Comment

ILLUSTRATION BY DAVID PARKINS

Reproducibility: expect less of the scientific paper

Olavo B. Amaral & Kleber Neves

Make science more reliable by placing the burden of replicability on the community, not on individual laboratories.

In 2018, we embarked on a journey to assess the reproducibility of biomedical research papers from Brazil. Thus began a multicentre collaboration of more than 60 laboratories to replicate 60 experiments from 2 decades of Brazilian publications[1]. We randomly selected experiments that used three common laboratory techniques: the MTT assay for cell viability, RT-PCR to measure specific messenger RNAs and the elevated plus maze to assess anxiety in rodents.

Each experiment will be repeated in three labs, and each lab has developed replication protocols based on the original article's written methods. The process of building, reviewing and preregistering these protocols has taken months of communication between the coordinating team and the labs performing replications. We had intense arguments around the meaning of positive and negative controls and the merits of different metrics to define replication success. We also spent many hours on mundane tasks, such as studying the nutritional content of different brands of bologna sausage to better emulate a cafeteria diet fed to rats in one experiment.

These are just some of the obstacles we

have faced so far as coordinators of the Brazilian Reproducibility Initiative: there was also the massive shutdown of labs due to the COVID-19 pandemic and the sharp decline of Brazil's currency, the real. With all of this, experiments are starting slowly, and the project is now set to finish by the end of 2022.

That said, we have already reached conclusions that apply beyond Brazilian science. As a broad solution, more rigorous protocols and better descriptions of methods are important, but insufficient for reproducibility – and might not be feasible for every paper. Current requirements for wide-ranging experiments in a single article are part of the problem. To solve these issues, expectations placed on the scientific paper must change.

Reproducibility is costly

Research articles in the life sciences are more ambitious than ever. The amount of data in high-impact journals has doubled over 20 years[2], and basic-science papers are increasingly expected to include evidence of how results will translate to clinical applications. An article in a journal such as *Nature* thus ends up representing many years of work by several people.

Still, that's no guarantee of replicability. The Reproducibility Project: Cancer Biology has so far managed to replicate the main findings in only 5 of 17 highly cited articles[3], and a replication of 21 social-sciences articles in *Science* and *Nature* had a success rate of between 57 and 67% (ref. 4).

Many calls have been made to improve this scenario. Proposed measures include increasing sample sizes, preregistering protocols and using stricter statistical analyses. Another proposal is to introduce heterogeneity in methods and models to evaluate robustness – for instance, using more than one way to suppress gene expression across a variety of cell lines or rodent strains. In our work on the initiative, we have come to appreciate the amount of effort involved in following these proposals for a single experiment, let alone for an entire paper.

Even in a simple RT-PCR experiment, there are dozens of steps in which methods can vary, as well as a breadth of controls to assess the purity, integrity and specificity of materials. Specifying all of these steps in advance represents an exhaustive and sometimes futile process, because protocols inevitably have to be adapted along the way. Recording the entire method in an auditable way generates spreadsheets with hundreds of rows for every experiment.

We do think that the effort will pay off in terms of reproducibility. But if every paper in discovery science is to adopt this mindset, a typical high-profile article might easily take an entire decade of work, as well as a huge budget. This got us thinking about other, more efficient ways to arrive at reliable science.

AMILCAR ORFALI/GETTY

A researcher prepares samples for RT-PCR, which measures specific messenger RNAs.

A stepwise process

There are typically three main expectations for a top-notch article in laboratory science: first, report original findings from exploratory research; second, confirm that they represent robust phenomena through further experiments using different methods; and, finally, suggest theoretical mechanisms to explain the results. However, these represent different aspects of the scientific process and do not have to be achieved all at once[5].

In fact, trying to live up to all three expectations in a five-page paper can be a recipe for not fulfilling any of them well. Forcing exploratory and confirmatory research into a single publication can undermine both, either by stifling the former or corrupting the latter. Pressure to confirm an initial, exciting observation can bias subsequent data and analysis, particularly if certain results are required in further experiments to get the paper accepted. Rather than being sceptical of their original observation, many researchers will naturally distrust or dismiss further data that refute their hypothesis and jeopardize publication.

