BMC Research Notes

COMMENTARY Open Access

Teaching reproducible research for medical students and postgraduate pharmaceutical scientists

Andreas D. Meid*

Abstract

In medicine and other academic settings, (doctoral) students often work in interdisciplinary teams together with researchers of pharmaceutical sciences, natural sciences in general, or biostatistics. They should be fundamentally taught good research practices, especially in terms of statistical analysis. This includes reproducibility as a central aspect. Acknowledging that even experienced researchers and supervisors might be unfamiliar with necessary aspects of a perfectly reproducible workflow, a lecture series on reproducible research (RR) was developed for young scientists in clinical pharmacology. The pilot series highlighted definitions of RR, reasons for RR, potential merits of RR, and ways to work accordingly. In trying to actually reproduce a published analysis, several practical obstacles arose. In this article, reproduction of a working example is commented to emphasize the manifold facets of RR, to provide possible explanations for difficulties and solutions, and to argue that harmonized curricula for (quantitative) clinical researchers should include RR principles. These experiences should raise awareness among educators and students, supervisors and young scientists. RR working habits are not only beneficial for ourselves or our students, but also for other researchers within an institution, for scientific partners, for the scientific community, and eventually for the public profiting from research findings.

Keywords: Reproducible research, Reproducibility, Heterogeneous treatment effects, Machine learning, Medical education

Introduction

In recent years, there has been a growing awareness that rigorous and transparent reporting of research is needed to ensure that study findings can be reproduced [1]. There is now consensus that the value of research can be enhanced by greater transparency and openness in the processes of research design, conduct, analysis, and reporting [1, 2]. At the same time, however, it became clear that simply registering study protocols, following reporting guidelines alone, or merely providing source data are not sufficient [3, 4]. For this reason, a lecture series on reproducible research (RR) for postgraduate students involved in several areas of clinical pharmacology at our institution has been established. When being offered to give this lecture series, it also appeared to me that even experienced researchers might be unfamiliar with certain aspects of a perfectly reproducible workflow, although reproducibility is of critical importance to any section of our multifaceted and dynamic discipline. This commentary honestly shares useful experiences from the pilot lecture series hoping to encourage reproducible working habits, to emphasize the crucial role of supervisors, and thus strengthen our awareness how important RR is when teaching good research practices. The course was divided into eight lectures, in which the topic was

*Correspondence: andreas.meid@med.uni-heidelberg.de
Department of Clinical Pharmacology and Pharmacoepidemiology, University of Heidelberg, Im Neuenheimer Feld 410, 69120 Heidelberg, Germany

introduced (1), technical requirements were identified (2), R/*RStudio* was presented as a suitable software (3), with which the participants' own projects (4) could be successively processed according to RR principles. Using pertinent methods for data management (5), data visualization (6), publishing and reporting (7), these projects could be finalized in a reproducible workflow (8). The commentary follows this structure and highlights important findings along the way.

Main text

Introducing reproducible research

In the first lecture of the course, RR was introduced by addressing the key questions about what it is by definition, why we should work accordingly, by which means we can conduct reproducible analyses, and how we can profit from them. The very first finding from the course was indeed that different researchers use the terms "reproducible" and "replicable" differently and sometimes interchangeably [5]. Following the "new lexicon for research reproducibility" [6], (methods) reproducibility is based on the same research conditions and is a fundamental requirement for successful replication (results reproduction). Full replication instead involves independent investigators, independent data, and optionally independent methods [7]. If original study results can be reproduced at all (which is rarely enough the case), then an application of the analysis procedure to new data (replication) is all the more meaningful. That such replication studies are very rare [8] shows the current difficulties of an open science philosophy. For example, DeBlanc and co-workers found limited availability of source code to analyse Medicare data in general medical journals [9]. While the authors meticulously explored potential reasons, they also emphasize that this clearly impedes RR. With data and source code at hand, however, all options are possible and allow all opportunities.

There are indeed positive examples for well-reproduced analyses and their added value in the current literature. In particular, in their re-analysis pursuing inferential reproducibility of the PACE ("paracetamol for acute low back pain") trial data, Schreijenberg et al. were able to confirm the results from earlier analyses of the same data [10]. This re-analysis demonstrated that trial conclusions were reproducible even after replication with a different methodological approach.

