Davos, 09.01.24

Invitation to Lecture Michael Kammer, PhD, from the Medical University of Vienna

When:	Wednesday, 31 January 2023, 9:30h st
Where:	Herman-Burchard-Strasse 1, 7265 Davos Wolfgang – in Room Seehorn
Room:	Campus Schulungsraum Jakobshorn
Topic:	An overview of R software tools to support simulation studies: towards standardizing coding practices
Register:	Please register your participation with info@cardio-care.ch

An overview of R software tools to support simulation studies: towards standardizing coding practices

Michael Kammer[1,2], Lorena Hafermann[3], Georg Heinze[1]

1 Medical University of Vienna, Center for Medical Data Science, Institute of Clinical Biometrics, Vienna, Austria

2 Medical University of Vienna, Department of Medicine III, Division of Nephrology and Dialysis, Vienna, Austria

3 Charité–Universitätsmedizin Berlin, corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Institute of Biometry and Clinical Epidemiology, Berlin, Germany

Abstract

Biostatistical method development is partly driven through biomedical applications and the complexities of real world datasets. However, sharing these datasets is often difficult because of legal or ethical concerns. The creation of synthetic data closely reproducing the real world data is an alternative circumventing such issues. Similarly, generating realistic data is important for method comparison studies, which are crucial to establishing the evidence base for biostatistical methods.

However, the design of data generating mechanisms is not a trivial task, and there 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 simulation studies, with a focus on the coding of the data generating mechanism. We found that there are many powerful and general simulation packages available, but only few of them were accompanied by peer-reviewed publications. Most packages adopted approaches that explicitly specify the data using distributional assumptions, in contrast to methods that create variations of an existing dataset e.g. similar to fully conditional specification.

In addition, we developed 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 full 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 example analyses using real-world datasets for which we derived plausible data-generating models for simulations through the different approaches. Simulation studies are very diverse and therefore a single tool is not enough to perform all kinds of such studies. Nevertheless, software packages may facilitate the standardization and exchange of code, thereby providing a framework essential to design better simulation studies.