The Journal of Open Source Education

r-cubed: Guiding the overwhelmed scientist from random wrangling to Reproducible Research in R

Luke W. Johnston[1], **Helene Baek Juel**[2], **Bettina Lengger**[3], **Daniel R. Witte**[1,4], **Hannah Chatwin**[5], **Malene Revsbech Christiansen**[2], **and Anders Aasted Isaksen**[4]

1 Steno Diabetes Center Aarhus, Aarhus, Denmark **2** Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Copenhagen, Denmark **3** Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark **4** Department of Public Health, Aarhus University, Aarhus, Denmark **5** University of Southern Denmark, Odense, Denmark

DOI: 10.21105/jose.00122

Software

- Review
- Repository
- Archive

Submitted: 08 February 2021
Published: 05 October 2021

Summary

The amount of biological data created increases every year, driven largely by technologies such as high-throughput -omics, real-time monitoring, or high resolution imaging in addition to greater access to routine administrative data and larger study populations. This not only presents operational challenges, but also highlights considerable needs for the skills and knowledge to manage, process, and analyze this data (Brownson et al., 2015). Along with the open science movement on the rise, methods and analytic processes are also increasingly expected to be open and transparent and for scientific studies to be reproducible (Watson, 2015).

Unfortunately, training in modern computational skills has not kept pace, which is particularly evident in biomedical research (Attwood et al., 2017; Cooper, 2017), where training tends to focus on clinical, experimental, or wet-lab skills. The computational learning module we have developed and described below aims to introduce and improve skills in R, reproducibility, and open science for researchers in the biomedical field, with a focus on diabetes research.

The `r-cubed` (**R**eproducible **R**esearch in **R** or R3) learning module is structured as a three-day workshop, with five sub-modules. We have specifically designed the module as an open educational resource that: 1) instructors can make use of directly or modify for their own lessons; and, 2) learners can use independently or as a reference after participating in the workshop. All content is available for re-use under CC-BY License.

Statement of Need

Reproducibility is a key component to verifying scientific findings. Unfortunately, the reproducibility of scientific studies is difficult to estimate (Considine et al., 2017; Leek & Jager, 2017) as researchers may be unaware of, have no training in, or lack incentives to conduct reproducible research. Improving reproducibility requires an awareness and training in multiple areas, including version control, project and data management, and reproducible reporting alongside the traditional curriculum of statistical analysis. Training in these skill areas is critical to tackling the modern demands of reproducibility.

JOSE
The Journal of Open Source Education

The learning module provides training in reproducibility, open science, and collaboration by teaching introductions to Git, GitHub, R Markdown, and data wrangling and visualization in R. While other resources for learning R and Git exist (Bryan, 2019; Chester Ismay, 2019; Lee, 2019), this module places a greater emphasis on reproducibility and the general workflow for doing data analysis. It also includes instructions targeted to other instructors to assist with re-use of the material. This module is based on the experiences and needs of biomedical researchers, as it is an area that currently lacks sufficient training for these skills. The authors, all of whom work or have worked in biomedical research, used their experience in creating and shaping the content.

The learning module was specifically designed with *re-use and adoption* in mind. Both the *Welcome* and *For Instructors* sections give more details on how this module can be re-used.

The *target audience* for this learning resource is described in *Welcome* page. Briefly, we suggest: *learners* use the material during and after the workshop; *instructors of the workshop* use the module as a reference while teaching; and those *interested in teaching*, but who are still new to the knowledge or skills themselves, can use the content to build their own workshop. The target learner of the workshop is detailed in the *syllabus*.

Description of Learning Modules

The *learning objectives* of this module are to provide a broad introduction to reproducible research practices, in the context of RStudio, Git, and GitHub. A detailed description of the learning objectives is found in the *syllabus*. Briefly, upon completing the learning module, learners are expected to have:

1. A basic level of proficiency in using R, a statistical programming language.
2. Improved data and code literacy.
3. The ability to conduct a modern and reproducible data analysis project.
4. Insight into the main challenges impeding open and reproducible research.

The learning module encompasses *five sub-modules* (Table 1), with each sub-module representing about a half-day of lessons. A general schedule is given in the *Schedule* section. Details on how these lessons can be used or structured are found in *For Instructors* section.

Table 1: An overview of the five sub-modules that form the `r-cubed` learning module.

Sub-module	Description
Management of R Projects	Introducing RStudio and R Projects; using packages, data, and file paths; and learning how to troubleshoot.
Version Control with Git	Using Git with RStudio; synchronizing Git with GitHub; dealing with file conflicts; and using Git as a collaboration tool.
Data Management and Wrangling	Introducing good working practices; loading data and packages into RStudio; and transforming data using select, rename, filter, arrange, and split-apply-combine functions.

Sub-module	Description
Analytically Reproducible Documents	Introducing and using R Markdown to insert code, tables, and figures to make a reproducible document.
Data Visualization	Plotting various combinations of variables and using formatting effectively.

Sub-modules are designed to be completed as a series in the order given, since concepts taught in later sub-modules are dependent on earlier lessons. Principles and applications of reproducibility are highlighted throughout sub-modules. Sub-modules are accompanied by three stand-alone lectures (and one introductory lecture) aimed at drawing together challenges in reproducibility and the RStudio skills taught during the workshop (see *Lecture Slides* section).

