Consensus-based guidance for conducting and reporting multi-analyst studies

Authors' names: Balazs Aczel[1*], Barnabas Szaszi[1*], Gustav Nilsonne[2,3], Olmo R van den Akker[4], Casper J Albers[5], Marcel A L M van Assen[4,6], Jojanneke A Bastiaansen[7,8], Dan Benjamin[9,10], Udo Boehm[11], Rotem Botvinik-Nezer[12], Laura F Bringmann[5], Niko A Busch[13], Emmanuel Caruyer[14], Andrea M Cataldo[15,16], Nelson Cowan[17], Andrew Delios[18], Noah N N van Dongen[11], Chris Donkin[19], Johnny B van Doorn[11], Anna Dreber[20,21], Gilles Dutilh[22], Gary F Egan[23], Morton Ann Gernsbacher[24], Rink Hoekstra[5], Sabine Hoffmann[25], Felix Holzmeister[21], Juergen Huber[21], Magnus Johannesson[20], Kai J Jonas[26], Alexander T Kindel[27], Michael Kirchler[21], Yoram K Kunkels[7], D Stephen Lindsay[28], Jean-Francois Mangin[29,30], Dora Matzke[11], Marcus R Munafò[31], Ben R Newell[19], Brian A Nosek[32,33], Russell A Poldrack[34], Don van Ravenzwaaij[5], Jörg Rieskamp[35], Matthew J Salganik[27], Alexandra Sarafoglou[11], Tom Schonberg[36], Martin Schweinsberg[37], David Shanks[38], Raphael Silberzahn[39], Daniel J Simons[40], Barbara A Spellman[33], Samuel St-Jean[41,42], Jeffrey J Starns[43], Eric L Uhlmann[44], Jelte Wicherts[4], Eric-Jan Wagenmakers[11]

Affiliations: [1]ELTE, Eotvos Lorand University, Budapest, Hungary, [2]Karolinska Institutet, Stockholm, Sweden, [3]Stockholm University, Stockholm, Sweden, [4]Tilburg University, Tilburg, The Netherlands, [5]University of Groningen, Groningen, The Netherlands, [6]Utrecht University, Utrecht, The Netherlands, [7]University of Groningen, University Medical Center Groningen, Groningen, The Netherlands, [8]Friesland Mental Health Care Services, Leeuwarden, The Netherlands, [9]University of California Los Angeles, Los Angeles, CA, USA, [10]National Bureau of Economic Research, Cambridge, MA, USA, [11]University of Amsterdam, Amsterdam, The Netherlands, [12]Dartmouth College, Hanover, NH, USA, [13]University of Münster, Münster, Germany, [14]University of Rennes, CNRS, Inria, Inserm, Rennes, France, [15]McLean Hospital, Belmont, MA, USA, [16]Harvard Medical School, Boston, MA, USA, [17]Department of Psychological Sciences, University of Missouri, MO, USA, [18]National University of Singapore, Singapore, [19]University of New South Wales, Sydney, Australia, [20]Stockholm School of Economics, Stockholm, Sweden, [21]University of Innsbruck, Innsbruck, Austria, [22]University Hospital Basel, Basel, Switzerland, [23]Monash University, Melbourne, Victoria, Australia, [24]University of Wisconsin-Madison, Madison, WI, USA, [25]Ludwig-Maximilians-University, Munich, Germany, [26]Maastricht University, Maastricht, The Netherlands, [27]Princeton University, Princeton, NJ, USA, [28]University of Victoria, Victoria, Canada, [29]Université Paris-Saclay, Paris, France, [30]Neurospin, CEA, France, [31]University of Bristol, Bristol, UK, [32]Center for Open Science, USA, [33]University of Virginia, Charlottesville, USA, [34]Stanford University, Stanford, USA, [35]University of Basel, Basel, Switzerland, [36]Tel Aviv University, Tel Aviv, Israel, [37]ESMT Berlin, Germany, [38]University College London, London, UK, [39]University of Sussex, Brighton, UK, [40]University of Illinois at Urbana-Champaign, USA, [41]University of Alberta, Edmonton, Canada, [42]Lund University, Lund, Sweden, [43]University of Massachusetts Amherst, USA, [44]INSEAD, Singapore

[*]Correspondence: aczel.balazs@ppk.elte.hu; szaszi.barnabas@ppk.elte.hu

Abstract

Any large dataset can be analyzed in a number of ways, and it is possible that the use of different analysis strategies will lead to different results and conclusions. One way to assess whether the results obtained depend on the analysis strategy chosen is to employ multiple analysts and leave each of them free to follow their own approach. Here, we present consensus-based guidance for conducting and reporting such multi-analyst studies, and we discuss how broader adoption of the multi-analyst approach has the potential to strengthen the robustness of results and conclusions obtained from analyses of datasets in basic and applied research.

Introduction

Empirical investigations often require researchers to make a large number of decisions about how to analyze the data. However, the theories that motivate investigations rarely impose strong restrictions on how the data should be analyzed. This means that empirical results typically hinge on analytical choices made by just one or a small number of researchers, and raises the possibility that different – but equally justifiable – analytical choices could lead to different results (Figure 1).

