UNIVERSITÄT LEIPZIG
Medizinische Fakultät

Universitätsklinikum Leipzig
Medizin ist unsere Berufung.

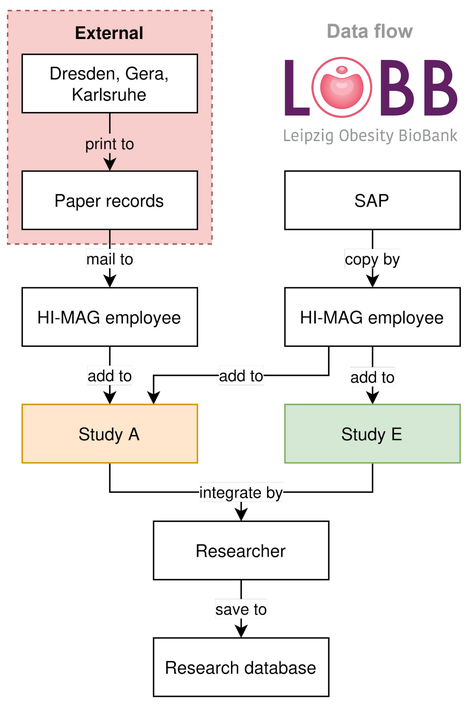
# Improving record linkage quality on identification data in the Leipzig Obesity BioBank

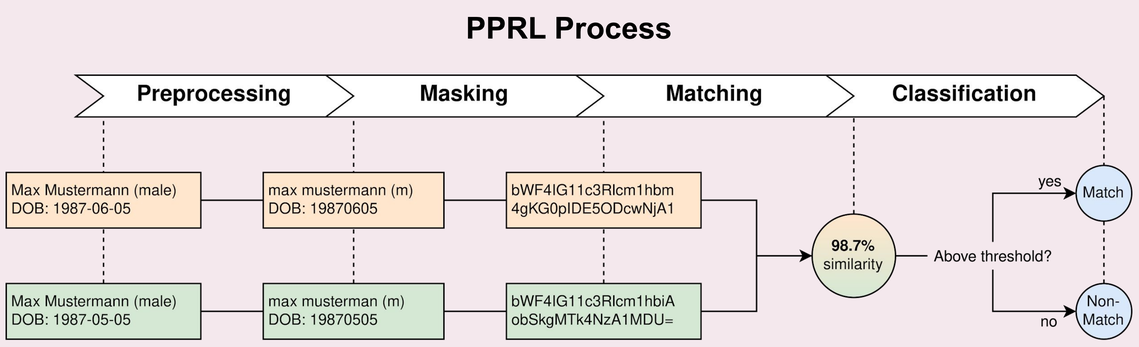
Maximilian Jugl[1,2], Anne Hoffmann[3], Matthias Kern[3,4], Matthias Blüher[3,4], Thomas Ebert[4], Toralf Kirsten[1,2]

[1]University of Leipzig Medical Center, Medical Informatics Center, Dept. Medical Data Science, Leipzig, [2]Leipzig University, Institute of Medical Informatics, Statistics, and Epidemiology, Leipzig, Germany, [3]Helmholtz Institute for Metabolic, Obesity and Vascular Research (HI-MAG), Helmholtz Zentrum München, University of Leipzig and University Hospital Leipzig, Leipzig, Germany, [4]University of Leipzig Medical Center, Department of Endocrinology, Nephrology and Rheumatology, Leipzig, Germany

## Background

- Leipzig Obesity BioBank (LOBB): longitudinal medical study researching diseases related to obesity with 8,000+ patients
- LOBB seeks to integrate medical data from multiple sources with varying input methods
- Privacy-preserving record linkage (PPRL) requires real-world data to test algorithms[1]
- Few studies based on real-world identification data (IDAT)[2,3,4] ⇒ generation of realistic data feasible[5]
- **Use LOBB to analyse error sources and generate realistic data generation workflow**



