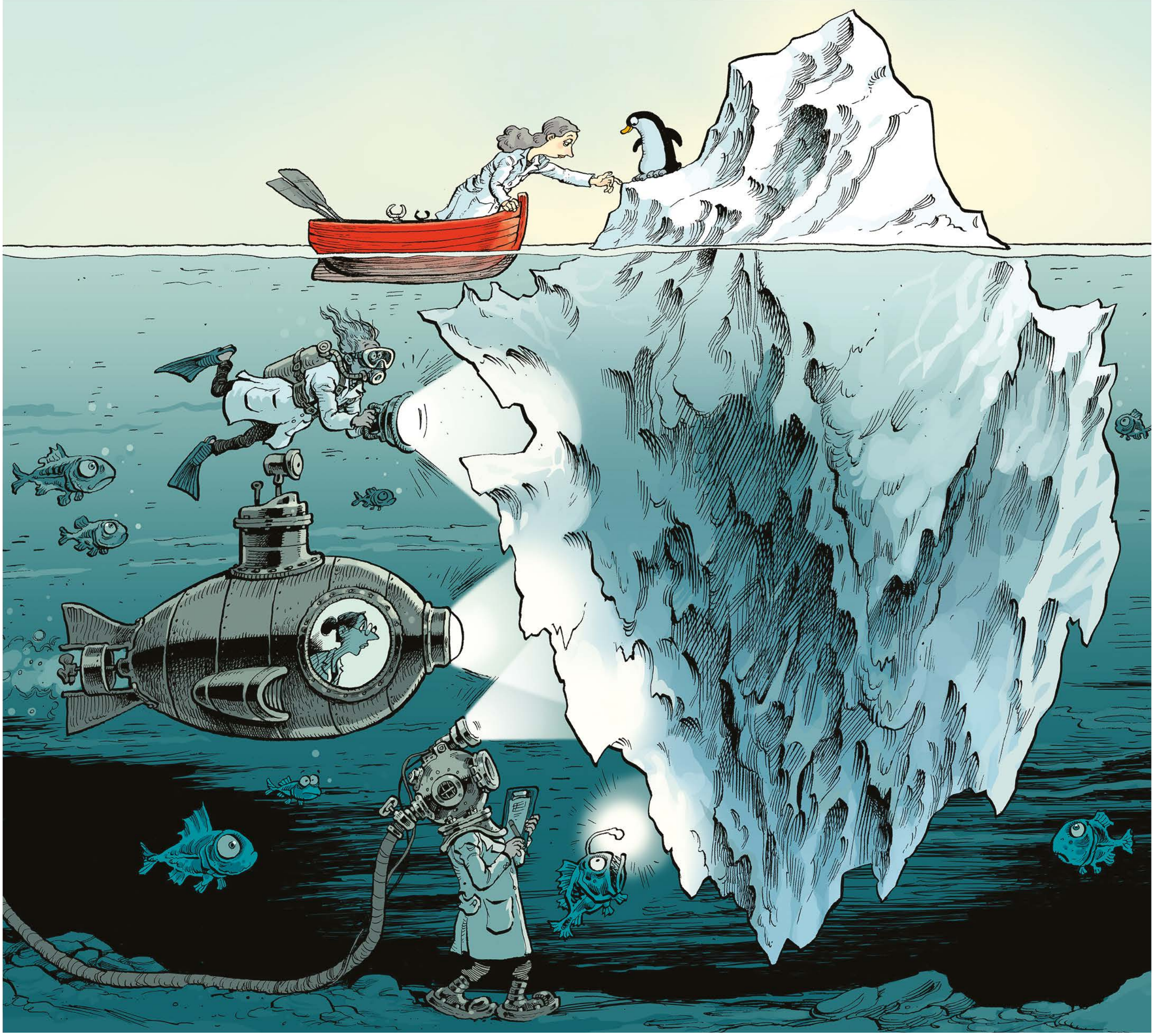


ILLUSTRATION BY DAVID PARKINS

One statistical analysis must not rule them all

Eric-Jan Wagenmakers, Alexandra Sarafoglou & Balazs Aczel

Any single analysis hides an iceberg of uncertainty. Multi-team analysis can reveal it.

A typical journal article contains the results of only one analysis pipeline, by one set of analysts. Even in the best of circumstances, there is reason to think that judicious alternative analyses would yield different outcomes.