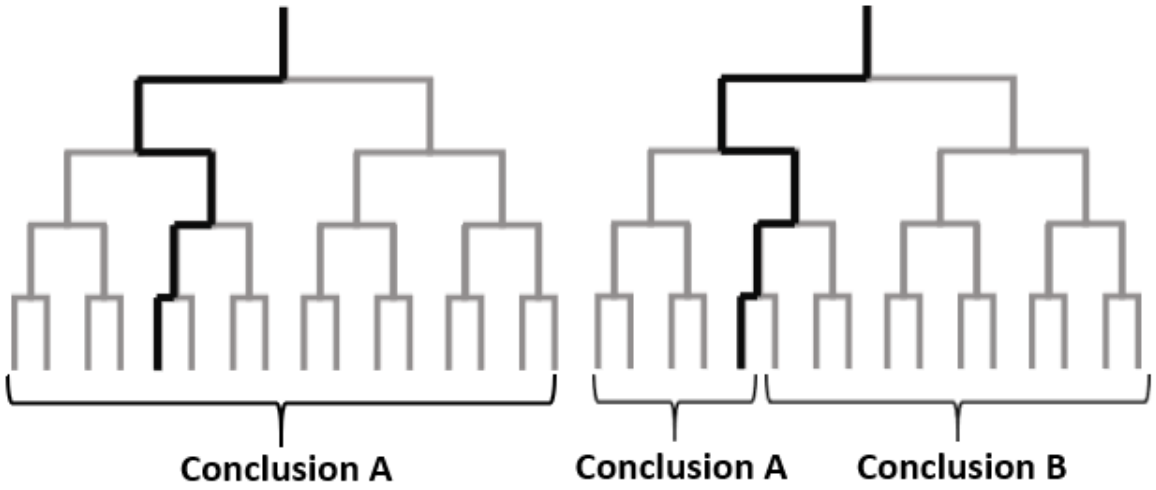

Figure 1.
Analysis choices and alternative plausible paths.
The analysis of a large dataset can involve a sequence of analysis choices, as depicted in these schematic diagrams. Tthe analyst first must decide between two options at the start of the analysis (top), and must make three additional decisions during the analysis: this leads to 16 possible paths for the analysis (grey lines). The left panel shows an example in which all possible paths lead to the same conclusion; the right panel shows an example in which some paths lead to conclusion A and other paths lead to conclusion B. Unless we can test alternative paths, we cannot know if the results obtained by following one particular path (thick black line) are robust, or if other plausible paths would lead to different results.

This "analytical variability" may be particularly high for datasets that were not initially collected for research purposes (such as electronic health records) because data analysts might know relatively little about how those data were collected and/or generated. However, when analyzing such datasets – and when making decisions based on the results of such analyses – it is important to be aware that the results will be subject to higher levels of analytical variability than the results obtained from analyses of data from, say, clinical trials.

A recent example of the perils of analytical variability is provided by two articles in the journal *Surgery* that used the same dataset to investigate the same question: does the use of a retrieval bag during laparoscopic appendectomy reduce surgical site infections? Each paper used reasonable analysis, but there were notable differences between them in how they addressed inclusion and exclusion criteria, outcome measures, sample sizes, and covariates. As a result of these different analytical choices, the two articles reached opposite conclusions: one paper reported that using a retrieval bag reduced infections (1), and the other reported that it did not (2; see also 3). This and other medical examples (4–6) illustrate how independent analysis of the same data can reach different, yet justifiable, conclusions. tr

The robustness of results and conclusions can be studied by evaluating multiple distinct analysis options simultaneously (e.g., vibration of effects (7) or multiverse analysis (8)), or by employing a "multi-analyst approach" that involves engaging multiple analysts to independently analyze the same data. Rather than exhaustively evaluating all plausible analyses, the multi-analyst approach examines analytical choices that are deemed most appropriate by independent analysts. Botvinik-Nezer et al. (13), for example, asked 70 teams to test the same hypotheses using the same functional magnetic resonance imaging dataset. They found that no two teams followed the same data preprocessing steps or analysis strategies, which resulted in substantial variability in the teams' conclusions. This and other work (9–12, 14–18) confirms how results can depend on analytic choices.

Although the multi-analyst approach will be new to many researchers, it has been in use since the 19th century. In 1857, for example, the Royal Asian Society asked four scholars to independently translate a previously unseen inscription to verify that the ancient Assyrian language had been deciphered correctly. The almost perfect overlap between the solutions indicated that "they have Truth for their basis" (19). The same approach can be used to analyze data today. With just a few co-analysts, the multi-analyst approach can be informative about the analytic robustness of results and conclusions. When the results of independent data analyses converge, more confidence in the conclusions is warranted. However, when the results diverge, confidence will be reduced, and scientists can examine the reasons for these discrepancies and identify potentially meaningful moderators of the results. With enough co-analysts, it is possible to estimate the variability among analysis strategies and attempt to identify factors explaining this variability.