Instructional Design

The module is designed for in-person settings, where the instructional design uses a combination of teaching methods (Table 2). Content taught using the methods below are described in the *Welcome* section. We also incorporate the use of sticky notes, as pioneered by the Carpentries, to help with troubleshooting and to facilitate a positive learning experience overall.

Table 2: Description of pedagogical methods used throughout the workshop.

Method	Description	Advantages
Participatory live-coding lessons	Participants join with instructors to write and troubleshoot code step-by-step.	Encourages participants to actively engage with the material, to build muscle memory through typing, and to learn how to handle mistakes, rather than passively observing content.
Independent reading of specific sections	Participants read more concept-heavy content at their own pace.	Including reading activities provides diverse learning opportunities aside from listening skills, can be advantageous for non-native English speakers, and can slow the pace of learning to enhance retention.
Completion of brief exercises	Exercises are interspersed throughout the live-coding sessions to complement the content.	Hands-on, practical exercises help reinforce what was previously learned and provides an opportunity to work through code at the learners' own pace.

Method	Description	Advantages
Group assignment work	Participants collaborate with others to apply skills taught over the workshop in terms of cleaning, manipulating, analyzing, and visualizing data.	Helps to reinforce learning by applying the skills and knowledge to a new problem and task, thus building confidence in using the skills.

The *For Instructors* section further describes the teaching approach that instructors can adopt to improve learning outcomes.

Experience of use in teaching and learning situations

The learning module has been delivered several times as 3-4 day workshops. Instructors were graduate students or postdoctoral researchers from diverse fields within diabetes research, of whom most had relatively recently learned R themselves. Including late-novice or early-intermediate R users as instructors makes communicating and relating to the learners easier and we believe it provides a better learning environment. A ratio of about 4-6 learners for every instructor/helper has proven to be most effective in past workshops.

Workshop participants were mostly PhD and postdoctoral researchers in the field of diabetes, most of whom were beginners in terms of their exposure to R and understanding of reproducibility. Participants were grouped into groups of 4-5 people as part of their group assignment.

After each day, participants gave feedback on the structure and content of the workshop through a Google Forms survey. This anonymous feedback is saved in the repository and is used to improve and enhance `r-cubed`, and ensure its continual improvement, relevance, and beginner-friendliness to biomedical researchers.

Story of the Project

Luke Johnston taught and built the learning material from this workshop in smaller segments over many years. The material was initially created because of the near complete lack of relevant training in academic settings on data management, coding and general workflows in data analysis, version control, and project management of scientific research. The teaching material was compiled into a more structured form when the Danish Diabetes Academy hosted the workshop, where the current author team has since heavily updated and revised it. We greatly enjoy teaching the workshop and, based on the feedback from the surveys, the participants do too!

Acknowledgements

We want to specifically thank the Danish Diabetes Academy for hosting, organizing, and providing a space to build and grow this workshop. We also thank the instructors and helpers of the first version of the workshop (João Santiago, Anna Schritz, Omar Silverman) for their early feedback during the first iteration.

References

Attwood, T. K., Blackford, S., Brazas, M. D., Davies, A., & Schneider, M. V. (2017). A global perspective on evolving bioinformatics and data science training needs. *Briefings in Bioinformatics*, *20*(2), 398–404. https://doi.org/10.1093/bib/bbx100

Brownson, R. C., Samet, J. M., Chavez, G. F., Davies, M. M., Galea, S., Hiatt, R. A., Hornung, C. A., Khoury, M. J., Koo, D., Mays, V. M., Remington, P., & Yarber, L. (2015). Charting a future for epidemiologic training. *Annals of Epidemiology*, *25*(6), 458–465. https://doi.org/10.1016/j.annepidem.2015.03.002

Bryan, J. (2019). *STAT 545: Data wrangling, exploration, and analysis with R.* https://stat545.com/

Chester Ismay, A. Y. K. (2019). *Statistical inference via data science: A ModernDive into R and the tidyverse.* Taylor & Francis Ltd. ISBN: 0367409828

Considine, E. C., Thomas, G., Boulesteix, A. L., Khashan, A. S., & Kenny, L. C. (2017). Critical review of reporting of the data analysis step in metabolomics. *Metabolomics*, *14*(1). https://doi.org/10.1007/s11306-017-1299-3

Cooper, et al., D. (2017). *Supporting the changing research practices of public health scholars.* https://doi.org/10.18665/sr.305867

Lee, M. (2019). Happy belly bioinformatics: An open-source resource dedicated to helping biologists utilize bioinformatics. *Journal of Open Source Education*, *2*(19), 53. https://doi.org/10.21105/jose.00053

Leek, J. T., & Jager, L. R. (2017). Is most published research really false? *Annual Review of Statistics and Its Application*, *4*(1), 109–122. https://doi.org/10.1146/annurev-statistics-060116-054104

Watson, M. (2015). When will 'open science' become simply 'science?' *Genome Biology*, *16*(1). https://doi.org/10.1186/s13059-015-0669-2