To keep attention high during the course, a cautionary negative example was also given with the "Duke-scandal" eventually leading to termination of clinical trials and lawsuits after article retraction [11]. With these drastic consequences in mind, the advantages of RR could be quickly worked out, either for the single researcher (e.g., streamlined working habits, strengthened confidence, easier adaptations), for the research team (e.g., higher research impact), or for the (scientific) community (e.g., better public and inter-professional recognition).

At this early stage of the lecture series, a working example was introduced providing a published analysis to be reproduced. This illustrative example was followed throughout the course. The particular publication of Duan et al. [12] headed into a current direction with high prospects for personalized medicine, namely predicting individual benefit by exploring heterogeneous treatment effects in the SPRINT [13] and ACCORD trials [14]. Based on the available baseline information, the authors developed models predicting individual treatment response to intensive antihypertensive treatment so that (better) treatment decisions could be made. Interestingly, a machine learning approach called *X*-learner [15] outperformed several alternative methods including the conventional logistic regression with interaction terms. In the publication, a publicly available Github repository was cited where the project is shared [12].

Requirements for RR

In the second lecture, workflow systems were approached to explicitly reflect the structure of projects, automate recurring steps, and transparently record the origin ('provenance') of (intermediate) results. Specifically, two possible solutions were assessed, namely the use of so-called visual workflows [16] and clearly defined folder structures. The latter applied to our working example [12], which was well-structured and was generally a very positive example bypassing several known barriers to proper reproducibility. Among the possible barriers are poor standardization of model building, lacking or insufficient documentation, incomplete transparency, or coding and typing errors. These common difficulties can be addressed by following five best practices in statistical computing [17], namely (1) best practices in code writing and commenting that also (2) documents workflow and key analytic decisions, (3) careful version control, (4) good data management, and (5) regular testing and review. All these aspects are easy to implement and understand from a standardized folder structure. Following our positive working example, Table 1 shows an exemplary pattern of how a research project could be structured. Centrally located in the main folder should be a readme-file in which important project information is documented. This includes, for example, the software packages with the corresponding version. Mainly, however, this concerns which files have to be executed in which order to get intermediate and final results. The main project folder should also include a file with which the entire project can be analysed. In this (make-) file separate analysis scripts are executed (i.e., "sourced"),

Table 1 Sample principle of a workflow for RR with a standardized folder structure

Main folder	Contents	Comment on item
Project name		Naming according to recurring pattern, which may include initials of the researcher and the date
└	Readme	Mandatory (text) file with important project information on prerequisites, scientific and technical background, and an instruction how to run the project code. Can also include a list of necessary software (package) versions, if not supplied as a separate file
└	Folder "data"	Folder with (raw) data or preprocessed data
└	Folder "lib"	Folder for storing literature or cross-project scripts (e.g., R functions, R packages, …)
└	Folder "results"	Folder for saving results of any kind (tables, figures, R-images, …)
└	Folder "src"	Folder with all executable ["source()"] files
└	Folder "paper"	Folder for storing publication drafts of any kind (e.g., Word documents, Markdown results, Shiny apps, …)
└	Folder "old"	Optional collection folder for old version of scripts or similar
└	Make	Central executable files for reproduction of the project

which run the necessary code lines for data preparation, statistical modelling or result generation (here collected in the subfolder "src"). (Raw) data are read in from a corresponding subfolder. If cross-project or recurring functions have to be used for the analysis, they can be stored in a library folder (here called subfolder "lib") (as well as other documents). Intermediate and final results can be stored automatically in a separate subfolder. For the publication and dissemination of the results, a separate subfolder is useful, which can be used for technical reports, presentations, manuscripts or even interactive apps. The Github repository of our working example was created accordingly [12].

Practical reproduction of a working example

From the third lecture on, the basics in using the software R were established. In addition to general techniques for data management and plotting, more specific skills were taught, as they may be necessary for RR. Thus, key concepts and their technical solutions were introduced (e.g., literate programming with R-markdown), and the advantages of common standards were emphasized. All course participants were able to bring in their own projects, which were worked on during the course. Inspired by the advantageous starting position of our working example, the published findings should now be reproduced, as well. Due to the fact that all lecture contents were now practically carried out on the working example in parallel, this working example accompanied the course until the end.

