MEDICAL UNIVERSITY OF VIENNA

Wiener Biometrische Sektion
der Internationalen Biometrischen Gesellschaft
Region Österreich – Schweiz

Please join the Biometric Colloquium

MICHAEL KAMMER

Medical University of Vienna,

- Center for Medical Data Science, Institute of Clinical Biometrics,
- Department of Medicine III, Division of Nephrology and Dialysis

AN OVERVIEW OF R SOFTWARE TOOLS TO SUPPORT THE GENERATION OF SYNTHETIC DATA FROM REAL-WORLD DATA

January 15th, 2025 at 9:00 am

Seminarraum Center for Medical Data Science (previously CeMSIIS),
Spitalgasse 23, Room 88.03.513
Medical University of Vienna, 1090 Wien

HOST: Florian Frommlet

ABSTRACT

Biostatistical method development is partly driven through biomedical applications and the complexities of real-world datasets. Such datasets are also crucial for method comparison studies establishing the evidence base for biostatistical methods. On the applied side, open science principles call for the reproducibility and replicability of results facilitated by sharing data and code. However, sharing datasets is often difficult because of legal or ethical concerns. The creation of synthetic data closely reproducing real-world data is an alternative circumventing such issues, but the design of data generating mechanisms is not a trivial task. There also seems to be little consensus on how to standardize the actual coding of such data generators. Consequently, authors of publications often develop their own ad-hoc simulation code. Well-designed and easy to use software tools can help addressing these concerns.

As a step towards the standardization of coding practices and code sharing, we provide an overview of existing software packages in the programming language R to support the generation of synthetic data, with a focus on tabular data for the use in simulation studies. We found that there are many useful packages available, but only few of them were accompanied by peer-reviewed publications. We identified two general approaches to generate data: either

by explicitly specifying the data using distributional assumptions, or methods that create variations of an existing dataset e.g. similar to fully conditional specification.

In addition, we develop an R package for data generation intended to be easy to use, thus lowering the barrier to conducting proper comparison studies for newly developed methods. To complement the existing ecosystem a key goal of our work is to build a library of interesting data generating models derived from real-world datasets, which are then directly and easily available to other users. Such a library of presets serves as starting point for comparison studies and facilitates replicability and data protection, as well as the standardization of simulation setups by sharing configurations, rather than by sharing full datasets.

We demonstrate a selection of the identified packages including our own by simulations and example analyses using real-world datasets for which we derived plausible data generating models through the different approaches. Due to the complexities of real-world data and different potential research questions of interest a single tool is not enough to create data for all kinds of studies. Nevertheless, software packages may facilitate the standardization and exchange of code for data generation, thereby providing a framework that supports open science principles while accounting for privacy and data ownership concerns.