] on most considered body sites. In particular, one observes an increase of the percentage of true positive identifications of 30% on the gut microbiome, averaged over four different metagenomic feature types.

The present paper is concerned with the mitigation of privacy risks resulting from the PMI techniques, based on the generation of synthetic data. The workflow of synthetic data generation tools may be summarized as follows:

1. Learn the properties of the original dataset, i.e. the distribution of its features and the correlations between them, and store them in some kind of model.
2. Use the model to generate synthetic data samples, e.g. by drawing random data entries from the learned distributions.

By generating enough synthetic samples, we can obtain a synthetic dataset of length equal to the original dataset (or any other arbitrary size). While there is no 1-to-1 correspondence between synthetic samples and real individuals, the hope is that we can use synthetic data for analysis without a significant loss of performance. One of the earliest usages of synthetic data was in the partial synthetic data approach by Rubin [18], where certain columns are generated synthetically. An overview on more than 20 approaches is given in [19], categorizing the approaches into fully or partially synthetic.

For selecting the synthetic data generation tools included in our analysis, we focused on recent methods that are open-source implementations and utilize generative models to ensure robust and high-quality data. In addition, the

techniques need to be able to deal with the relatively large (hundreds to thousands) number of attributes in metagenomic microbiome profiles. We identified the following two tools.

The **Synthetic Data Vault** (SDV) has been developed by Patki et al. in 2016 ([15]) [3]. The SDV allows to synthesize both single table datasets and relational databases. The user can choose between different models for learning the patterns on the original data, such as fitting Gaussian Copulas or using GANs.

The **synthpop** (SP) package for the statistical analysis language *R* ([13]) [4] uses as the default synthesis method the CART (Classification and Regression Trees) algorithm. However, the user is able to specify a different method.

Another tool, the Data Synthesizer ([16]), is not included in our analysis due to efficiency issues during the construction of its Bayesian network. Possible speed-ups for this tool are a prospect for future work (see Section 5).

## 3 Threat Model & Goal of Synthetization

We now discuss the notions related to a privacy violation of microbiome data. In addition, we will consider our goal to mitigate these threats by publishing synthetic instead of original microbiome datasets.

**Victim**. An individual who provided their microbiome samples in the course of medical studies, diagnosis and therapy of diseases, personal health and fitness advise, or similar. Their microbiome data, and possible analysis results or metadata, are publicly available, e.g. in the form of (possibly deidentified) samples.

**Adversary**. A party in possession of microbiome samples with the intention to link them to other microbiome samples. The purpose is to reveal the identity of the underlying individuals and/or to obtain more information about them. The adversary's options to obtain microbiome samples include public microbiome databases, cyberattacks against healthcare and research facilities, data exfiltration via insiders, and potentially, directly from the victim (e.g., saliva).

**Threats**. Assume that an adversary possesses a sample of a certain individual. We discuss four reasons for the adversary to match the sample against another database.

(i) To find out if an individual participated in a certain study, i.e. a form of membership disclosure. This might allow them to obtain sensitive information, e.g. in the case where the study has been conducted in the context of specific diseases.
(ii) The attacker may be able to obtain previously unknown metadata linked to the identified sample from the new database (e.g., medical and personal data provided in the course of a study or treatment, or questionnaires).
(iii) Even in the absence of metadata, the attacker may be able to get hold of new microbiome samples from the same individual and could thereby learn about changes over time in the individual's human microbiome. Such changes could, e.g., point to physical diseases, depression, and changes in diet.

[3] `https://github.com/HDI-Project/SDV`
[4] `https://www.synthpop.org.uk/` resp. `https://github.com/bnowok/synthpop`

(iv) Ongoing research efforts increasingly associate microbiome samples with several other individual traits, such as the age or geographical background [24]. Therefore, collecting and linking multiple samples from the same individual could also aid an attacker in identifying the person behind a sample.

