Reproducible workflows and collaboration with domain-knowledge experts

Charles T. Gray[1], Hollie Birkinshaw[2], Matthew Grainger[3], Tamar Pincus[2], and Gavin Stewart[1]

[1]Newcastle University
[2]Royal Holloway
[3]NINA

July 28, 2021

```
library(tidyverse)
library(targets)
library(gt)
knitr::opts_chunk$set(echo = TRUE)

knitr::opts_knit$set(root.dir = here::here())
```

```
source("R/hpp_themes.R")
```

1 Unboxing the black box of data wrangling

Proponents of open science commendably underscore reuse and extensibility of scientific research components, such as data and code; techniques that facilitate the incorporation of these components into future analyses [7, 10, 13]. Less explicit attention, however, is given to the benefits of reproducible workflows, where results can be readily calculated by another researcher, and computational transparency *during* the research project, as opposed to beyond publication. Of course, we can appeal to an analyst's commitment to scientific civic duty, but by explicitly examining how reproducible workflows can facilitate collaboration with domain-knowledge experts we begin to answer the arguably more pertinent question, *What's in reproducibility for me*?

1.1 Benefits of reproducibility and open science

Much has been said about the benefits of reproducibility, wherein precisely the same computational results are readily produced by another analyst, and open science. We differentiate between explicit replication of results, where the same scientific conclusion is drawn from a repeated experiment, as was the focus of the semantically confusing Reproducibility Project [4]. However, it is worth emphasising that reproducibility in and of itself does not necessarily further the 'process of scientific discovery' [5]. And in the context of open science, preregistratation, where a clear outline of a scientific methodology is published prior to embarking on the work, does not protect against poor conclusions [12]. These discussions are central and ongoing in the meta-research community; a weighing of merits of pathways to scientific discovery, how to find truths or draw conclusions.

In this manuscript, we do not seek to contribute to these worthy discussions. Instead, our focus is on scientific computational *practice*, as opposed to conclusion. Setting aside which path leads us best to a result, let us instead focus on an aspect of computational reproducibility that provides utility for the analyst in avoiding pitfalls and foibles of answering scientific questions. For science has necessarily become ever more collaborative, as the demand for more complex algorithms rises.

In the example discussed in this manuscript, we examine the implementation of a network meta-analysis in a Cochrane review of the treatment of chronic pain with antidepressants [2]. A project such as this is necessarily collaborative; it is unrealistic to expect a single academic to provide expert-level data science, statistical analysis, and domain-knowledge insights. Scientific fraud is, of course, something we seek to guard against, but here we are more concerned with garden-variety errors the data-analyst and statistician, who by nature are discipline agnostic, make in absence of a genuine understanding of the domain. In so doing, we hope to contribute to the growing literature on selfish reasons for an analyst to adopt computational reprudicbility [9].

1.2 The problem of black box analysis: pitfalls, foibles, and outright fraud

Reproducibility can certainly guard against fraud. By reproducing analyses, scientists have uncovered cases of outright manipulation of data and results [1, 11]. Reproducibility provides a transparency that assists in uncovering academic misconduct of this kind. However, there is also a grey area of **Questionable Research Practices**, such as choosing variables to maximise statistical significance, wherein a scientist diligently follows methods they were trained in but lead to spurious results [6]. Here we will focus not on questionable practices in statistical modelling, but in the preparation of data for modelling, variously known as **cleaning** or **wrangling**.

Much focus is given to correct model design. However, there is a deceptively mundane requirement before modelling can begin. Commonly, variables of in-

terest must be encoded with finite levels in columns in a flat-form delimited data format, such as `.csv`. For this Cochrane review, we wish to build network meta-analyses, but in addition, we wish to create summary of findings tables, as well as provide the other scientists with an accessibly formatted delimited dataset that they can readily open in whatever spreadsheet software they are comfortable with.

```
tar_read(p_output_raw) %>% gt()
```

[table]labelformat=empty,skip=1pt lllllllll outcome study arm obs obs_info type design condition class main_aim
mood, pain, etc. unique study identifier unique arm identifier columns with mean, sd, counts, sample size, etc. information extracted from column headers: timepoint, scale, etc. Intervention Type Group Chronic pain conditions(s) Intervention Class Main aim (Pain...

1.3 A reproducible workflow for collaboration

From here we will detail an R-specific workflow using the `targets::` package [8], Google sheets, with data-scraping provided by the `googlesheets4::` package [3].

2 Labelling scales and interpreting dosage

3 Discussion

References

[1] Keith A. Baggerly and Kevin R. Coombes. Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. 3(4):1309–1334.

[2] Hollie Birkinshaw, Claire Friedrich, Peter Cole, Christopher Eccleston, Marc Serfaty, Gavin Stewart, Simon White, R. Andrew Moore, and Tamar Pincus. Antidepressants for pain management in adults with chronic pain: A network meta-analysis. (4).

[3] Jennifer Bryan. *Googlesheets4: Access Google Sheets Using the Sheets API V4*.

[4] Open Science Collaboration. Estimating the reproducibility of psychological science. 349(6251).

[5] Berna Devezer, Luis G. Nardin, Bert Baumgaertner, and Erkan Ozge Buzbas. Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. 14(5):e0216125.

[6] Hannah Fraser, Tim Parker, Shinichi Nakagawa, Ashley Barnett, and Fiona Fidler. Questionable research practices in ecology and evolution. 13(7):e0200303.

[7] Heidi Laine. Open science and codes of conduct on research integrity. 37(4).

[8] Will Landau. *The Targets R Package User Manual*.

[9] Florian Markowetz. Five selfish reasons to work reproducibly. 16(1):274.

[10] R. D. Peng. Reproducible Research in Computational Science. 334(6060):1226–1227.

[11] Amy Schleunes. Top Spider Biologist's Research Under Fire.

[12] Aba Szollosi, David Kellen, Danielle Navarro, Richard Shiffrin, Iris van Rooij, Trisha Van Zandt, and Chris Donkin. Is preregistration worthwhile?

[13] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The FAIR Guiding Principles for scientific data management and stewardship. 3(1):160018.