For example, in 2020, the UK Scientific Pandemic Influenza Group on Modelling asked nine teams to calculate the reproduction number *R* for COVID-19 infections[1]. The teams chose from an abundance of data (deaths, hospital admissions, testing rates) and modelling approaches. Despite the clarity of the question, the variability of the estimates across teams was considerable (see 'Nine teams, nine estimates').

On 8 October 2020, the most optimistic estimate suggested that every 100 people with COVID-19 would infect 115 others, but perhaps as few as 96, the latter figure implying that the pandemic might actually be retreating. By contrast, the most pessimistic estimate had 100 people with COVID-19 infecting 166 others, with an upper bound of 182, indicating a rapid spread. Although the consensus was that the trajectory of disease spread was cause for concern, the uncertainty across the nine teams was considerably larger than the uncertainty within any one team. It informed future work as the pandemic continued.

Flattering conclusion

This and other 'multi-analyst' projects show that independent statisticians hardly ever use the same procedure[2–6]. Yet, in fields from ecology to psychology and from medicine to materials science, a single analysis is considered sufficient evidence to publish a finding and make a strong claim.

Over the past ten years, the concept of *P*-hacking has made researchers aware of how the ability to use many valid statistical procedures can tempt scientists to select the one that leads to the most flattering conclusion. Less understood is how restricting analyses to a single technique effectively blinds researchers to an important aspect of uncertainty, making results seem more precise than they really are.

To a statistician, uncertainty refers to the range of values that might reasonably be taken by, say, the reproduction number of COVID-19 or the correlation between religiosity and well-being[6], or between cerebral cortical thickness and cognitive ability[7], or any number of statistical estimates. We argue that the current mode of scientific publication — which settles for a single analysis — entrenches 'model myopia', a limited consideration of statistical assumptions. That leads to overconfidence and poor predictions.

To gauge the robustness of their conclusions, researchers should subject the data to multiple analyses; ideally, these would be carried out by one or more independent teams. We understand that this is a big shift in how science is done, that appropriate infrastructure and incentives are not yet in place, and that many researchers will recoil at the idea as being burdensome and impractical. Nonetheless, we argue that the benefits of broader, more-diverse approaches to statistical inference could be so consequential that it is imperative to consider how they might be made routine.

Charting uncertainty

Some 100 years ago, scholars such as Ronald Fisher advanced formal methods for hypothesis testing that are now considered indispensable for drawing conclusions from numerical data. (The *P* value, often used to determine 'statistical significance', is the best known.) Since then, a plethora of tests and methods have been developed to quantify inferential uncertainty. But any single analysis draws on a very limited range of these. We posit that, as currently applied, uncertainty analyses reveal only the tip of the iceberg.

The dozen or so formal multi-analyst projects completed so far (see Supplementary information) show that levels of uncertainty are much higher than that suggested by any single team. In the 2020 Neuroimaging Analysis Replication and Prediction Study[2], 70 teams used the same functional magnetic resonance imaging (MRI) data to test 9 hypotheses about brain activity in a risky-decision task. For example, one hypothesis probed how a brain region is activated when people consider the prospect of a large gain. On average across the hypotheses, about 20% of the analyses constituted a 'minority report' with a qualitative conclusion opposite to that of the majority. For the three hypotheses that yielded the most ambiguous outcomes, around one-third of teams reported a statistically significant result, and therefore publishing work from any of one these teams would have hidden considerable uncertainty and the spread of possible conclusions. The study's coordinators now advocate that multiple analyses of the same data be done routinely.

> "Formal methods cannot cure model myopia, because they are firmly rooted in the single-analysis framework."

Another multi-analyst project was in finance[3] and involved 164 teams that tested 6 hypotheses, such as whether market efficiency changes over time. Here again, the coordinators concluded that differences in findings were due not to errors, but to the wide range of alternative plausible analysis decisions and statistical models.

All of these projects have dispelled two myths about applied statistics. The first myth is that, for any data set, there exists a single, uniquely appropriate analysis procedure. In reality, even when there are scores of teams and the data are relatively simple, analysts almost never follow the same analytic procedure.

The second myth is that multiple plausible analyses would reliably yield similar conclusions. We argue that whenever researchers report a single result from a single statistical analysis, a vast amount of uncertainty is hidden from view. And although we endorse recent science-reform efforts, such as large-scale replication studies, preregistration and registered reports, these initiatives are not designed to reveal statistical fragility by exploring the degree to which plausible alternative analyses can alter conclusions. In summary, formal methods, old and new, cannot cure model myopia, because they are firmly rooted in the single-analysis framework.