The multi-analyst approach is still rarely used, but we argue that many disciplines could benefit from its broader adoption. To help researchers overcome practical challenges, we provide consensus-based guidance (including a checklist) to help researchers surmount the practical challenges of preparing, conducting, and reporting multi-analyst studies.

Methods

To develop this guidance, we recruited a panel of 50 methodology experts who followed a preregistered 'reactive-Delphi' expert consensus procedure (20). We adopted this procedure to ensure that the resulting guidance represents the shared thinking of relevant experts and that it incorporates their topic-related insights. The applied consensus procedure and its reporting satisfy the recommendations of CREDES (21), a guidance on conducting and reporting Delphi studies. A flowchart of the Delphi expert consensus procedure is available at https://osf.io/pzkcs/.

Preparation

Preregistering the project

Before the start of the project, on 11 November 2020, a research plan was compiled and uploaded to a time-stamped repository at https://osf.io/dgrua. During the project, we followed the preregistered plan in all respects except implementing slight changes in the wording of the survey questions to improve comprehension and not using R to analyze our results. We declared that we would share the R code and codebook of our analyses, but the project ultimately did not require us to conduct analyses in R. Instead, we shared our code in Excel and ODS format at https://osf.io/h36qy/.

Creating the initial Multi-Analyst Guidance draft

Before the expert consensus process, the first three authors and the last author (henceforth: proposers) created an initial multi-analyst guidance draft after brainstorming and reviewing all the previously published multi-analyst-type projects they were aware of (9–18). This initial document is available here: https://osf.io/kv8jt/

Recruiting experts

The proposers contacted 81 experts to join the project. The contacted experts included all the organizers of previous multi-analyst projects known at the time (9–18), as well as the members of the expert panel from another methodological consensus project (22). The previous projects were identified by conducting an unsystematic literature search and by surveying researchers in social media. Of the 81 experts, 3 declined our invitation and 50 accepted the invitation and participated in the expert consensus procedure (their names are available at https://osf.io/fwqvp/), while 28 experts did not respond to our call.

Preparatory rounds

Upon joining the project, the experts received a link to the preparatory online survey (available at https://osf.io/kv8jt/) which included the initial Multi-Analyst Guidance draft where they had the option to comment on each of the items and the overall content of the guidance.

Based on the feedback received from the preparatory online survey, the proposers updated and revised the initial Multi-Analyst Guidance. This updated document was uploaded to an online shared document and was sent out to the experts who had the option to

edit and comment on the content. Again, based on feedback, the proposers revised the content of the document, and this new version was included in the expert consensus survey.

Consensus survey

The expert consensus questionnaire was sent out individually to each expert first on 8 February 2021 in the following Qualtrics survey available at https://osf.io/wrpnq/. The consensus survey approach had the advantage of minimizing potential biases in the experts' judgments: the questions were posed in a neutral way, experts all received the same questions, and experts did not see the responses of the other experts or any reaction of the project organizers. The survey contained the ten recommended practices grouped into the following five stages: i) recruiting co-analysts; ii) providing the dataset, research questions, and research tasks; iii) conducting the independent analyses; iv) processing the results; v) reporting the methods and results. The respondents were asked to rate each of the ten recommended practices on a nine-point Likert-type scale ('I agree with the content and wording of this guidance section' ranging from "1-Disagree" to "9-Agree"). Following each section, the respondents could leave comments regarding the given item.

The preregistration indicated consensus on the given item if the interquartile range of its ratings was two or smaller. It defined support for an item if the median rating was six or higher (as in 22).

Each recommended practice found support and consensus from the 48 experts who completed ratings in our first round. For each item, the median rating was eight or higher with an interquartile range of two or lower. Thus, following our preregistration, there was no need to conduct additional consensus-survey rounds; all of the items were eligible to enter the guidance with consensual support. This high level of consensus might have been due to the experts' involvement in the preparatory round of the project. The summary table of the results is available at https://osf.io/qc7a8/.

Finalising the manuscript

The proposers drafted the manuscript and supplements. All texts and materials were sent to the expert panel members. Each contributor was encouraged to provide feedback on the manuscript, the report, and the suggested final version of the guidance. After all discussions, minor wording changes were implemented, as documented at https://osf.io/e39j4/. No contributor objected to the content and form of the submitted materials and all approved the final item list.

Multi-analyst guidance

The final guidance includes ten recommended practices (Table 1) concerning the five main stages of multi-analyst studies. To further assist researchers in documenting multi-analyst projects, we also provide a modifiable reporting template (Supplementary file 1), as well as a reporting checklist (Supplementary file 2).