Moreover, requiring a large number of experiments in a single article can work against rigour: it shifts the workload towards many fragile experiments rather than a few robust ones. Studies have shown that neither statistical power[6] nor quality of reporting of individual experiments[7] improve as journal impact increases. And the amount and variety of data from many experiments can overwhelm reviewers' capacity to scrutinize evidence.

Finally, because a research group working on its own is inevitably limited in how much it can vary methods, models or conditions, most articles end up basing their conclusions on constrained data, without assessing generalizability[8]. In our work in the initiative, we were repeatedly surprised by the different ways researchers filled in the gaps in descriptions from original articles' protocols. Take experiments on macrophages obtained from the peritoneal cavity of mice, for example. Some labs used drugs to boost the number of these white blood cells, while others refrained out of concern that this would alter cellular responses. Most teams assessed the fraction of cultured cells that are viable macrophages – but there was little agreement on what fraction is high enough for experiments to proceed. Obtaining similar results under these different conditions can boost confidence that a phenomenon is robust; however, introducing such variability in methods is often beyond what a lab can do on its own.

Articles by individual research groups should thus be regarded as preliminary by default. If the expectation is that results of every publication hold true in other settings, models or populations, a reproducibility crisis seems inevitable. Instead of asking every author to conduct a decade's worth of confirmatory experiments, the scientific enterprise might be better served by other mechanisms to establish the validity of a claim – perhaps beyond the scope of a paper.

Paths to reproducible science

What other ways are there to assess whether findings are robust enough? One option is to

synthesize the published literature, drawing on results from studies by different research groups. This already happens for most clinical guidelines, which are typically derived from a meta-analysis of existing evidence. This approach, however, is marred by publication bias and incomplete reporting in primary studies. Thus, assessing reliability by this method still requires widespread problems to be fixed.

A potentially better approach is to organize confirmatory experiments that are specifically designed to assess robustness and generalizability. These will ideally incorporate multiple methods and experimental models (such as mouse strains or cell types) in different laboratories. Coordination between groups can standardize data collection and guarantee access to results, thus facilitating synthesis and eliminating publication bias.

Diverse types of collaboration have been set up across various areas of science. The pharmaceutical industry has managed multicentre clinical trials for decades. Consortia working in genetic epidemiology pool samples from different populations to increase statistical power. Academic psychology labs have joined forces for community efforts such as the Reproducibility Project: Psychology, the Many Labs Projects and the Psychological Science Accelerator. And initiatives in neuroscience include the International Brain Laboratory, the Human Connectome Project and the ENIGMA consortium.

Such endeavours are intensive in terms of cost and labour, and cannot be conducted for every published finding. Still, they are a more efficient way to confirm key phenomena than waiting for data to accrue from uncoordinated efforts. Moreover, investing effort to increase rigour in selected confirmatory projects is probably more feasible than demanding that every biomedical publication be replicable, generalizable and clinically relevant.

Divide labour, foster collaboration

Other authors have argued that exploratory research that generates tentative findings should be more clearly separated from confirmatory projects that evaluate them, as a way of improving both ends of the process[9]. Independence between exploratory and confirmatory work can allow greater freedom for scientists to explore hypotheses, while upholding rigour and preventing bias when they are put to the test. Moreover, each approach requires a different set of abilities and should be evaluated by distinct metrics.

Basic exploratory science would be helped if editorial policies reduced requests for new experiments and refrained from demanding evidence of clinical potential. Exploration can also benefit from forums that publish isolated findings of limited scope, as long as experiments and analyses are impartial. This can aid review, reduce bias and accelerate dissemination, while reducing incentives to dress up exploratory research as confirmatory work by cutting corners — or descriptions of unsuccessful experiments — to tell a coherent story.

Large-scale confirmatory science, by contrast, requires supportive infrastructure that is rarely available. There needs to be training, funding and rewards for researchers to focus on managing collaborations, participating in large experiments and synthesizing data — especially because this involves sacrificing academic freedom to some extent. If coordinated efforts to confirm published findings become routine, they can also stimulate the average scientist to be more rigorous in evaluating findings before publication, ultimately improving the quality of exploratory research.