We need something else. The obvious treatment for model myopia is to apply more than one statistical model to the data. High-energy physics and astronomy have a strong tradition of teams carrying out their own analyses of other teams' research once the data are made public. Climate modellers routinely perform 'sensitivity analyses' by systematically removing and including variables to see how robust their conclusions are.

For other fields to make such a shift, journals, reviewers and researchers will have to change how they approach statistical inference. Instead of identifying and reporting the result of a single 'correct' analysis, statistical inference should be seen as a complex interplay of different plausible procedures and processing pipelines[8]. Journals could encourage this practice in at least two ways. First, they could adjust their submission guidelines to recommend the inclusion of multiple analyses (possibly reported in an online supplement)[9]. This would motivate researchers to either conduct extra analyses themselves or to recruit more analysts as co-authors. Second, journals could invite teams to contribute their own analyses in the form of comments on a recently accepted article.

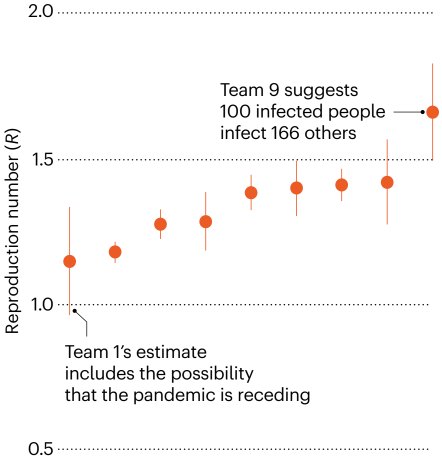
False alarm?

Certainly, large-scale changes in how science is done are possible: expectations surrounding the sharing of data are growing. Medical journals now require that clinical trials be registered at launch for the results to be published. But proposals for change inevitably prompt critical reactions. Here are five that we've encountered.

Won't readers get confused? Currently, there are no comprehensive standards for, or conventions on, how to present and interpret the

NINE TEAMS, NINE ESTIMATES

Comparing models of the rate of COVID-19's spread in the United Kingdom in early October 2020 revealed a degree of uncertainty masked by any one model.



SOURCE: REF. 1

results of multiple analyses, and this situation could complicate how results are reported and make conclusions more ambiguous. But we argue that potential ambiguity is a key feature of multi-team analysis, not a bug. When conclusions are supported only by a subset of plausible models and analyses, readers should be made aware. Facing uncertainty is always better than sweeping it under the rug.

Aren't other problems more pressing? Problems in empirical science include selective reporting, a lack of transparency around analyses, hypotheses that are divorced from the theories they are meant to support, and poor data sharing. It is important to make improvements in these areas — indeed, how data are collected and processed, and how variables are defined, will greatly influence all subsequent analyses. But multi-analyst approaches can still bring insight. In fact, multi-analyst projects usually excel in data sharing, transparent reporting and theory-driven research. We view the solutions to these problems as mutually reinforcing rather than as a zero-sum game.

Is it really worth the time and effort? Even those who see benefit in multiple analyses might not see a need for them to happen at the time of publication. Instead, they would argue that the original team be encouraged to pursue multiple analyses or that shared data can be reanalysed by other interested researchers after publication. We agree that both would be an improvement over the status quo (sensitivity analysis is a severely underused practice). However, they will not yield the same benefits as multi-team analyses done at the time of publication.

Post-publication analyses are usually published only if they drastically undercut the original conclusion. They can give rise to squabbles more than constructive discussion, and would come out after the authors and readers have already drawn conclusions based on a single analysis. Information about uncertainty is most useful at the time of analysis. However, we doubt whether a single team can muster the mental fortitude needed to reveal the fragility of their findings; there might be a strong temptation to select those analyses that, together, present a coherent story. In addition, a single research team usually has a somewhat narrow expertise in data analysis. For instance, each of the nine teams that produced different estimates for *R* would probably feel uncomfortable if they had to code and produce estimates using the other teams' models. Even for simple statistical scenarios (that is, a comparison of two outcomes — such as the proportions of people who improve after receiving a drug or placebo — and a test of a linear correlation), several teams can apply widely divergent statistical models and procedures[10].

Some sceptics doubt that multi-team analyses will consistently find broad enough ranges of results to make the effort worthwhile. We think that the outcomes of existing multi-analyst projects counter that argument, but it would be useful to gather evidence from yet more projects. The more multi-analyst approaches are undertaken, the clearer it will be as to how and when they are valuable.