Table 1
Recommended Practices for the Main Stages of the Multi-Analyst Method

Stage	Recommended practices
Recruiting co-analysts	1. Determine a minimum target number of co-analysts and outline clear eligibility criteria before recruiting co-analysts. We recommend that the final report justifies why these choices are adequate to achieve the study goals. 2. When recruiting co-analysts, inform them about (a) their tasks and responsibilities; (b) the project code of conduct (e.g., confidentiality/ non-disclosure agreements); (c) the plans for publishing the research report and presenting the data, analyses, and conclusion; (d) the conditions for an analysis to be included or excluded from the study; (e) whether their names will be publicly linked to the analyses; (f) the co-analysts' rights to update or revise their analyses; (g) the project time schedule; and (h) the nature and criteria of compensation (e.g., authorship).
Providing datasets, research questions, and research tasks	3. Provide the datasets accompanied with a codebook that contains a comprehensive explanation of the variables and the datafile structure. 4. Ensure that co-analysts understand any restrictions on the use of the data, including issues of ethics, privacy, confidentiality, or ownership. 5. Provide the research questions (and potential theoretically derived hypotheses that should be tested) without communicating the lead team's preferred analysis choices or expectations about the conclusions.
Conducting the independent analyses	6. To ensure independence, we recommend that co-analysts should not communicate with each other about their analyses until after all initial reports have been submitted. In general, it should be clearly explained why and at what stage co-analysts are allowed to communicate about the analyses (e.g., to detect errors or call attention to outlying data points).
Processing the results	7. Require co-analysts to share with the lead team their results, the analysis code with explanatory comments (or a detailed description of their point-and-click analyses), their conclusions, and an explanation of how their conclusions follow from their results. 8. The lead team makes the commented code, results, and conclusions of all non-withdrawn analyses publicly available before or at the same time as submitting the research report.
Reporting the methods and results	9. The lead team should report the multi-analyst process of the study, including (a) the justification for the number of co-analysts; (b) the eligibility criteria and recruitment of co-analysts; (c) how co-analysts were given the data sets and research questions; (d) how the independence of analyses was ensured; (e) the numbers of and reasons for withdrawals and omissions of analyses; (f) whether the lead team conducted an independent analysis; (g) how the results were processed; (h) the summary of the results of co-analysts; (i) and the limitations and potential biases of the study. 10. Data management should follow the FAIR principles (23), and the research report should be transparent about access to the data and code for all analyses (22).

In addition to the Multi-analyst Guidance and Checklist, we provide practical considerations that can support the organization and execution of multi-analyst projects. This section contains various clarifications, recommendations, practical tools, and optional extensions, covering the five main stages of a multi-analyst project.

Recruiting co-analysts

Choosing co-analysts

The term co-analyst refers to one researcher or team of researchers working together in a multi-analyst project. Researchers can collaborate on the analyses, but if they do, we recommend that they submit the analyses as one co-analyst team, in order to ensure the independence of the analyses across teams. Researchers from the same lab or close collaborators should be able to submit separate reports in the multi-analyst project as long as they do not discuss their analyses with each other until the project rules allow that. The lead team may conduct an analysis themselves depending on the study goals and the design of the project (e.g., to set a performance baseline for comparing submitted models). Alternatively, the lead team may choose not to conduct an analysis themselves; in any case, they are expected to be transparent about their level of involvement as well as the timing (e.g., whether they conducted their analyses with or without knowing the results of the crowd of analysts).

Researchers should carefully consider both the breadth and depth of statistical and research-area expertise required for their project and should justify their choices about the required qualifications, skills, and credentials for analysts in the project. If the aim of the study is to explore what factors influence researchers' analytical choices, then it can be useful to seek "natural variation" (representativeness) within an expert community or to maximize diversity of the co-analysts along the dimensions where they might differ the most in their choices (e.g., experience, background, discipline, interest in the findings, intellectual allegiance to different theories, paradigmatic viewpoints).

Deciding on the number of co-analysts

To decide on the desired number of co-analysts, one has to consider which of the two main purposes of the multi-analyst method applies to the given project:

(A) Checking the robustness of the conclusions

The aim here is solely to check whether different analysts obtain the same conclusions. Confidence in the stability of the conclusions decreases with divergent results and increases with convergent results. Many projects can achieve this aim by recruiting only one additional analyst, or a handful of further analysts. For example, the above-mentioned two analyses of the same dataset published in the journal *Surgery* (1,2) were sufficient to detect that the analytical space allows for opposite conclusions.

(B) Assessing the variability of the analyses

Those who wish to estimate the variability among the different analysis strategies often need to satisfy stricter demands. For example, studies that aim to assess how much the results vary among the analysts will require a larger number of co-analysts. When determining the number of co-analysts in such cases, the same factors need to be taken into consideration as in standard sample size estimation methods. For example, Botvinik-Nezer et al. (13)

presented the analyses of 70 teams to demonstrate the divergence of results when analyzing a functional magnetic resonance imaging dataset.

Recruiting co-analysts

Depending on the specific goal of the research, the recruitment of co-analysts can happen in several ways. Co-analysts can be recruited before or after obtaining the dataset. With stricter eligibility criteria, co-analysts can be invited individually from among topic experts or statistical experts. Follow-up open invitations can ask experts to suggest others to be invited. Alternatively, the lead team can open the opportunity to anyone to join the project as a co-analyst within the expert community (e.g., in professional society mailing lists and on social media), where expertise can be defined as the topic requires it.

It is important to note that whenever the co-authors' behavior is the subject of the study then they should be regarded similarly to human participants respecting ethical and data protection regulations. Useful templates for project advertisement and analyst surveys can be found in (12,24).

