
SBA
Research

SBA Research

COMET

Competence Centers for Excellent Technologies

Utility and Privacy Assessment of Synthetic Microbiome Data

M. Hittmeir, R. Mayer, A. Ekelhart

36th DBSEC, 18.07.2022, Newark, USA

Bundesministerium Verkehr, Innovation und Technologie

Bundesministerium Digitalisierung und Wirtschaftsstandort

FFG Forschung wirkt.

wirtschaftsagentur wien Ein Fonds der Stadt Wien

FWF Der Wissenschaftsfonds.

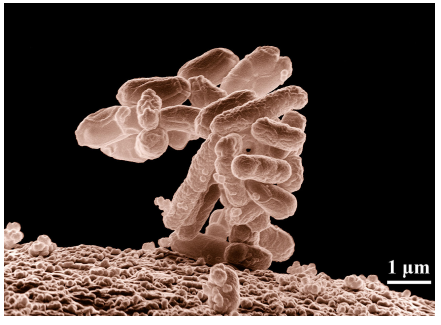
netidee OPEN INNOVATIONS

The Human Microbiome

Microorganisms in and on the human body, such as bacteria, fungi and viruses

Examples for body sites hosting microorganisms are:

- Organs such as the skin and the lung
- The mouth: teeth, gums and saliva
- The gastrointestinal tract



Our body needs the microbiome to **function properly**
Dysfunctions in the microbiome are linked to several diseases

Microbiome Data

	158337416	158499257	158883629	158802708	158944319
|o__Actinomycetales|f__Actinomycetaceae|g__Actinomyces|s__Actinomyces_odontolyticus	0	0.00158368	0.00469286	0.000165606	0.000459244
|o__Actinomycetales|f__Actinomycetaceae|g__Actinomyces|s__Actinomyces_oris	0.002494	0.0092369	0.00176714	0.000564819	0.000773189
|o__Actinomycetales|f__Actinomycetaceae|g__Actinomyces|s__Actinomyces_urogenitalis	0	0	0	0	0
|o__Actinomycetales|f__Actinomycetaceae|g__Actinomyces|s__Actinomyces_viscosus	0.0183326	0.00331386	0.00320359	0.000767926	0.00163145
|o__Actinomycetales|f__Corynebacteriaceae|g__Corynebacterium|s__Corynebacterium_accolens	0	0	0.000277568	0.000135905	0
|o__Actinomycetales|f__Corynebacteriaceae|g__Corynebacterium|s__Corynebacterium_matruchotii	0.00131085	0.00259662	0.000875916	0.000187606	0.00113293
|o__Actinomycetales|f__Corynebacteriaceae|g__Corynebacterium|s__Corynebacterium_tuberculostearicum	0	0	0	0	0
|o__Actinomycetales|f__Micrococcaceae|g__Rothia|s__Rothia_dentocariosa	0.000544351	0.0267706	0.0247112	0.00516867	0.00410293
|o__Actinomycetales|f__Micrococcaceae|g__Rothia|s__Rothia_mucilaginosa	0.011687	0.0137408	0.0187899	0.0028885	0.000370662
|o__Actinomycetales|f__Micrococcaceae|g__Rothia|s__Rothia_unclassified	0	0	0	0	0
|o__Actinomycetales|f__Mycobacteriaceae|g__Mycobacterium|s__Mycobacterium_unclassified	0	0.000153635	0.000226156	0	0
|o__Actinomycetales|f__Propionibacteriaceae|g__Propionibacterium|s__Propionibacterium_acnes	0.00291646	0.00109022	0.013879	0.00277459	0.00015081
|o__Actinomycetales|f__Propionibacteriaceae|g__Propionibacterium|s__Propionibacterium_unclassified	0	0.00112488	0.000709575	0.000331311	0.000100607
|o__Bifidobacteriales|f__Bifidobacteriaceae|g__Bifidobacterium|s__Bifidobacterium_adolescentis	0	0	0	0	0
|o__Bifidobacteriales|f__Bifidobacteriaceae|g__Bifidobacterium|s__Bifidobacterium_dentium	0	0	0	0	0
|o__Bifidobacteriales|f__Bifidobacteriaceae|g__Bifidobacterium|s__Bifidobacterium_longum	0	0	0	0	0
|o__Bifidobacteriales|f__Bifidobacteriaceae|g__Bifidobacterium|s__Bifidobacterium_unclassified	0	0	0	0	0
|o__Bifidobacteriales|f__Bifidobacteriaceae|g__Gardnerella|s__Gardnerella_vaginalis	0	0	0	0	0
|o__Bifidobacteriales|f__Bifidobacteriaceae|g__Parascardovia|s__Parascardovia_denticolens	0	0	0	0	0
|o__Coriobacteriales|f__Coriobacteriaceae|g__Atopobium|s__Atopobium_parvulum	0	0.000174234	0.000171442	0	1.51311e-05
|o__Coriobacteriales|f__Coriobacteriaceae|g__Atopobium|s__Atopobium_rimae	0	0.000112538	0.000215353	0	0
|o__Coriobacteriales|f__Coriobacteriaceae|g__Atopobium|s__Atopobium_vaginae	0	0	0	0	0
|o__Coriobacteriales|f__Coriobacteriaceae|g__Collinsella|s__Collinsella_aerofaciens	0	0	0	0	0
|o__Coriobacteriales|f__Coriobacteriaceae|g__Cryptobacterium|s__Cryptobacterium_curtum	0	0	0	0	0
|o__Coriobacteriales|f__Coriobacteriaceae|g__Olsenella|s__Olsenella_uli	0	0	0	0	0