"A better mechanism might be to build formal systems to manage collaborative projects."

All of this, however, requires reorganizing scientific labour, and one thing our initiative has taught us is that academic researchers are not used to being told what to do. Large-scale collaborations thus need to be centralized enough to guarantee rigour and adherence to guidelines, but should maintain some flexibility to accommodate each lab's own work routines.

Key to our strategy has been asking the right questions rather than being prescriptive. Requiring researchers to register how they will blind their study is more flexible than enforcing how they do it, but still serves to eliminate bias. Another key point has been to develop tools that enable best practices — from automatically randomizing sample distribution on plates to standardizing spreadsheets for data collection.

Despite all this, we worry that grassroots efforts such as ours might not be scalable. Not only has the initiative kept the coordinating team absorbed for the past three years, but it has also frequently collided with other priorities in our collaborating labs.

A better mechanism might be to build formal systems to manage collaborative projects, driven by institutions or funders. Such collaborations already exist in specific areas, as exemplified by efforts from the US National Institute on Aging[10], the US Defense Advanced Research Projects Agency (DARPA)[11] and the German Federal Ministry of Education and Research[12]. Still, there is room for them to become much more widespread, and perhaps as much a part of biomedical science as grant applications or peer review.

Changing our expectations

Although there is scope to make the average paper more rigorous, an overemphasis on individual papers and their reproducibility should not detract us from other means of arriving at sound conclusions. Instead of expecting that every paper will establish reliable phenomena, it might be more feasible to improve systematic confirmation of preliminary findings.

For this to happen, the biomedical science community needs to be convinced that some resources should be diverted to larger projects investigating fewer ideas. Funders and institutions must be more proactive in coordinating the scientific workforce to select and address key research questions, rather than scattering resources between competing labs. This involves building incentive systems — in terms of funding, career advancement and credit — to encourage researchers to take on less autonomous roles in larger projects. Scientific societies and journals can also play a part in determining which findings in a given research field are considered crucial for replication — a tough decision that requires extensive input from the scientific community.

Moving the burden of reproducibility from individual researchers to organized communities can ultimately raise the bar of what is considered scientific fact, and could also have a salutary effect on the public communication of science. The ideal way to achieve all of this remains an open question. But we can at least agree that it is larger than what fits in a paper.

The authors

Olavo B. Amaral and **Kleber Neves** are neuroscientists working in meta-research at the Federal University of Rio de Janeiro, Brazil, and coordinators of the Brazilian Reproducibility Initiative.
e-mails: olavo@bioqmed.ufrj.br; kleber.neves@bioqmed.ufrj.br

1. Amaral, O. B., Neves, K., Wasilewska-Sampaio, A. P. & Carneiro, C. F. *eLife* **8**, e41602 (2019).
2. Cordero, R. J. B., de León-Rodriguez, C. M., Alvarado-Torres, J. K., Rodriguez, A. R. & Casadevall, A. *PLoS ONE* **11**, e0156983 (2016).
3. Errington, T. M. *et al. eLife* **3**, e04333 (2014).
4. Camerer, C. F. *et al. Nature Hum. Behav.* **2**, 637–644 (2018).
5. Haig, B. Preprint at PsyArXiv https://doi.org/10.31234/osf.io/v784s (2020).
6. Brembs, B., Button, K. & Munafò, M. *Front. Hum. Neurosci.* **7**, 291 (2013).
7. Carneiro, C. F. D. *et al. Res. Integr. Peer Rev.* **5**, 16 (2020).
8. Yarkoni, T. *Behav. Brain Sci*. https://doi.org/10.1017/s0140525x20001685 (2020).
9. Kimmelman, J., Mogil, J. S. & Dirnagl, U. *PLoS Biol.* **12**, e1001863 (2014).
10. Lithgow, G. J., Driscoll, M. & Phillips, P. *Nature* **548**, 387–388 (2017).
11. Raphael, M. P., Sheehan, P. E. & Vora, G. J. *Nature* **579**, 190–192 (2020).
12. Drude, N. I., Gamboa, L. M., Danziger, M., Dirnagl, U. & Toelch, U. *eLife* **10**, e62101 (2021).

The authors declare no competing interests.