Providing the dataset, research questions, and research tasks

Providing the dataset

The lead team can invite the co-analysts to conduct data preprocessing (in addition to the main analysis). If the lead team decides to conduct the preprocessing themselves, showing their preprocessing methods can be informative to the co-analysts, but also has the potential to influence them if the preprocessing reflects some preference of methods or expectations of outcomes.

Before providing the dataset, the lead team should ensure that data management will comply with legal (e.g., the General Data Protection Regulation (GDPR) in the European Union) and ethical regulations applying to all teams (see 25). If the dataset contains personal information, a version should be provided where data can no longer be related to an individual. An alternative is to provide a simulated dataset and ask the co-analysts to provide code to analyze the data (26,27). The lead team can then run the code on the actual data.

It is important that the co-analysts understand not just the available dataset but also any ancillary information that might affect their analyses (e.g., prior exclusion of outliers or handling of missing data in the blinded dataset). Providing a codebook that is accessible and understandable for researchers with different backgrounds is essential (28).

Providing the research question

The provided research question(s) should motivate the analysis conducted by the co-analysts. The research questions should be conveyed without specifying preferred analysis choices or expectations about the conclusions. Depending on the purpose of the project, the research questions can be more or less specific. While more specific research questions limit the analytical freedom of the co-analysts, less specific ones better explore the ways researchers can diverge in their operationalization of their question. A research question (e.g., "Is happiness age-dependent?") can be more specific when, for example, it is formulated as a directional hypothesis (e.g., "Are young people more happy than old ones?") or when the constructs are better operationalized (e.g., by defining what counts as young and happy).

Providing the task

The multi-analyst approach can leave the operationalization of the research question to the co-analysts so that they can translate the theoretical question into the measurement. Taking this approach can reveal the operational variations of a question, but it can also make it difficult to compare the statistical results.

Requesting results in terms of standardized metrics (e.g., *t*-values, standardized beta, Cohen's *d*) makes it easier to compare results between co-analysts. The requested metric can be determined from the aim of the analysis (e.g., hypothesis testing, parameter estimation). It needs to be borne in mind, however, that this request might bias the analysis strategies towards using methods that easily provide such a metric. [A practical tool for instructions on reporting effect estimates: (29).]

Co-analysts should be asked to keep a record of any code, derivatives etc. that were part of the analysis, at least until the manuscript is submitted and all relevant materials are (publicly) shared.

As an extension, the co-analysts can be asked to record considered but rejected analysis choices and the reasoning behind their choices (e.g., by commented code, log-books, or dedicated solutions such as DataExplained (24)). These logs can reflect where and why co-analysts diverge in their choices.

Robustness, or multiverse analyses (in the sense that each team is free to provide a series of outcomes instead of a single one) can also be part of the task of the co-analysts so that multiple analyses are conducted under alternative data analysis preprocessing choices.

Communication with co-analysts

In projects with many co-analysts, keeping contact via a dedicated email address and automating some of the messages (e.g., automated emails when teams finished a stage in the process) can help streamline the communication and make the process less prone to human errors. For co-analyst teams with multiple members, it can be helpful for each team to nominate one member as the representative for communications.

If further information is provided to a co-analyst following specific questions, it can be useful to make sure the same information is provided to all teams, for example via a Q&A section of the project website, hosting weekly office hours where participants could ask questions, or via periodic email with updates.

Conducting the independent analyses

Preregistering the process and statistical analyses

We can distinguish *meta-* and *specific preregistrations*. Meta-preregistrations concern the plan of the whole multi-analyst project. It is good practice for the lead team to preregister how they would process, handle, and report the results of the co-analysts in order to prevent result-driven biases. This can be done in the form of a Registered Report at journals that invite such submissions (30). Any metascientific questions, such as randomization of co-analysts to different conditions with variations in instructions or data, or covariates of interest for studying associations to analytic variability, should be specified.

Specific preregistrations concern the analysis plans of the co-analysts. Requiring co-analysts to prepare a specific preregistration for each analysis can be a strategy to prevent

overfitting and undisclosed flexibility. It makes sense to require it from either all or none of the teams in order to maintain equal treatment among them (unless the effect of preregistration is a focus of the study).

Requiring specific preregistrations may be misaligned with the goals of the project when the aim is to explore how the analytic choices are formed during the analyses, independent of initial plans. Under such circumstances, requiring specific preregistrations may be counterproductive. Nevertheless, the lead team can record their meta-preregistration that lays down the details of the multi-analyst project.

There are alternative solutions to prevent researchers from being biased by their data and results. For example, co-analysts could be provided with blinded datasets (14,16,31), simulated datasets (27), or with a subset of the data (e.g., 11).

Processing the results

Collecting the results

To facilitate summarizing the co-analysts' methods, results, and conclusions, the lead team can collect results through provided templates or survey forms that can structure analysts' reports. It is practical to ask the co-analysts at this stage to acknowledge that they did not communicate or cooperate with other co-analysts regarding the analysis in the project. It can also be helpful for the lead team if the co-analysts explain how their conclusions were derived from the results. In case preregistration was employed for any analyses, the template can also collect any deviations from the preregistered plan for inclusion in an online supplement.