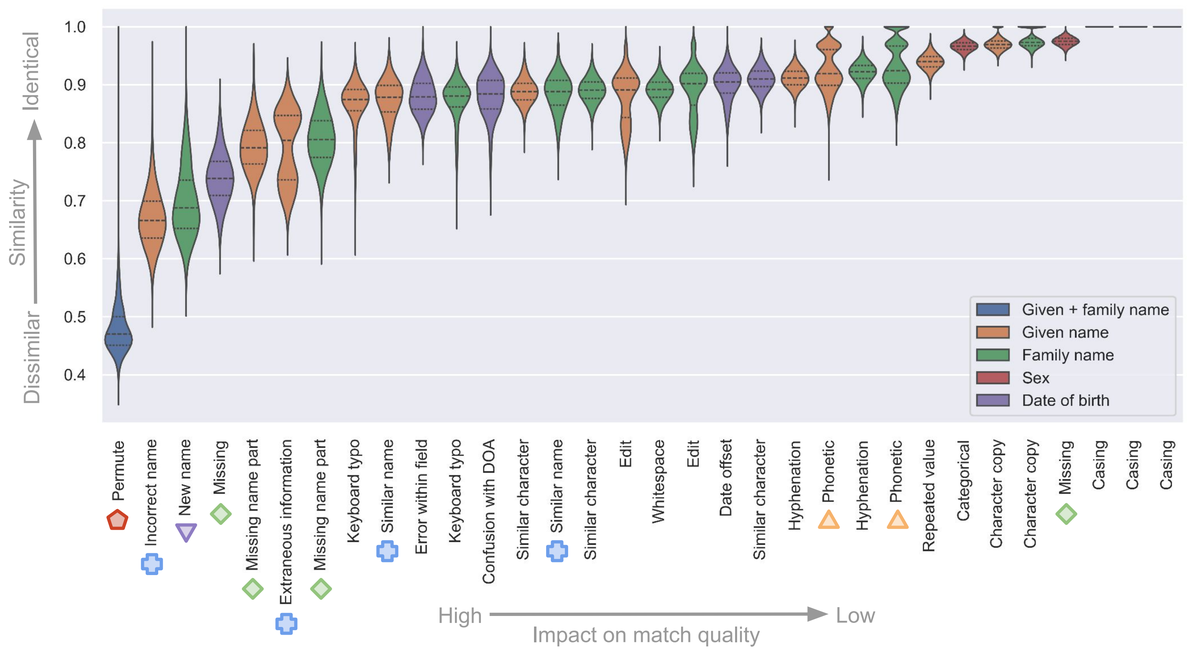

- Preprocessing and Masking happen at Data Holders, Matching and Classification at Trust Unit
- Masked records are non-reversible ⇒ Trust Unit has no insights into how similarities come to be
- **Solution:** Generate data similar to real-world data to simulate Preprocessing and Masking steps

## Methods

- Determine sources of typographic errors between personal record pairs in two LOBB sub-studies
- Derive error configuration for realistic data generation tool *Gecko*[6], then generate and mutate data
- Compare computed similarities between record pairs in LOBB and those generated by *Gecko*

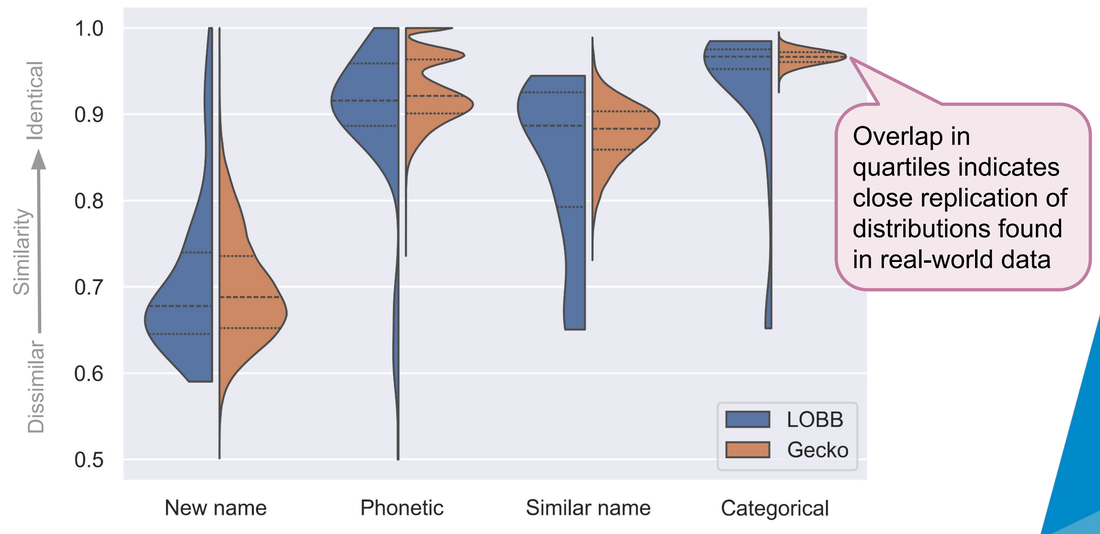
## Results

- 275 erroneous records in total ⇒ **1.94% error rate**
- 22 error classes with phonetic errors△ and new family names▽ taking up **over 48% of observed errors**
- Erroneous✚ or missing values◇ and permutation of given and last names⬠ rarely occur, but have a **drastic impact on match quality**
- *Gecko*-generated data closely replicates LOBB data in computed similarities of clean to mutated records




## Conclusion

- Reproducible workflow for generating realistic data in a medical study from multiple separate data sources
- Overview of error classes from an authentic source
- Proof of *Gecko*'s capabilities to generate and mutate realistic data found in the real world




LOBB Leipzig Obesity BioBank

MDS Medical Data Science

[1] Christen, Peter. The data matching process. Springer Berlin Heidelberg, 2012.
[2] Nguyen, Long, et al. "Privacy-preserving record linkage of deidentified records within a public health surveillance system: evaluation study." Journal of Medical Internet Research 22.6 (2020): e16757.
[3] Randall, Sean M., et al. "Privacy-preserving record linkage on large real world datasets." Journal of biomedical informatics 50 (2014): 205-212.
[4] Yigzaw, Kassaye Yitbarek, Antonis Michalas, and Johan Gustav Bellika. "Secure and scalable deduplication of horizontally partitioned health data for privacy-preserving distributed statistical computation." BMC medical informatics and decision making 17 (2017): 1-19.
[5] Christen, Peter, Thilina Ranbaduge, and Rainer Schnell. "Linking sensitive data." Methods and techniques for practical privacy-preserving information sharing. Cham: Springer (2020).
[6] Jugl, Maximilian, and Toralf Kirsten. "Gecko: A Python library for the generation and mutation of realistic personal identification data at scale." SoftwareX 27 (2024): 101846.