Extract from a report on microbial species found at ‘buccal mucosa’ (inside of the cheek)

Relative abundance: Each column (sample vector) sums up to 1.

Personal Microbiome Identification

Q: Is it possible to identify individuals in a microbiome database?

We consider two datasets like above:

- D_1 with samples at some initial point in time
- D_2 with samples (from the same individuals) at a later time

Task of PMI

For each sample in D_2, identify samples from the same individual in D_1.

Two main approaches:

1. Franzosa et al. (2015): Based on comparison of most abundant and stable features
2. H. et al. (2022): Based on computation of distances between sample vectors ("nearest-neighbors")

Results

- Up to 94% correct re-identifications on gut microbiomes
- High temporal stability and individual uniqueness

Data Synthetization

Q: Can we prevent PMI and still make the data available?

We are not always interested in local details of the data.
The analysis often focuses more on **global trends**.

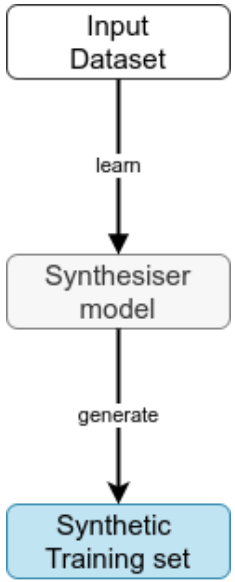
Idea: Publish some data that resembles the real data

- Preserve global characteristics:

  Distribution of attributes, correlations between them

- Published data does not contain real individuals

General workflow of data synthesizers:



1. Data Description
    - Original data is used to build a model
    - Information about distributions and correlations, etc.
2. Data Generation
    - Model is used to generate data samples
    - Global properties of resulting synthetic dataset are similar to the original…
    - …but the samples do not represent real individuals (No 1-to-1 correspondence)

Data Synthesizer Tools

We considered two freely available tools.

1. **Synthetic Data Vault** (Python): N. Patki et al., 2016
    - Different models for learning
    - We used method based on Gaussian Copulas
2. **Synthpop** (R): B. Nowok et al., 2016
    - Highly customizable
    - We used the default synthesis method: CART

Experimental Setup

- We used six datasets from the “Knights-lab” repository[1]
- 128-172 gut MB samples and 557-943 features
- Classification tasks concerning diseases

Preparation

- Preprocessing specific to MB data (filtering, binning)
- Stratified 5-fold cross validation to get train and test data
- ML models: Random Forest and Support Vector Machine

