comment

Corrected: Publisher Correction

A visual tool for defining reproducibility and replicability

Reproducibility and replicability are fundamental requirements of scientific studies. Disagreements over universal definitions for these terms have affected the interpretation of large-scale replication attempts. We provide a visual tool for representing definitions and use it to re-examine these attempts.

Prasad Patil, Roger D. Peng and Jeffrey T. Leek

Reproducibility crisis?

Reproducibility and replicability are at the centre of heated debates across scientific disciplines[1]. One of the central issues is that almost no one agrees upon the specific meaning of these terms[2]. A major initiative in psychology used the term 'reproducibility' to refer to completely redoing experiments, including data collection[3]. In cancer biology, reproducibility has been used to refer to the recalculation of results using a fixed set of data and code[4]. These disagreements in terminology may seem purely semantic, but they have major scientific and political implications.

A prominent back-and-forth in the pages of the journal *Science* regarding the psychology replication attempt by the Open Science Collaboration (OSF) mentioned above[3] hinged on the definitions of 'reproduction' and 'replication'[5]. Partially due to differing definitions for these terms, the sides came to opposing conclusions based on the results of the original study, with Gilbert et al. expressing optimism about scientific replicability[5]. Nonetheless, the press, government officials and even late-night comedy hosts have pointed out 'irreproducibility' as the fundamental problem with the scientific process. But they use this term to encompass the more insidious problems of false discoveries, missed discoveries, scientific errors and scientific misconduct. Others have suggested conceptual, verbal frameworks to help define these terms[2], but when the terms are actually used in conducting a study of replicability within a scientific discipline, it can remain unclear which definition is being applied and to what extent it is being followed during a replication attempt of an originally published study.

Visualizing definitions

Because of the many field-specific definitions and resulting semantic disagreements, we have developed a mode of visual representation that can be used to establish definitions for reproducibility, replicability and related terms in the context of a scientific study itself. We use icons to represent basic components of a scientific study that tend to vary across studies and disciplines: the intent of a study (research question, experimental design, analysis plan) and what was actually performed in the conduct of the study (data collected, analysis conducted, estimates made and conclusions asserted). Applying our visual representation tool can be used both to display the definitions of reproducibility and replicability that we adhere to and to illustrate a set of recent, related scientific publications in reference to these definitions (Fig. 1). The R package scifigure is freely available for creating these visualizations (https://CRAN.R-project.org/package=scifigure). We encourage those who intend to discuss issues of replicability in science to first display how they are defining their terminology by using these visualizations.

We anticipate three primary usages of the scifigure R package: to provide a precise point of reference for discussions of reproducibility and replicability in the literature (especially across fields), to compare the conduct of a replication or reproduction study to a pre-established definition of the term, and to compare differences in protocol across multiple studies. Figure 1 shows a comparison of published reproduction and replication efforts to our pre-specified definitions of reproducibility and replicability. Although we compare pairs of studies here, the number of comparisons per figure is only constrained by the size of the R graphical device (up to 20 studies can be visualized in one figure). There are many substeps and specifics in scientific protocols that could have been visualized by icons, especially in the details of experimental design and differences in methodology. We have made it possible for users to input custom icons to represent the steps in their experiment however they wish.

The advantage of representing a reproduction and replication attempt visually in this format is that both the study authors and study readers can confirm (i) to what degree these definitions were followed in the course of the study and (ii) whether or not any differences from the definitions were accounted for when making comparisons to the original publication in question. Although we will use the definitions for reproducibility and replicability displayed in Fig. 1a to address ongoing discussions, we provide the scifigure R package so that all definitions may be represented and discussed in a standardized manner. For example, in our definition of replicability, we rely on some form of formal statistical assessment to compare original and replication estimates and resulting claims, as noted by the approximation sign. Some context-specific approaches have been proposed[6] to make these comparisons, but these may not always be applicable or necessary. The main goal of the scifigure package is to allow for visual comparisons of study protocols to an established baseline or original definition, and these need not necessarily be the ones we use in Fig. 1a.