We will use synthetization to reduce the risks of the threats (i)-(iv). As we have already discussed in Section 2, there is no 1-to-1 correspondence between synthetic and real samples. An adversary may still try to apply PMI techniques to match a sample in their possession against a synthetic dataset. However, any match that is potentially found must be considered incidental, as no synthetic sample is related to a particular individual. In addition, no metadata is linked to the samples in a synthetic dataset, which eliminates threat (ii). Since identity disclosure in the form discussed above does not appear to be possible, previous publications about privacy risks on synthetic data ([20], [6]) have focused on the notion of *attribute disclosure*. It refers to the risk that, on structured tabular datasets, an adversary might learn the victim's value of a sensible attribute. It has been concluded that attribute disclosure can happen without identity disclosure, and that there remains a risk for attribute disclosure on synthetic data. However, it appears that this conclusion does not translate to microbiome datasets. The attributes in these tables are particularly numerous. While the complete composition of the microbiome is rich with information, it is much less obvious that knowledge about a single, specific feature in a sample could be particularly sensitive. While we do not completely disregard the possibility of attribute disclosure, we will not investigate it further in the present paper.

The promise of synthetic data is a close similarity of its global properties to the original data. From a privacy-preserving point of view, this would ideally come with a substantial dissimilarity of its local properties. Synthetic samples that are very similar to real ones might still allow to deduce local information about the original dataset, even without the immediate threat of identity disclosure. In previous publications (e.g., [5]), this question has been investigated by considering the distance between synthetic samples and their nearest-neighbor in the original dataset. We will conduct a similar analysis in Section 4.2.

## 4 Evaluation

Our experiments are based on the publicly available datasets from "Knightslab"'s microbiome machine learning repository[5] [21]. In total, the repository contains data for 33 curated machine learning tasks, mainly for binary classification. We utilized six medium-sized datasets: two for distinguishing healthy microbiome samples from those where the hosts suffer from Morbus Crohn (the "Gevers" datasets in the repository), two further datasets with tasks on Irritable Bowel Disease (the "Morgan" datasets), one for distinguishing lean from obese individuals (the "Turnbaugh" dataset), and one for detecting tumors (the "Kostic" dataset). In each case, we will use data containing RefSeq-based OTU

[5] `https://knights-lab.github.io/MLRepo/`

abundance counts to compare our results directly to the baselines shown in [21, p.5]. Table 1 provides an overview of the original dataset properties.

Table 1: Dataset characteristics prior to any preprocessing

| Dataset | Task | # Samples | # Features |
|---|---|---|---|
| Gevers Ileum | control vs crohn's disease | 140 | 943 |
| Gevers Rectum | control vs crohn's disease | 160 | 943 |
| Morgan CD | healthy vs crohn's disease | 128 | 829 |
| Morgan UC | healthy vs ulcerative colitis | 128 | 829 |
| Turnbaugh | lean vs obese | 142 | 557 |
| Kostic | healthy vs tumor biopsy | 172 | 908 |

For their case study benchmarking, the authors of [21] used an experimental setup including feature preprocessing steps and a repeated 5-fold cross-validation. We will use a similar setup and briefly summarize it here:

1. The OTU counts in the original tables are converted to relative abundance, such that the sum of all values in each sample vector equals 1. In addition, the counts were filtered at a minimum of 10% prevalence across samples and collapsed at a complete-linkage correlation of 95%.
2. The relative abundances in the tables are transformed by using the same feature abundance limits as in [7], namely $t_i = 0.00005 \cdot 10^i$ for $i = 0, 1, 2, 3, 4$. For each cell value $x$, we then set $x \leftarrow 0$ if $x < t_0$, $x \leftarrow i$ if $t_{i-1} \leq x < t_i$ for some $1 \leq i \leq 4$, and $x \leftarrow 5$ if $x \geq t_4$.
3. Like the authors of the original publication [21], we apply a 5-fold cross validation to split up our datasets into training and test data. Note that the samples are assigned to the folds such that the target class is equally balanced within each fold, i.e. the folds are stratified.
4. For each fold, we applied the tools discussed in Section 2 to the training dataset to generate synthetic training data of equal length. For synthpop, we used the CART ("Classification and Regression Trees") model and set the "minbucket" parameter[6] to 1. For the Synthetic Data Vault, we used the GaussianCopula model, and passed an argument to the model that a categorical transformer should be used for the target column.[7]
5. Finally, we train our machine learning models on both the original and the synthetic training datasets, and compare their performances on the test dataset of the respective fold. We applied Random Forest and Support Vector Machines (SVM) with radial and linear kernel and default parameters.
6. This entire process is repeated ten times, and the mean class probabilities are used to calculate the ROC-AUC score (see Section 4.1).