The Github repository of our working example provided the source code written in the Python programming language [12]. Before the code could be run, the data repository storing the source data required that ethical committee approval had to be obtained before data access could be granted [18]. After that, a workspace environment for the Python programming language had to be set up. This did not appear straightforward, but an intuitive readme file was helpful together with other 'best practice' ideas, standardized folders and files to be sourced for automatic data loading and running of the analysis code. The repository in particular provided several conceptual benefits when it comes to making distinctions between source data and derived files. It also helped to recognize the dependencies between code elements, different files, or libraries. In the end, the project could indeed be run and yielded results in the identical fashion to the published paper.

The published analysis apparently satisfied all three levels of reproducibility, full depth (i.e., code for data preparation, analytical results, and figures), portability to another computer system, and full coverage (i.e., all published results) [19]. Nevertheless, slightly diverging results and the fact that Python module versions have changed since the original publication were the reason to translate the Python code to more familiar R code as a more common programming language in medical research. Any colleague being not familiar with dedicated analysis software could have tried to reproduce the findings by this means. Interestingly, resulting estimates from this particular attempt to reproduce the original analyses were even more intriguing. Figure 1 illustrates several metrics from the original publication describing the performance of the *X*-learner in predicting individual treatment benefits to intensified blood pressure control.

All performance metrics relied on bootstrapped samples to derive estimates for their mean and 95% confidence intervals. For the most part, the Python and R results clearly overlapped suggesting good agreement and thus reproducibility (or "inferential" reproduction when considering that two different analytical methods were actually used). Nevertheless, there were also relatively large differences, especially in the absolute risk reduction in the group ("bucket") with predicted benefit