Won't journals baulk? One sceptical response to our proposal is that multi-analyst projects will take longer, be more complicated to present and assess, and will even require new article formats — complications that will make journals reluctant to embrace the idea. We counter that the review and publication of a multi-analyst paper do not require a fundamentally different process. Multi-team projects have been published in a variety of journals, and most journals already publish comments attached to accepted manuscripts. We challenge journal editors to give multi-analyst projects a chance. For instance, editors might test the waters by organizing a special issue consisting of case studies. This should make it readily apparent whether the added value of the multi-analyst approach is worth the extra effort.

> **"Journals, governments and philanthropists should actively recruit or support multi-analysis teams."**

Won't it be a struggle to find analysts? One response to our proposal is that the bulk of multi-team analyses published so far are the product of demonstration projects wrapped into a single paper. These papers encompass several analyses with long author lists comprised mainly of enthusiasts for reform; most other researchers would see little benefit in being a minor contributor to a multi-analyst paper, especially one at the periphery of their core research interest. But we think enthusiasm has a broad base. In our multi-analyst projects, we have been known to receive more than 700 sign-ups in about 2 weeks.

Moreover, a range of incentives could attract teams of analysts, such as gaining co-authorship and the chance to work on important questions or simply to collaborate with specialists. Further incentives and catalysts are easy to imagine. In a forthcoming special issue of the journal *Religion, Brain & Behavior*, several teams will each publish their own conclusions and interpretations of the research question addressed by the main article[6], and this means each teams' contribution is individually recognized. When a question is particularly urgent, journals, governments and philanthropists should actively recruit or support multi-analysis teams.

Yet another approach would be to incorporate multiple analyses into training programs, which would be both useful for the research community and eye-opening for statisticians. (At least one university has incorporated replication studies into its curricula[11].) Ideally, participating in multiple analyses will be seen as part of being a good science 'citizen', and be rewarded through better prospects for hiring and promotion.

Whatever the mix of incentives and formats, the more that multiple analyses efforts are implemented and discussed, the easier they will become. What makes such multi-team efforts work well should be studied and applied to improve and expand the practice. As the scientific community learns how to run multi-team analyses and what can be learnt, acceptance and enthusiasm will grow.

We argue that rejecting the multi-analyst vision would be like Neo opting for the blue pill in the film *The Matrix*, and so continuing to dream of a reality that is comforting but false. Scientists and society will be better served by confronting the potential fragility of reported statistical outcomes. It is crucial for researchers and society to have an indication of such fragility from the moment the results are published, especially when these results have real-world ramifications. Recent many-analyst projects suggest that any single analysis will yield conclusions that are overconfident and unrepresentative. Overall, the benefit of increased insight will outweigh the extra effort.

The authors

Eric-Jan Wagenmakers is a methodologist and **Alexandra Sarafoglou** a postdoctoral fellow at the University of Amsterdam, the Netherlands. **Balazs Aczel** is vice dean for science at Eötvös Loránd University in Budapest, Hungary.
e-mails: ej.wagenmakers@gmail.com; alexandra.sarafoglou@gmail.com; balazs.aczel@gmail.com

1. Scientific Pandemic Influenza Group on Modelling. *SPI-M-O: Consensus statement on COVID-19, 8 October 2020* (2020).
2. Botvinik-Nezer, R. *et al. Nature* **582**, 84–88 (2020).
3. Menkveld, A. J. *et al.* Preprint at SSRN https://doi.org/10.2139/ssrn.3961574 (2021).
4. Silberzahn, R. *et al. Adv. Methods Pract. Psychol. Sci.* **1**, 337–356 (2018).
5. Steegen, S., Tuerlinckx, F., Gelman, A. & Vanpaemel, W. *Perspect. Psychol. Sci.* **11**, 702–712 (2016).
6. Hoogeveen, S. *et al.* Preprint at PsyArXiv https://doi.org/10.31234/osf.io/pbfye (2022).
7. Marek, S. *et al. Nature* **603**, 654–660 (2022).
8. Wagenmakers, E.-J. *et al. Nature Hum. Behav.* **5**, 1473–1480 (2021).
9. Aczel, B. *et al. eLife* **10**, e72185 (2021).
10. van Dongen, N. N. N. *et al. Am. Stat.* **73**, 328–339 (2019).
11. Button, K. *Nature* **561**, 287 (2018).

The authors declare no competing interests.
Supplementary information accompanies this Comment online (see go.nature.com/37hdvkb).