Reviewing replication attempts

We can use our visual representation to resolve arguments and misconceptions around some controversial discussions of reproducibility and replicability. Consider the case of the claim that only 6 of 53 preclinical studies were replicated by scientific teams at a pharmaceutical company[7]. Compared to the definition we establish in Fig. 1a, the paper from Begley and Ellis describing this replication effort reported a hypothesis: that most studies do not replicate. It also reported a claim: that 47 out of the 53 studies could not be replicated by scientists at the company. However, the population, hypothesis, experimental design, experimenter, data, analysis plan,

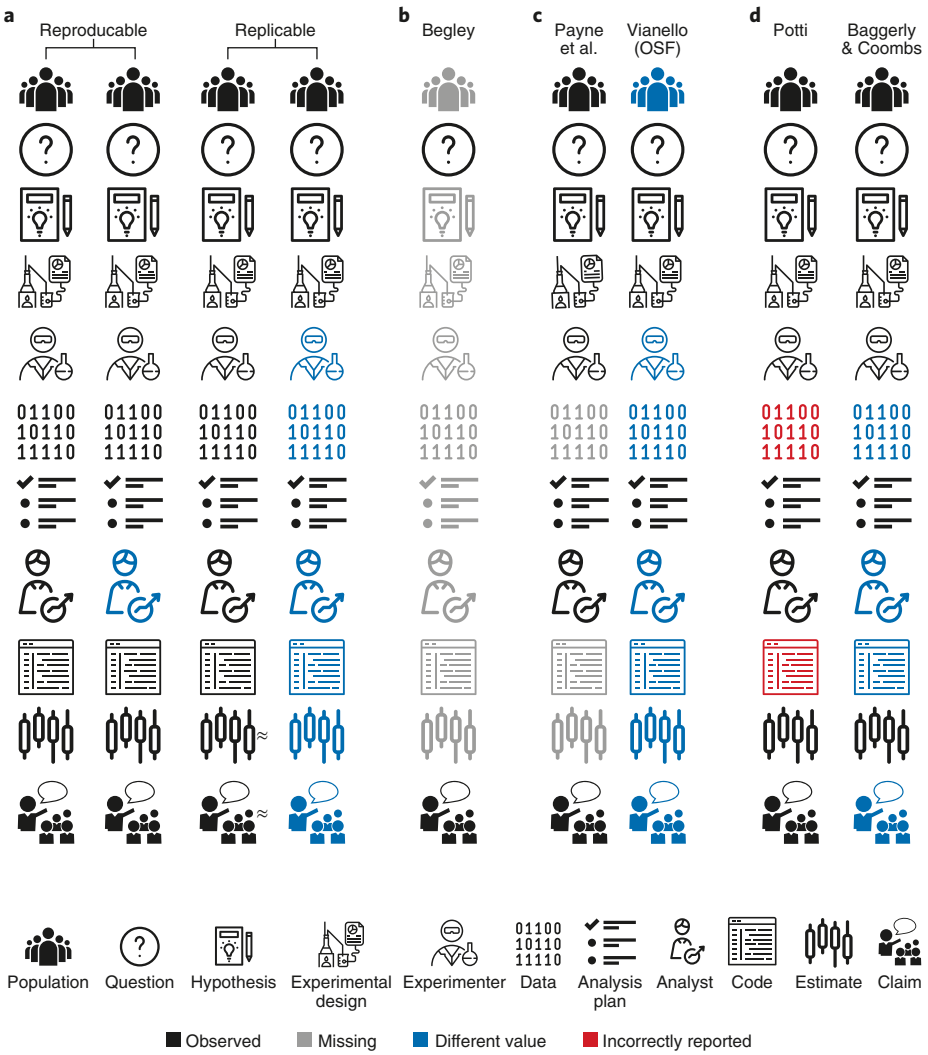

Fig. 1 | A graphic representation for the scientific process a, We define reproducibility as re-performing the same analysis with the same code using a different analyst, and we define replicability as re-performing the experiment and collecting new data. **b**, The paper in which only 6 of 53 preclinical discoveries in oncology and cancer haematology replicated[7] only reported a question and a claim; the rest of the scientific components of the study (hypothesis, experimental design, analysis plan, etc.) were missing. **c**, Scientists disagree over the interpretation of the results from the RP:P[3,5], in part because a replication was not performed as the population changed. **d**, In the case of the publication of fraudulent gene signatures for chemosensitivity[10], reproducibility wasn't the main issue; the issue was that the original study contained incorrect code and data. Icons made by Freepik, OCHA, Phatplus, Linh Pham and Gregor Cresnar are found on https://www.flaticon.com/; icon from PNG Repo licensed by CC BY 4.0.
Population: the complete collection of units for which information is sought[11].
Question: the interrogative statement we wish to address in the population of interest.
Hypothesis: the proposed explanation of our question that we wish to test.
Experimental design: our stated procedure for sampling and measuring units from our population of interest.
Experimenter: the scientist who will carry out the experimental design.
Data: the manifestation of the experimenter carrying out the experimental design to his or her best ability.
Analysis plan: suggested by our hypothesis, the manner in which we intend to extract information from our data to answer our research question.
Analyst: the scientist who will carry out the analysis plan.
Code: the manifestation of the analyst carrying out the analysis plan to his or her best ability; this includes any decisions made in the course of statistical analysis.
Estimate: the statistical result(s) obtained from the code.
Claim: the conclusion about the research question implied by the estimate.
Observed: this step in the replication attempt is identical to its counterpart in the original study.
Missing: this step in the replication attempt is unreported or unknown and cannot be compared to its counterpart in the original study.
Different value: this step in the replication attempt differs from its counterpart in the original study.
Incorrectly reported: this step in the replication attempt is not accurately represented or described.