To collect analytic code, it may be useful to require a container image (32,33) or a portable version of the code that handles issues like software package availability (34) (for a guideline see 35).

Validating the results

The lead team is recommended to ensure that each analyst's codes/procedures reproduce that analyst's submitted results. Computational reproducibility can be ascertained by running the code or repeating the analytic process by the lead team, but independent experts or the other co-analysts can also be invited to undertake this task (36,37).

The project can leverage the crowd by asking co-analysts to review others' analyses, or the lead team can employ external statistical experts to assess analyses and detect major errors. The lead team can decide to omit analyses with major errors. In that case, the reasons for omission should be documented, and for transparency, the results of the omitted analyses should be included in an online supplement.

After all the analyses have been submitted and validated, the co-analysts could have the option in certain projects to inspect the work of the other analysts and freely withdraw their own analyses. This can be appropriate if seeing other analyses makes them aware of major mistakes or shortcomings in their analytic procedures. A potential bias in this process is that co-analysts might lose confidence in their analyses after seeing other, more senior, or more expert co-analysts' work. One way to decrease this potential bias is to follow a multi-stage process: after the first round of analyses is submitted, co-analysts could be allowed to see each other's analysis steps/code without knowing the identity of the co-analyst or the

results of their analysis. It is the lead team's decision whether they allow co-analysts to correct or update their analyses after an external analyst or the co-analysts themselves find issues in their analyses.

Importantly, it is a minimum expectation that from the start of the project, the co-analysts should know about the conditions for their analyses to be included in, or omitted from, the study. All withdrawals, omissions, and updates of the results should be transparent in subsequent publications, for example in the supplementary materials.

Reporting the methods and results

Recording contributorship

Using CRediT taxonomy can transparently record organizers' and co-analysts' contributions to the study. Practical tools (e.g., tenzing 38) can make this task easier. Co-analysts can be invited to be co-authors and/or be compensated for their contribution in other ways (e.g., prizes, honorariums). Expectations for contribution and authorship should be communicated clearly at the outset.

Presenting the methods and results

Beyond a descriptive presentation of results in a table or graph, the reporting of the results of multi-analyst projects is not straightforward and remains an open area of research. Published reports of multi-analyst projects have adopted several effective methods for presenting results. For binary outcomes, Botvinik-Nezer et al. (39) used a table with color coding (i.e., a binary heat map) to visualize outcomes across all teams. They overlaid each teams' confidence in their findings and added additional information about analytical paths in adjacent columns (Supplementary Table 1). For a project with a relatively small number of effect sizes for continuous outcomes, Schweinsberg et al. (24) used interval plots combined with an indication of analytical choices underlying each estimate (Figure 3). Olsson Collentine et al. (40) (Figure 2) used funnel plots and Patel et al. (7) (Figures 1 and 2) used volcano plots to depict numerous, diverse outcomes with an intuitive depiction of clustering (akin to a multiverse analysis).

If the main purpose is to estimate variability of analyses, it is interesting to investigate and report factors that might influence variability in the chosen analytic approaches and in the results obtained by these analytical approaches. If, on the other hand, the main purpose is to investigate the robustness of conclusions by assessing the degree to which different analysts obtain the same results, it is advisable to focus more on methods that produce only a single answer to the research question of interest. When each analysis team can provide multiple, distinct responses to the same research question, it becomes more difficult to explore how conclusions depend on the analysis choices because the individual analyses are no longer independent of each other.

The analytical approach of each co-analyst can be divided into discrete choices concerning, for instance, data preprocessing steps and decisions in model specification. If it is possible to recombine the individual choices (which will not always be the case as certain data preprocessing steps or method choices may only make sense if the aim is to fit a certain class of models), it may be worthwhile to create a larger set of possible analytical approaches that is made up of all possible combinations. In this case, the descriptive results of the multi-

analyst project can be combined with a multiverse type approach (e.g., vibration of effects 7, multiverse analysis 8, or specification curve 41) to quantify and compare the variability in results that can be explained by the different analytical choices (7,42). Additionally, this larger set of possible combinations can be helpful to present the results in an interactive user interface in which readers can explore how the results change as a function of certain analytical choices (42,43). Finally, dividing the co-analysts' analytical approaches into individual choices may ultimately help in providing a unique answer to the research question of interest while accounting for the uncertainty in the choice of the analytical approach. While there are so far no approaches that would allow the derivation of a unique result that integrates all uncertain decisions, it may be a promising area of research to extend Bayesian approaches that account for model uncertainty (44) and measurement error (45).

To support the reporting of Multi-Analyst projects, we provide a freely modifiable *Reporting Template* available from here: https://osf.io/h9mgy/

Limitations

The present work does not cover all aspects of multi-analyst projects. For instance, the multi-analyst approach outlined here entails the independent analysis of one or more datasets, but it should be acknowledged that other crowd-sourced analysis approaches might not require such independence of the analyses. Some of our practical considerations reflect disagreement and/or uncertainty within our expert panel, so they remain underspecified. Those include how to determine the number or eligibility of co-analysts for a project, how best to assess the validity of each analysis; and how to measure robustness of conclusions. Therefore, we emphasize that this consensus-based guidance is a first step towards the broader adoption of the multi-analyst approach in empirical research, and we hope and expect that our recommendations will be developed further in response to user feedback. Users of this guidance can provide feedback and suggestions for revisions at https://forms.gle/2fVqZAD3KKHVUDKq7.