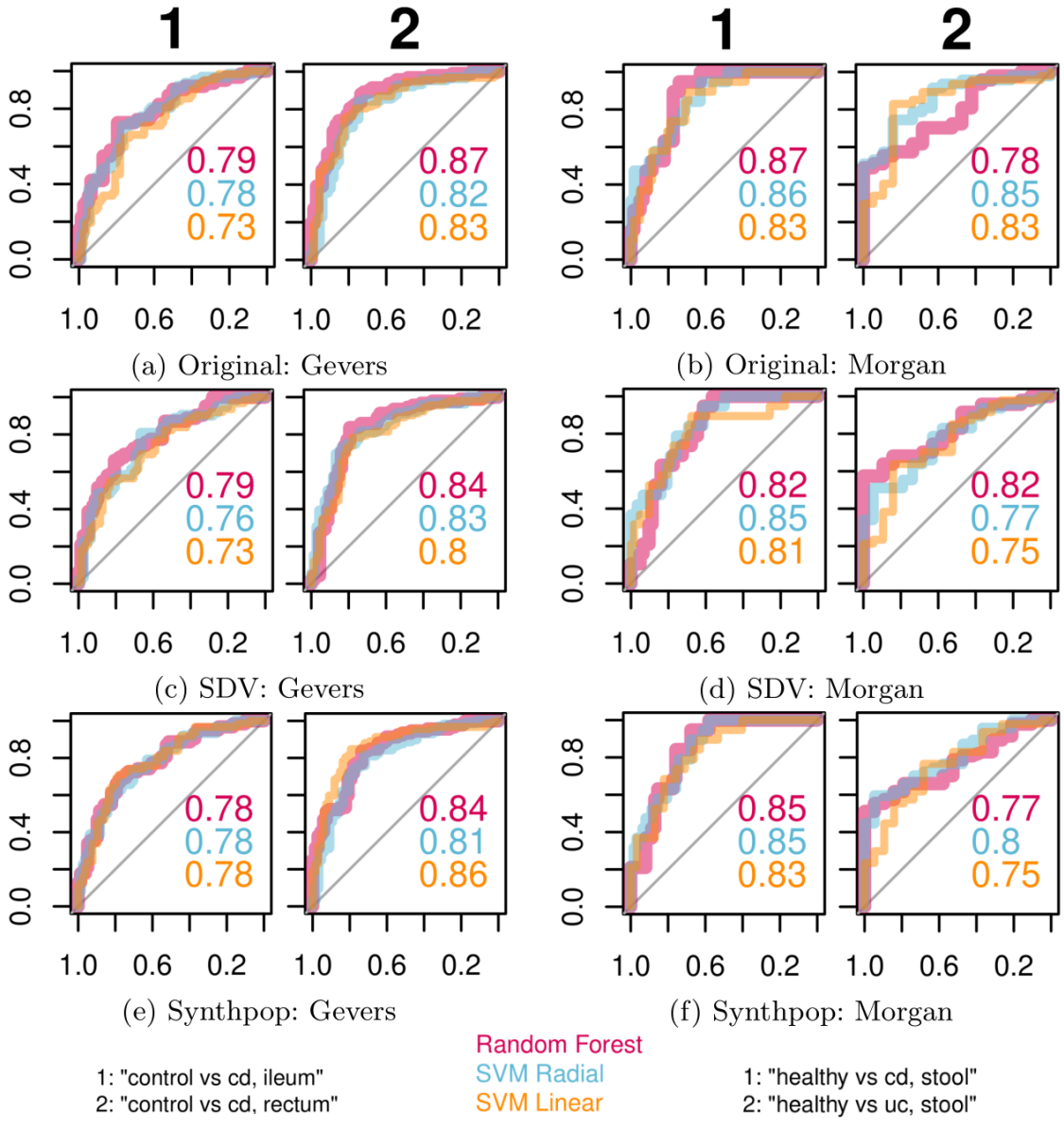
[1] https://knights-lab.github.io/MLRepo/

For each split:

1. Apply the data synthesizers to the training data
2. Use the original and the synthetic training datasets as input for the ML models
3. Evaluate their performance on the same test data, using ROC-AUC