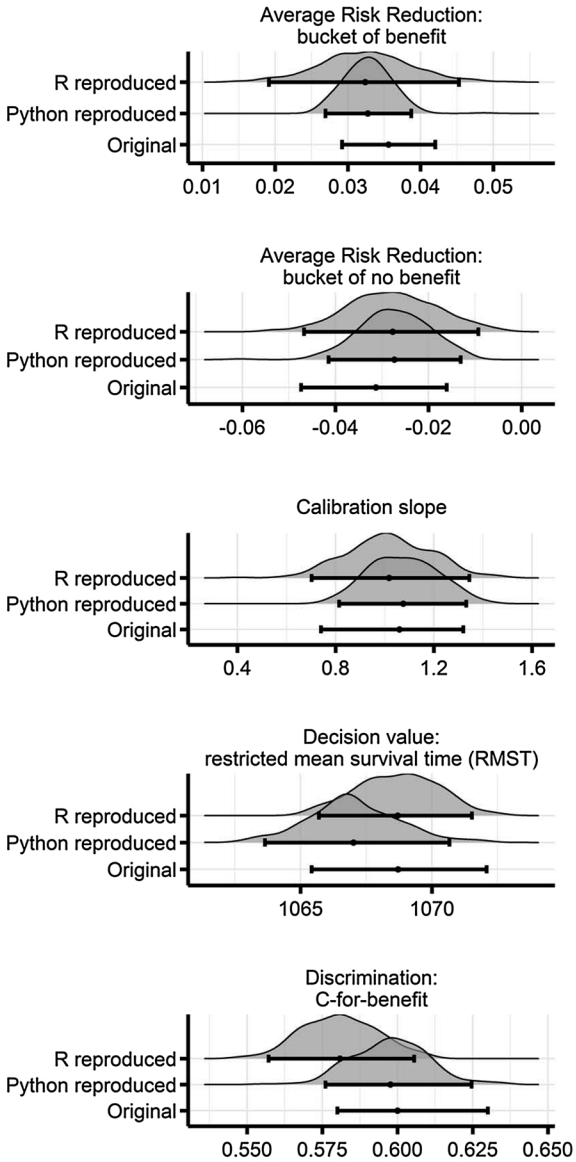

Fig. 1 Bootstrapped performance metrics used to derive mean estimates and 95% confidence intervals from the original publication [12], from the reproduction using the supplied code written in Python, and from the supplied code translated to R. Average risk reductions were calculated for two subgroups (buckets) of those patients with predicted benefit in absolute risk reduction (ARR >0) and those patients without predicted benefit (ARR ≤ 0). A calibration line was fitted between quintiles of ARRs and predicted risk, whose slope is chosen for this set of performance metrics. As a decision value, the model predicted restricted mean survival time [RMST (days)] indicates the mean time to event if treatment choice would have been based on the predicted individual benefit [and is thus to be compared with the baseline value of 1061.2 days, 95% confidence interval: (1057.4; 1064.1)]. The c-for-benefit is a metric reflecting the model's ability to predict treatment benefit (rather than risk for an outcome) [20]. Using the Python implementation to calculate this metric, the individual risk estimates reproduced in R yielded an estimate of 0.61 (0.55; 0.72). Of note, we restrict the presentation of results to distributions from resampling and their summary parameters; further numerical metrics to quantify reproducibility are left out for simplicity. Analyses were using the Anaconda distribution of Python version 3.7.3 (Anaconda Software Distribution, version 2–2.4.0) and the R software environment version 3.6.1 (R Foundation for Statistical Computing, Vienna, Austria)

of intensified blood pressure reduction or the c-for-benefit metric [20] with significant shifts in the distribution of bootstrap estimates. If concrete conclusions in terms of treatment recommendations are drawn from this, these differences could become relevant. A closer look yet revealed more aspects that are likely to explain the differences, especially between the two analysis software packages. One possible explanation is that Duan and coworkers used inverse probability of censorship weighting [21] to handle the time-to-event nature of the source data. For predicting the (survival) probability of no censorship, different external functions were loaded into Python and R that produced slightly different probabilities. But also custom generic code can be an issue upon translation. In order to calculate the c-for-benefit [20] in R, the R code provided in the supplements of the original publication was used and yielded a slightly smaller estimate. Among general explanations, the fact that random forests are indeed based on a random process is the most compelling argument, especially when they have to be reproduced with different programs on different operating systems (and thus different seeds) [22].

Lessons learned and implications

In the final lecture with project presentations, the lessons learned from the attempts to reproduce the working example were discussed. These experiences from the lecture series' working example expose several noteworthy aspects. Good documentation, standardized workflows, available data, and a freely available software solution facilitate RR. In this particular case, this framework likewise helped to teach and better understand cutting-edge methods. By deciding to reproduce this excellent work, a prototypical positive example was provided that simultaneously elucidated barriers that any of us could have trying to reproduce these findings. Considering the potential benefits of RR and our surprising observations and obstacles, the key question is how to incorporate these insights into typical workflows of medical research.

Obviously, data availability is a central aspect for reproducibility. Potential barriers might relate to data ownership, data privacy, or information enabling to identify patients. While considering the risk for misuse, protections should not preclude the incredible potential of

accessible data for research. There are several pertinent solutions ranging from public or private archives (enabling full data access) to public or private enclaves (allowing to obtain only aggregated results according to specific queries) [23].

Commonly available software tools and a common analysis language are further decisive points facilitating reproducible working habits. The endeavors of the pharmacometric community towards standards in Modeling and Simulation are interesting to follow [24]. Here, repositories sometimes include simulated data sets in accordance with actual distributions and correlations as a pragmatic solution to data privacy and security requirements. Pseudo-code can also be helpful, although a description alone will mostly not be sufficient for RR [4]. If pseudo-code is used, it should be as precise as possible, because the devil is in the details [25]. This makes the idea of a common analysis language across certain software packages seem all the more appealing, as it was approached in the field of Pharmacometrics [26]. As inter-operability of programming languages is more and more common, sharing our research projects in accordance with RR principle would be a helpful step forward to facilitated reproduction and hopefully successful replication with independent data in the end.

Outlook

Hands-on education and scientific interaction stand on top of all requirements. In addition to publications, tutorials, or courses, the principles of RR should also appear in harmonized curricula for quantitative medical researchers. While a clear roadmap for instructors has yet to be defined, a course in RR should consider the following insights from our pilot lecture series:

- As our pilot lecture series profited from different perspectives from inter-professional teams, either the audience or the lecturers should come from different disciplines to better recognize the aims, means, and potential merits of RR.
- As our hands-on experiences from a fully reproduced analysis illustrated good practices and practical obstacles of RR, such a course should be oriented towards practical examples.
- As scientific interactions were very enriching, a course of RR should include enough room for discussions.
- As skills in statistical programming are fundamental to apply best practices, such a course should provide the basics or be accompanied by another course.

If these efforts further supported reproducible projects, it would not only build confidence and credibility within our broad discipline, but also towards other disciplines and decision-makers. In order for our findings to impact regulatory decisions or patient care, results must be replicable in independent settings as the ultimate standard, for which reproducible projects are fundamental [7].