Conclusions

This guidance document aims to facilitate adoption of the multi-analyst approach in both basic and clinical research. Although the multi-analyst approach is at an incipient stage of adoption, we believe that the scientific benefits greatly outweigh the extra logistics required, especially for projects with high relevance for clinical practice and policy making. The approach should have particular relevance when it indicates that applying different analysis strategies to a given dataset may lead to conflicting results. The multi-analyst approach allows a systematic exploration of the analytical space to assess whether the reported results and conclusions are dependent on the chosen analysis strategy, ultimately improving the transparency, reliability, and credibility of research findings.

We hope that our guidance here and in guideline databases will make it easier for researchers to adopt this approach to empirical analyses. We encourage journals and funders to consider recommending or requesting independent analyses whenever it is crucial to know whether the conclusions are robust to alternative analysis strategies.

Acknowledgements: This research was not funded. AS was supported by a talent grant from the Netherlands Organisation for Scientific Research (NWO) to AS (406-17-568). RB-N is an Awardee of the Weizmann Institute of Science – Israel National Postdoctoral Award Program for Advancing Women in Science. BAN was supported by grants from the John Templeton Foundation, Templeton World Charity Foundation, Templeton Religion Trust, and Arnold Ventures. SSt-J is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) [funding reference number BP–546283–2020] and the Fonds de recherche du Québec - Nature et technologies (FRQNT) [Dossier 290978]. JMW and ORvdA were supported by a Consolidator Grant (IMPROVE) from the European Research Council (ERC; grant no. 726361). YKK was supported by a grant from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (ERC-CoG-2015; No 681466 to M Wichers). DvR was supported by a Dutch scientific organization VIDI fellowship grant (016.Vidi.188.001). LFB was supported by a Dutch scientific organization VENI fellowship grant (Veni 191G.037). MJS was supported by the US National Science Foundation (1760052). ELU was supported by an R&D grant from INSEAD.

Author contributors: BA and BS are joined first authors and guarantors. BA, BS, GN, and E-JW were responsible for the study conception and design. ORvdA, CJA, MALMvA, JAB, DB, UB, RB-N, LFB, NB, EC, AMC, NC, A Delios, NNNvD, CD, JBvD, A Dreber, GD, GFE, MAG, RH, SH, FH, JH, MJ, KJJ, ATK, MK, YKK, DSL, J-FM, DM, MRM, BRN, BAN, RAP, DvR, JR, MJS, AS, TS, MS, DS, RS, DJS, BAS, SSt-J, JJS, ELU, and JW served as expert panel in the development of the guidance and checklist. All authors participated in drafting and critically revising the manuscript. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Competing interests: We have read the journal's policy and the authors of this manuscript have the following competing interests: BAN is Executive Director of the Center for Open Science, a non-profit technology and culture change organization with a mission to increase openness, integrity, and reproducibility of research. The other authors declare no competing interest.

Data and materials availability: All anonymized data as well as the survey materials are publicly shared on the Open Science Framework page of the project: https://osf.io/4zvst/. Our methodology and data-analysis plan were preregistered. The preregistration document can be accessed at: https://osf.io/dgrua.

Transparency declaration: The lead author affirms that the manuscript is an honest, accurate, and transparent account of the work being reported; that no important aspects of the study have been omitted; and that any discrepancies from the work as planned have been explained.