[6] This parameter refers to the minimum number of observations in any terminal note. A higher value leads to synthetic data with larger differences to original samples.

[7] The default transformation for the GaussianCopula model is One Hot Encoding. We noticed that, with this default setting, it often occurred that only one of the two possible values of the target column is generated, leading to a constant column in the synthetic training data. Using the categorical transformation solves this problem.

To summarize, we use the very same experimental setup as the authors of [21][8], with one notable difference: we apply the feature abundance limits in Step 2 to transform the floating point numbers in the sample vectors to integer values that represent the abundance. The main idea behind this is to split up the complete range of values in an (geometrically) even way, emphasizing the corresponding levels of abundance. Comparing the performance baselines established in [21, p.5] to our results on the original data in the subsequent subsection, we observe increased AUC-scores for Support Vector Machines. This indicates a slight benefit from this data preprocessing step.

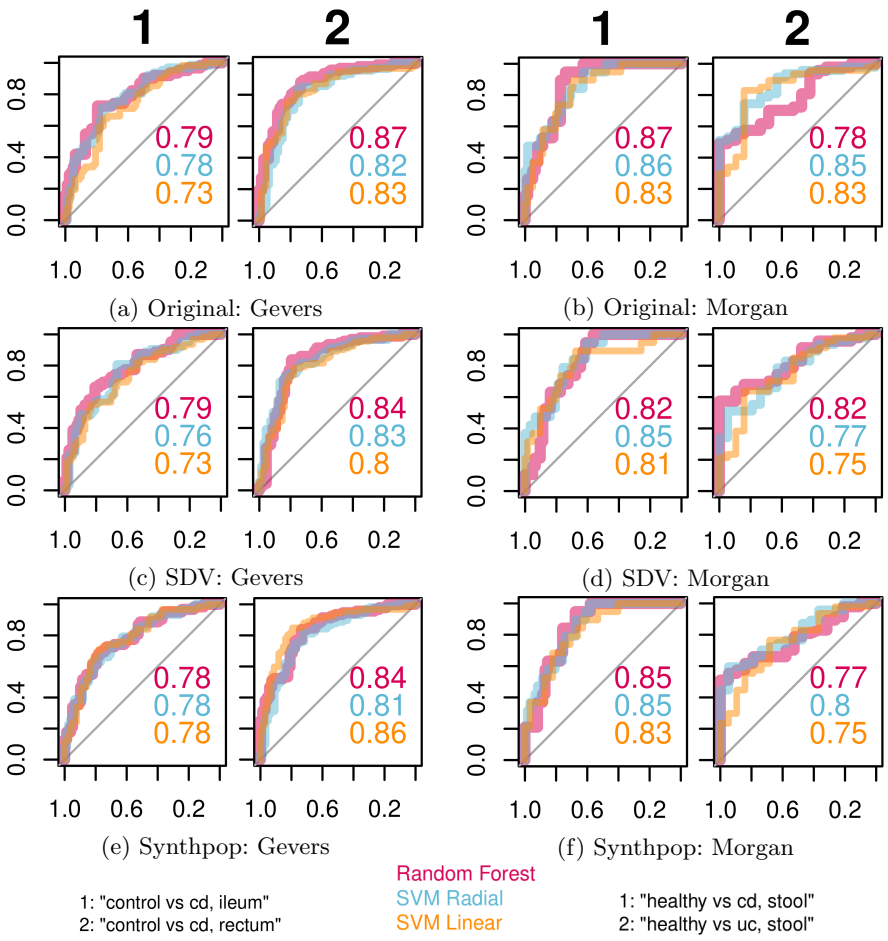
### 4.1 Task Utility

We now consider the results of the experiment described above. For comparing the performance of the machine learning algorithms, we used a Receiver Operating Characteristic (ROC) curve just like in the original publication ([21]). The following plots compare sensitivity (or "true-positive rate") on the y-axis to specificity (or "true-negative rate") on the x-axis for various probability thresholds. The Area Under the Curve (AUC) scores are listed within the plots and colored according to the ML model they refer to.