Abbreviations

ACCORD: Action to control cardiovascular risk in diabetes; PACE: "Paracetamol for acute low back pain" trial; RR: Reproducible research; SRPINT: Systolic blood pressure intervention trial.

Acknowledgements

The author would like to specially thank Tony Duan for his excellent support concerning the queries on the Python code. Similarly, the author thanks Lukas Arnecke for his on-site support on any issue related to Python. Last but not least, the author would like to express thanks to all participants of the lecture series and for all the stimulating thoughts.

Authors' contributions

The author is responsible for the content and writing of the article alone. The author read and approved the final manuscript.

Funding

The author is funded by the Physician-Scientist Programme of Heidelberg University, Faculty of Medicine. The funder has no role in any actions related to this analysis, or preparation of the paper.

Availability of data and materials

Access to the underlying data of the published analysis [12] being reproduced for teaching purposes can be requested at https://biolincc.nhlbi.nih.gov/home. As specified in published analysis [12], the corresponding analysis code is available from the repository https://github.com/tonyduan/hte-prediction-rcts. The mapped project folder with translated R code is available from the repository https://github.com/andreasmeid/Duan_reproduction.

Declarations

Ethics approval and consent to participate

This manuscript reports the experiences from a lecture series about reproducible research. For teaching purposes, a published analysis of particular trial data was reproduced, for which ethics committee approval (reference number V-223/2019 at the University of Heidelberg) and permission from the data keeping institution BioLINCC (https://biolincc.nhlbi.nih.gov/home/) were obtained.

Consent for publication

Not applicable.

Competing interests

The author declares no competing interests.

Received: 7 October 2021 Accepted: 26 November 2021
Published online: 09 December 2021

References

1. Serghiou S, Contopoulos-Ioannidis DG, Boyack KW, Riedel N, Wallach JD, Ioannidis JPA. Assessment of transparency indicators across the biomedical literature: How open is open? PLoS Biol. 2021;19:e3001107.
2. Catalá-López F, Caulley L, Ridao M, Hutton B, Husereau D, Drummond MF, Alonso-Arroyo A, Pardo-Fernández M, Bernal-Delgado E, Meneu R, Tabarés-Seisdedos R, Repullo JR, Moher D. Reproducible research practices, openness and transparency in health economic evaluations: study protocol for a cross-sectional comparative analysis. BMJ Open. 2020;10:e034463.