analysts, code and estimates are not available in the paper describing the replication attempt (Fig. 1b). This makes it clear that the published report of the replication effort itself is missing most of the components that should be reported in all scientific studies.

Later, the same pharmaceutical company reported replication study results[8], though it was unclear whether these newly released replications were part of the originally reported 53. It was pointed out that some of the reported studies included experiments with different populations, which violates our definition of replication. Conducting identical studies that target different populations could yield effects of different magnitude and direction, due to the composition of each population. This may be misinterpreted as non-replication of an effect if the change in population is not documented. A similar issue was at the heart of a disagreement over several of the studies in the Reproducibility Project: Psychology (RP:P)[3]. In this project, independent investigators replicated 100 studies. In one case, a study originally performed in the United States on US college students was evaluated among a group of Italians (Fig. 1c). Although this could raise concerns because this violates our definition of replicability, it may not violate the definition used by the RP:P, who studied the impact of this discrepancy in a subsequent replication effort by having multiple labs conduct replications of the same original study simultaneously[9]. These details aside, the important point here is that the RP:P never explicitly defined replicability, which caused disagreement over the interpretation of their results.

Finally, consider one of the earliest and most egregious debates over reproducibility, namely the case of a predictor of chemosensitivity that ultimately fell apart. This lead to lawsuits, an Institute of Medicine conference and report, and ultimately the end of the lead author's scientific career[10]. In this case, both the code and the data produced by the original authors were made available; however, they were the wrong code and data. A team from MD Anderson was able to investigate and ultimately produce data and code that reproduced the original results (Fig. 1d). Ultimately, the study was reproducible, which is surprising given the focus on this study being a violation of reproducibility. The problem with the study was not that the data and code could not be produced; it was that these items, when produced, were wrong.

By explicitly visualizing which components of the scientific process differed and which were held constant from the original study to the repeat attempt(s), we can help resolve arguments and provide a solid foundation for journal and public policy around these complicated issues. We do not claim that science does not suffer from a replicability problem, but we hope that a tool to establish consensus before further discussion or measurement will help elucidate the extent to which this problem exists and will allow us to better evaluate whether new policies help address it.

Data and code availability

The scifigure package and its accompanying vignette can be downloaded from https://CRAN.R-project.org/package=scifigure ❐

Prasad Patil[1,2], Roger D. Peng[3] and Jeffrey T. Leek[3]*

[1]*Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA.* [2]*Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA, USA.* [3]*Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA.*
**e-mail: jtleek@gmail.com*

Published online: 17 June 2019
https://doi.org/10.1038/s41562-019-0629-z

References

1. Fanelli, D. *Proc. Natl. Acad. Sci. USA* **115**, 2628–2631 (2018).
2. Kenett, R. S. & Shmueli, G. *Nat. Methods* **12**, 699 (2015).
3. Open Science Collaboration. *Science* **349**, aac4716 (2015).
4. Peng, R. D. *Science* **334**, 1226–1227 (2011).
5. Gilbert, D. T., King, G., Pettigrew, S. & Wilson, T. D. *Science* **351**, 1037 (2016).
6. Heller, R., Bogomolov, M. & Benjamini, Y. *Proc. Natl. Acad. Sci. USA* **111**, 16262–16267 (2014).
7. Begley, C. G. & Ellis, L. M. *Nature* **483**, 531–533 (2012).
8. Cramer, P. E. et al. *Science* **335**, 1503–1506 (2012).
9. Klein, R. A. et al. *Adv. Methods Pract. Psychol. Sci.* **1**, 443–490 (2018).
10. Baggerly, K. A. & Coombes, K. R. *Ann. Appl. Stat.* **3**, 1309–1334 (2009).
11. Johnson, R. A. *Statistics: Principles and Methods* (Wiley, 2009).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-019-0629-z.