Let us start by summarizing the results of the two "Gevers" datasets (Figure 1). Comparing the scores of ML models trained on SDV data instead of the original data, we can see that there are only minor differences in the effectiveness. Data generated via synthpop also shows scores that are very close to the original data. It is notable that we even have a slightly better score of SVM Linear on both datasets. Let us proceed with the two "Morgan" datasets (Figure 1). On SDV data, we again have very close scores on the first dataset, with a slightly larger drop for the Random Forest model. On the second dataset, Random Forest performed even better than on the original data, but SVM Linear performed substantially worse and shows a drop of the AUC score of 8%. Similar conclusions can be drawn for synthpop, which performed slightly better on the first dataset and slightly worse on the second. On the "Turnbaugh" dataset (Figure 2), we again obtain good scores for both the SDV and synthpop. However, synthpop shows better results than the SDV, and the difference is particularly pronounced for SVM Linear. These roles are reversed on the "Kostic" dataset (Figure 2), where SDV shows a particular high score on the Random Forest model.

On average, we observe a satisfying performance on synthetic data that is on par with previous experimental results for the evaluation of synthetic microdata for machine learning tasks (e.g., see [5]). SDV and synthpop are on a similar performance level. However, both methods appear to work better or worse in certain scenarios. For example, SVM with a linear kernel appears to perform exceptionally well on some of the synthpop datasets, while Random Forest performs well on some of the SDV datasets.

[8] An implementation of their experiments in the programming language R can be found on Github (`https://github.com/knights-lab/MLRepo`). We have modified their code and called Python scripts (e.g. for the SDV) with the package 'reticulate' (`https://rstudio.github.io/reticulate/index.html`).




(a) Original: Gevers (b) Original: Morgan

(c) SDV: Gevers (d) SDV: Morgan

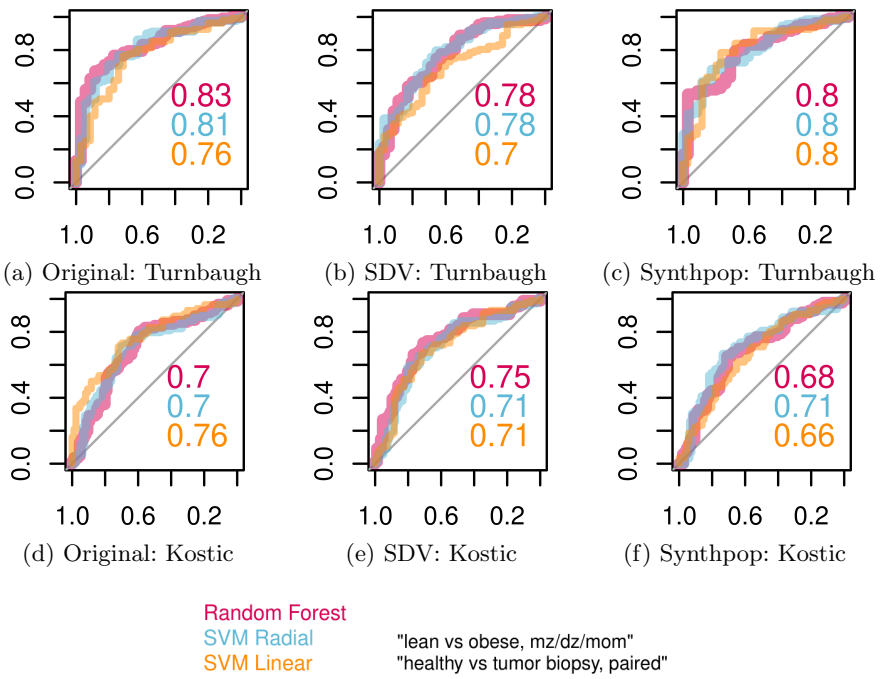
(e) Synthpop: Gevers (f) Synthpop: Morgan