3. Wang SV, Verpillat P, Rassen JA, Patrick A, Garry EM, Bartels DB. Transparency and reproducibility of observational cohort studies using large healthcare databases. Clin Pharmacol Ther. 2016;99:325–32.
4. Goldstein ND, Hamra GB, Harper S. Are descriptions of methods alone sufficient for study reproducibility? An example from the cardiovascular literature. Epidemiology. 2020;31:184–8.
5. Plesser HE. Reproducibility vs. replicability: a brief history of a confused terminology. Front Neuroinform. 2017;11:76.
6. Goodman SN, Fanelli D, Ioannidis JP. What does research reproducibility mean? Sci Transl Med. 2016;8:341ps12.
7. Peng RD. Reproducible research in computational science. Science. 2011;334:1226–7.
8. Iqbal SA, Wallach JD, Khoury MJ, Schully SD, Ioannidis JP. Reproducible research practices and transparency across the biomedical literature. PLoS Biol. 2016;14:e1002333.
9. DeBlanc J, Kay B, Lehrich J, Kamdar N, Valley TS, Ayanian JZ, Nallamothu BK. Availability of statistical code from studies using medicare data in general medical journals. JAMA Intern Med. 2020;180:905–7.
10. Schreijenberg M, Chiarotto A, Mauff KAL, Lin CC, Maher CG, Koes BW. Inferential reproduction analysis demonstrated that "paracetamol for acute low back pain" trial conclusions were reproducible. J Clin Epidemiol. 2020;121:45–54.
11. Potti A, Dressman HK, Bild A, Riedel RF, Chan G, Sayer R, Cragun J, Cottrill H, Kelley MJ, Petersen R, Harpole D, Marks J, Berchuck A, Ginsburg GS, Febbo P, Lancaster J, Nevins JR. Genomic signatures to guide the use of chemotherapeutics. Nat Med. 2006;12:1294–300.
12. Duan T, Rajpurkar P, Laird D, Ng AY, Basu S. Clinical value of predicting individual treatment effects for intensive blood pressure therapy. Circ Cardiovasc Qual Outcomes. 2019;12:e005010.
13. Wright JT Jr, Williamson JD, Whelton PK, Snyder JK, Sink KM, Rocco MV, Reboussin DM, Rahman M, Oparil S, Lewis CE, Kimmel PL, Johnson KC, Goff DC Jr, Fine LJ, Cutler JA, Cushman WC, Cheung AK, Ambrosius WT. A randomized trial of intensive versus standard blood-pressure control. N Engl J Med. 2015;373:2103–16.
14. Cushman WC, Evans GW, Byington RP, Goff DC Jr, Grimm RH Jr, Cutler JA, Simons-Morton DG, Basile JN, Corson MA, Probstfield JL, Katz L, Peterson KA, Friedewald WT, Buse JB, Bigger JT, Gerstein HC, Ismail-Beigi F. Effects of intensive blood-pressure control in type 2 diabetes mellitus. N Engl J Med. 2010;362:1575–85.
15. Künzel SR, Sekhon JS, Bickel PJ, Yu B. Metalearners for estimating heterogeneous treatment effects using machine learning. Proc Natl Acad Sci USA. 2019;116:4156–65.
16. VisTrails. http://www.vistrails.org. Accessed 9 Nov 2021.
17. Sanchez R, Griffin BA, Pane J, McCaffrey DF. Best practices in statistical computing. Stat Med. 2021. https://doi.org/10.1002/sim.9169.
18. BioLINCC. https://biolincc.nhlbi.nih.gov/home. Accessed 9 Nov 2021.
19. Freire J, Bonnet P, Shasha D. Computational reproducibility: state-of-the-art, challenges, and database research opportunities. In: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. New York: SIGMOD; 2012. p. 593–6.
20. van Klaveren D, Steyerberg EW, Serruys PW, Kent DM. The proposed 'concordance-statistic for benefit' provided a useful metric when modeling heterogeneous treatment effects. J Clin Epidemiol. 2018;94:59–68.
21. Vock DM, Wolfson J, Bandyopadhyay S, Adomavicius G, Johnson PE, Vazquez-Benitez G, O'Connor PJ. Adapting machine learning techniques to censored time-to-event health record data: a general-purpose approach using inverse probability of censoring weighting. J Biomed Inform. 2016;61:119–31.
22. Beam AL, Manrai AK, Ghassemi M. Challenges to the reproducibility of machine learning models in health care. JAMA. 2020;323:305–6.
23. Simon GE, Coronado G, DeBar LL, Dember LM, Green BB, Huang SS, Jarvik JG, Mor V, Ramsberg J, Septimus EJ, Staman KL, Vazquez MA, Vollmer WM, Zatzick D, Hernandez AF, Platt R. Data sharing and embedded research. Ann Intern Med. 2017;167:668–70.
24. Lippert J, Burghaus R, Edginton A, Frechen S, Karlsson M, Kovar A, Lehr T, Milligan P, Nock V, Ramusovic S, Riggs M, Schaller S, Schlender J, Schmidt S, Sevestre M, Sjögren E, Solodenko J, Staab A, Teutonico D. Open systems pharmacology community-an open access, open source, open science approach to modeling and simulation in pharmaceutical sciences. CPT Pharmacometrics Syst Pharmacol. 2019;8:878–82.
25. Petrone AB, DuCott A, Gagne JJ, Toh S, Maro JC. The Devil's in the details: reports on reproducibility in pharmacoepidemiologic studies. Pharmacoepidemiol Drug Saf. 2019;28:671–9.
26. Smith MK, Moodie SL, Bizzotto R, Blaudez E, Borella E, Carrara L, Chan P, Chenel M, Comets E, Gieschke R, Harling K, Harnisch L, Hartung N, Hooker AC, Karlsson MO, Kaye R, Kloft C, Kokash N, Lavielle M, Lestini G, Magni P, Mari A, Mentré F, Muselle C, Nordgren R, Nyberg HB, Parra-Guillén ZP, Pasotti L, Rode-Kristensen N, Sardu ML, Smith GR, Swat MJ, Terranova N, Yngman G, Yvon F, Holford N. Model description language (MDL): a standard for modeling and simulation. CPT Pharmacometrics Syst Pharmacol. 2017;6:647–50.

Publisher's Note