The overall process is repeated 10 times to get reliable results

Results



(a) Original: Gevers

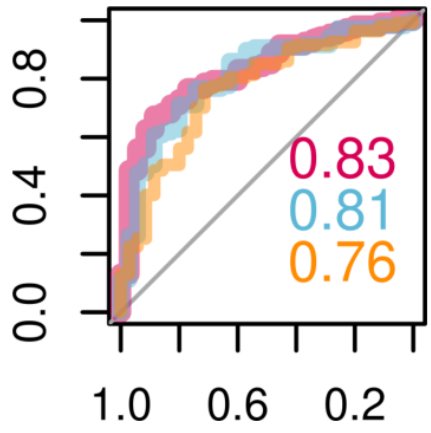
(b) Original: Morgan

(c) SDV: Gevers

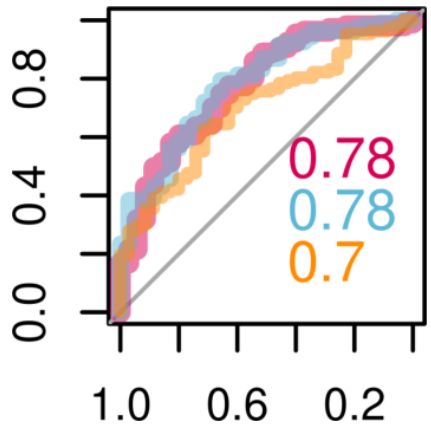
(d) SDV: Morgan

(e) Synthpop: Gevers

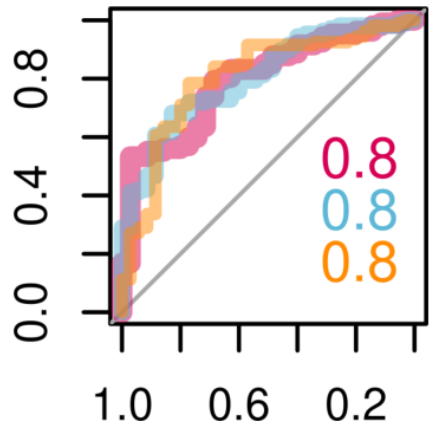
(f) Synthpop: Morgan



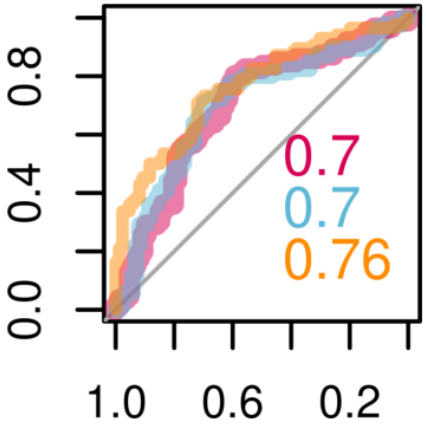
(a) Original: Turnbaugh



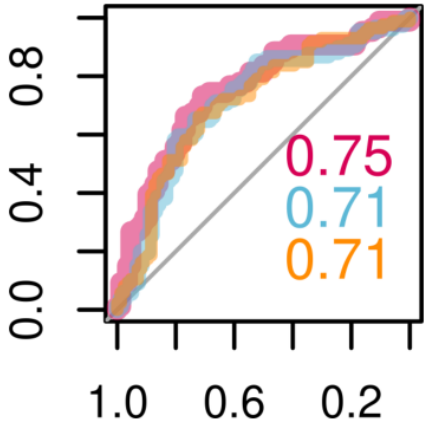
(b) SDV: Turnbaugh



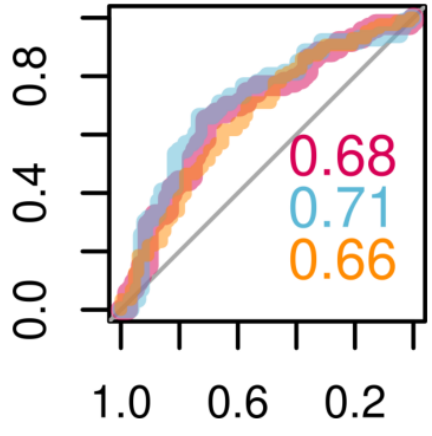
(c) Synthpop: Turnbaugh



(d) Original: Kostic



(e) SDV: Kostic



(f) Synthpop: Kostic

Random Forest
SVM Radial
SVM Linear

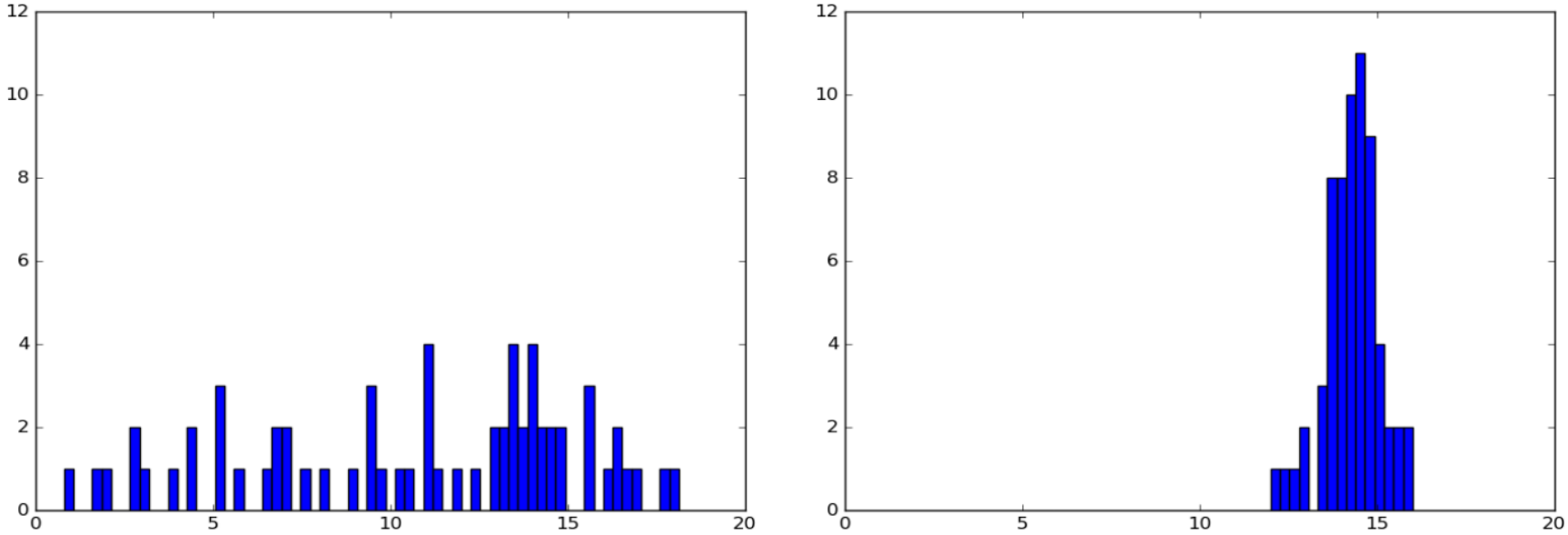
"lean vs obese, mz/dz/mom"
"healthy vs tumor biopsy, paired"

Privacy Assessment

- No 1-to-1 relation between synthetic and original samples
- However, are there close local similarities?
- If yes, there might be vulnerable original records

Sample Similarity Check

For each synthetic sample s: Find the minimal distance d_s to a sample in the original dataset ("nearest neighbor")



Morgan CD dataset; X-axis: minimum distance; Y-axis: number of records

- Synthpop generates samples close to original records
- SDV produces much larger differences on average

Summary

- Both SDV and synthpop performed well
- AUC scores mostly $\pm 5\%$ from original
- synthpop generates vulnerable samples
  SDV seems “safer”
- However, synthpop allows trade-off between utility and privacy risk reduction

References

1. Franzosa, E., Huang, K., Meadow, J., Gevers, D., Lemon, K., Bohannan, B.: Identifying personal microbiomes using metagenomic codes. PNAS 112(22), E2930–E2938 (2015)
2. Hittmeir, M., Mayer, R., Ekelhart, A.: Distance-based techniques for personal microbiome identification. ARES 2022, to appear, Link: https://tinyurl.com/5htduzfu
3. N. Patki, R. Wedge, K. Veeramachaneni, The Synthetic Data Vault, In: Proceedings of the 3rd DSAA (2016)
4. B. Nowok, G. M. Raab, C. Dibben, synthpop: Bespoke Creation of Synthetic Data in R, In: Journal of Statistical Software (2016)

Questions?

mhittmeir@sba-research.org