Fig. 1: Results on Gevers and Morgan: Final ROC curves and AUC scores

### 4.2 Sample Similarity and Privacy Risks

The goal of the synthetization is to keep the *global* dataset properties as close to the original data as possible. On the other hand, we do not want to preserve the *local* properties of the original data, as these may allow to deduce information about real individual samples (see Section 3). While there is no 1-to-1 relation between synthetic and original samples, there may be privacy risks related to close local similarities between original and synthetic data samples, leading to near-matches. We measure this similarity as follows:

1. For each synthetic sample $s$, we computed the nearest neighboring original sample, i.e., the sample with minimal euclidean distance $d_s$ to $s$.
2. We compute the mean and the variance of these minimal distances, i.e., the mean and the variance over $d_s$ for all samples $s$ in the synthetic dataset.




(a) Original: Turnbaugh (b) SDV: Turnbaugh (c) Synthpop: Turnbaugh

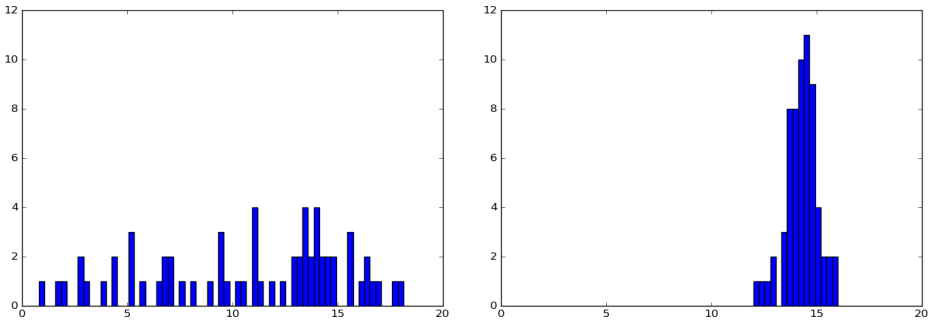
(d) Original: Kostic (e) SDV: Kostic (f) Synthpop: Kostic

Fig. 2: Results on Turnbaugh and Kostic

The results of a randomly chosen synthetic dataset for each of the processed original datasets are reported in Table 2. We can see that the distances $d_s$ are on average reasonably large on both the SDV and on synthpop. In a direct comparison, however, synthpop's samples are considerably closer to the original ones than those obtained from SDV. In addition, synthpop shows a large variance in these local similarities of the datasets, and some samples are close or even equal to original data samples. Hence, datasets produced by synthpop seem to contain local information that is very similar to original data, thus leading to increased privacy risks for those vulnerable samples. SDV does not show the same behavior and may hence provide an increased reduction of privacy vulnerabilities.

Table 2: Mean minimal distance between original and synthetic nearest neighbors

| Dataset | SDV | SPO |
|---|---|---|
| Gevers Ileum | $11.322 \pm 1.386$ | $8.688 \pm 24.527$ |
| Gevers Rectum | $10.925 \pm 1.240$ | $7.510 \pm 21.120$ |
| Morgan CD | $14.301 \pm 0.565$ | $10.744 \pm 20.577$ |
| Morgan UC | $14.048 \pm 0.735$ | $9.772 \pm 22.701$ |
| Turnbaugh | $12.824 \pm 0.588$ | $9.141 \pm 24.413$ |
| Kostic | $14.970 \pm 0.627$ | $11.299 \pm 30.380$ |



(a) Synthpop (b) SDV

Fig. 3: Distance plots for the Morgan CD dataset

In Figure 3, we provide a histogram showing the minimal distances for the "Morgan" CD dataset. The plot visualizes the tendencies observed in Table 2 and looks similar for the other considered datasets.

## 5 Conclusions and Future Work

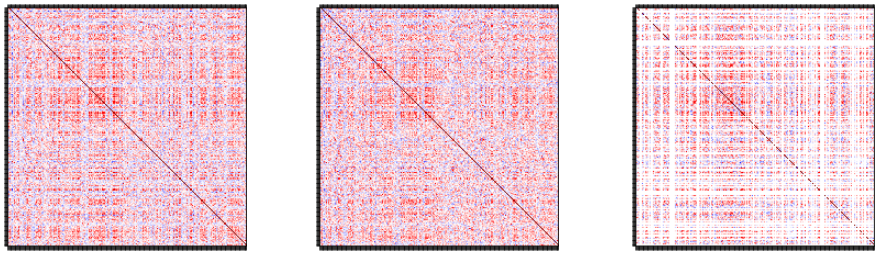
In this paper, we evaluated two data synthetization methods, the *Synthetic Data Vault* and *synthpop*, for machine learning tasks on microbiome data. Both the SDV and synthpop performed well on the considered tasks, with AUC scores mostly deviating about $\pm 3\%$ from those on the original data. In our assessment of the privacy risks, we observed that synthpop generates a number of synthetic microbiome samples very close to original ones, indicating that there may remain higher privacy risks on synthpop than on SDV. However, we point out that synthpop comes with parameters (e.g., the already discussed 'minbucket' parameter) allowing to generate datasets with smaller or larger deviations from the original data. Since there is a trade-off between the utility and the privacy risk reduction on synthetic data, the option to customize the synthetic datasets may be useful in specific scenarios.

Our future work will concern the development of speed-ups for the Data Synthesizer and similar tools based on Bayesian networks. While models based on these networks usually generate data of high quality, the large number of features in microbiome tables leads to problems related to the efficient construction of these networks. Moreover, a comparison with other privacy-preserving data publishing methods, such as k-anonymity or micro-aggregation, would help to put the achieved performance of synthetic data into perspective.

**Acknowledgements** This work was partially funded by the Austrian Research Promotion Agency FFG under grant 877173 (GASTRIC). SBA Research (SBA-K1) is a COMET Centre within the framework of COMET – Competence Centers for Excellent Technologies Programme and funded by BMK, BMDW, and the federal state of Vienna. The COMET Programme is managed by FFG.

## Appendix: Heatmap Comparison

Figure 4 shows heatmaps for the "Kostic" dataset to demonstrate the ability of the synthesizers to preserve correlations between attributes. Red color in the feature matrix indicates a positive (Pearson) correlation coefficient between the respective features, while blue color indicates a negative coefficient. Synthpop's heatmap shows great similarities to the original. The heatmap on the SDV dataset is also similar, but to a lesser extent. In addition, there are features on the SDV dataset that are constant, and for which the correlation coefficient is not defined. Similar patterns can also be observed on the other datasets.



(a) Original (b) Synthpop (c) SDV

Fig. 4: Heatmaps on the Kostic dataset

## References


1. Berger, B., Cho, H.: Emerging technologies towards enhancing privacy in genomic data sharing. Genome Biol **20**(128) (2019). https://doi.org/10.1186/s13059-019-1741-0
2. Distrutti, E., Monaldi, L., Ricci, P., Fiorucci, S.: Gut microbiota role in irritable bowel syndrome: New therapeutic strategies. World J Gastroenterol **22**(7) (2016)
3. Fierer, N., Lauber, C.L., Zhou, N., McDonald, D., Costello, E.K., Knight, R.: Forensic identification using skin bacterial communities. PNAS **107**(14) (2010)
4. Franzosa, E., Huang, K., Meadow, J., Gevers, D., Lemon, K., Bohannan, B.: Identifying personal microbiomes using metagenomic codes. PNAS **112**(22) (2015)
5. Hittmeir, M., Ekelhart, A., Mayer, R.: On the Utility of Synthetic Data: An Empirical Evaluation on Machine Learning Tasks. In: International Conference on Availability, Reliability and Security. ARES, ACM Press, Canterbury, United Kingdom (2019). https://doi.org/10.1145/3339252.3339281
6. Hittmeir, M., Mayer, R., Ekelhart, A.: A Baseline for Attribute Disclosure Risk in Synthetic Data. In: ACM Conference on Data and Application Security and Privacy. CODASPY, ACM, New Orleans LA USA (Mar 2020). https://doi.org/10.1145/3374664.3375722
7. Hittmeir, M., Mayer, R., Ekelhart, A.: Distance-based Techniques for Personal Microbiome Identification. In: International Conference on Availability, Reliability and Security. ARES, ACM, Vienna, Austria (Aug 2022). https://doi.org/10.1145/3538969.3538985

8. Ley, R., Turnbaugh, P., Klein, S., Gordon, J.: Microbial ecology: human gut microbes associated with obesity. Nature **444**(7122) (2006)
9. Li, G., Wang, Y., X., S.: Improvements on a privacy-protection algorithm for dna sequences with generalization lattices. Computer Methods and Programs in Biomedicine **108**(1) (2012)
10. Liu, Q., Mak, J.W.Y., Su, Q., Yeoh, Y.K., Lui, G.C.Y., Ng, S.S.S., Zhang, F., Li, A.Y.L., Lu, W., Hui, D.S.C., Chan, P.K., Chan, F.K.L., Ng, S.C.: Gut microbiota dynamics in a prospective cohort of patients with post-acute covid-19 syndrome. Gut **71**(3) (2022). https://doi.org/10.1136/gutjnl-2021-325989
11. Malin, B.A.: Protecting genomic sequence anonymity with generalization lattices. Methods Inf Med. **44**(5) (2005)
12. Musso, G., Gambino, R., Cassader, M.: Obesity, diabetes, and gut microbiota: the hygiene hypothesis expanded? Diabetes Care **33**(10) (2010)
13. Nowok, B., Raab, G., Dibben, C.: synthpop: Bespoke creation of synthetic data in r. Journal of Statistical Software, Articles **74**(11) (2016). https://doi.org/10.18637/jss.v074.i11
14. O'Leary, N., Wright, M., Brister, J., et al.: Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res **44** (2016)
15. Patki, N., Wedge, R., Veeramachaneni, K.: The synthetic data vault. In: IEEE International Conference on Data Science and Advanced Analytics (DSAA). Montreal, QC, Canada (October 17–19 2016)
16. Ping, H., Stoyanovich, J., Howe, B.: Datasynthesizer: Privacy-preserving synthetic datasets. In: International Conference on Scientific and Statistical Database Management. Chicago, IL, USA (June 27–29 2017)
17. Rogers, G., Keating, D., Young, R., Wong, M., Licinio, J., Wesselingh, S.: From gut dysbiosis to altered brain function and mental illness: mechanisms and pathways. Mol Psychiatry **21**(6) (2016)
18. Rubin, D.B. (ed.): Multiple Imputation for Nonresponse in Surveys. Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., Hoboken, NJ, USA (Jun 1987). https://doi.org/10.1002/9780470316696
19. Surendra, H., Mohan, H.S.: A review of synthetic data generation methods for privacy preserving data publishing. International Journal of Scientific & Technology Research **6**(3) (March 2017)
20. Taub, J., Elliot, M., Pampaka, M., Smith, D.: Differential Correct Attribution Probability for Synthetic Data: An Exploration. In: Domingo-Ferrer, J., Montes, F. (eds.) Privacy in Statistical Databases. Lecture Notes in Computer Science, Springer International Publishing, Valencia, Spain (2018). https://doi.org/10.1007/978-3-319-99771-1_9
21. Vangay, P., Hillmann, B.M., Knights, D.: Microbiome Learning Repo (ML Repo): A public repository of microbiome regression and classification tasks. GigaScience **8**(5) (2019). https://doi.org/10.1093/gigascience/giz042
22. Wagner, J., Paulson, J.N., Wang, X., Bhattacharjee, B., Corrada Bravo, H.: Privacy-preserving microbiome analysis using secure computation. Bioinformatics **32**(12) (02 2016). https://doi.org/10.1093/bioinformatics/btw073
23. Woerner, A.E., Novroski, N.M., Wendt, F.R., Ambers, A., Wiley, R., Schmedes, S.E., Budowle, B.: Forensic human identification with targeted microbiome markers using nearest neighbor classification. Forensic Science International: Genetics **38** (2019)
24. Yatsunenko, T.: Human gut microbiome viewed across age and geography. Nature **486**(7402) (2012)