Additional files

Supplementary file 1

Reporting template for multi-analyst studies

Supplementary file 2
Reporting checklist for multi-analyst studies

References

1. Fields AC, Lu P, Palenzuela DL, Bleday R, Goldberg JE, Irani J, et al. Does retrieval bag use during laparoscopic appendectomy reduce postoperative infection? Surgery. 2019;165(5):953–7.
2. Turner SA, Jung HS, Scarborough JE. Utilization of a specimen retrieval bag during laparoscopic appendectomy for both uncomplicated and complicated appendicitis is not associated with a decrease in postoperative surgical site infection rates. Surgery. 2019;165(6):1199–202.
3. Childers CP, Maggard-Gibbons M. Same Data, Opposite Results?: A Call to Improve Surgical Database Research. JAMA Surg. 2021;156(3):219–20.
4. de Vries M, Witteman CLM, Holland RW, Dijksterhuis A. The Unconscious Thought Effect in Clinical Decision Making: An Example in Diagnosis. Med Decis Making. 2010;30(5):578.
5. Jivanji D, Mangosing M, Mahoney SP, Castro G, Zevallos J, Lozano J. Association Between Marijuana Use and Cardiovascular Disease in US Adults. Cureus. 2020 12:e11868(12).
6. Shah S, Patel S, Paulraj S, Chaudhuri D. Association of Marijuana Use and Cardiovascular Disease: A Behavioral Risk Factor Surveillance System Data Analysis of 133,706 US Adults. Am J Med. 2021 May 1;134(5):614-620.e1.
7. Patel CJ, Burford B, Ioannidis JP. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. J Clin Epidemiol. 2015;68(9):1046–58.
8. Steegen S, Tuerlinckx F, Gelman A, Vanpaemel W. Increasing Transparency Through a Multiverse Analysis. Perspect Psychol Sci. 2016 Sep 1;11(5):702–12.
9. Bastiaansen JA, Kunkels YK, Blaauw FJ, Boker SM, Ceulemans E, Chen M, et al. Time to get personal? The impact of researchers choices on the selection of treatment targets using the experience sampling methodology. J Psychosom Res. 2020 Oct 1;137:110211.
10. Dongen NNN van, Doorn JB van, Gronau QF, Ravenzwaaij D van, Hoekstra R, Haucke MN, et al. Multiple Perspectives on Inference for Two Simple Statistical Scenarios. Am Stat. 2019 Mar 29;73(sup1):328–39.
11. Salganik MJ, Lundberg I, Kindel AT, Ahearn CE, Al-Ghoneim K, Almaatouq A, et al. Measuring the predictability of life outcomes with a scientific mass collaboration. Proc Natl Acad Sci. 2020 Apr 14;117(15):8398–403.
12. Silberzahn R, Uhlmann EL, Martin DP, Anselmi P, Aust F, Awtrey E, et al. Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. Adv Methods Pract Psychol Sci. 2018 Sep 1;1(3):337–56.

13. Botvinik-Nezer R, Holzmeister F, Camerer CF, Dreber A, Huber J, Johannesson M, et al. Variability in the analysis of a single neuroimaging dataset by many teams. Nature. 2020 Jun;582(7810):84–8.

14. Dutilh G, Annis J, Brown SD, Cassey P, Evans NJ, Grasman RPPP, et al. The Quality of Response Time Data Inference: A Blinded, Collaborative Assessment of the Validity of Cognitive Models. Psychon Bull Rev. 2019 Aug 1;26(4):1051–69.

15. Fillard P, Descoteaux M, Goh A, Gouttard S, Jeurissen B, Malcolm J, et al. Quantitative evaluation of 10 tractography algorithms on a realistic diffusion MR phantom. NeuroImage. 2011 May 1;56(1):220–34.

16. Starns JJ, Cataldo AM, Rotello CM, Annis J, Aschenbrenner A, Bröder A, et al. Assessing theoretical conclusions with blinded inference to investigate a potential inference crisis. Adv Methods Pract Psychol Sci. 2019;2(4):335–49.

17. Maier-Hein KH, Neher PF, Houde J-C, Côté M-A, Garyfallidis E, Zhong J, et al. The challenge of mapping the human connectome based on diffusion tractography. Nat Commun. 2017;8:1349

18. Poline J-B, Strother SC, Dehaene-Lambertz G, Egan GF, Lancaster JL. Motivation and synthesis of the FIAC experiment: Reproducibility of fMRI results across expert analyses. Hum Brain Mapp. 2006;27(5):351–9.

19. Fox Talbot WH, Hincks E, Oppert J, Rawlinson HC. 1861. Comparative translations of the inscription of Tiglath Pileser I. *Journal of the Royal Asiatic Society of Great Britain & Ireland*. **18**:150–219. DOI: https://doi.org/10.1017/S0035869X00013666

20. McKenna HP. The Delphi technique: a worthwhile research approach for nursing? J Adv Nurs. 1994 Jun 1;19(6):1221–5.

21. Jünger S, Payne SA, Brine J, Radbruch L, Brearley SG. Guidance on Conducting and REporting DElphi Studies (CREDES) in palliative care: Recommendations based on a methodological systematic review. Palliat Med. 2017;31(8):684–706.

22. Aczel B, Szaszi B, Sarafoglou A, Kekecs Z, Kucharský Š, Benjamin D, et al. A consensus-based transparency checklist. Nat Hum Behav. 2020 Jan;4(1):4–6.

23. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3:160018.

24. Schweinsberg M, Feldman M, Staub N, van den Akker OR, van Aert RCM, van Assen MALM, et al. Same data, different conclusions: Radical dispersion in empirical results when independent analysts operationalize and test the same hypothesis. Organ Behav Hum Decis Process. 2021 Jul 1;165:228–49.

25. Lundberg I, Narayanan A, Levy K, Salganik MJ. Privacy, Ethics, and Data Access: A Case Study of the Fragile Families Challenge. Socius. 2019 Jan 1;5:2378023118813023.

26. Drechsler J. Synthetic datasets for statistical disclosure control: theory and implementation. Vol. 201. Springer Science & Business Media; 2011.

27. Quintana DS. A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation.. eLife. 2020;9:e53275.

28. Kindel AT, Bansal V, Catena KD, Hartshorne TH, Jaeger K, Koffman D, et al. Improving metadata infrastructure for complex surveys: Insights from the Fragile Families Challenge. Socius. 2019;5:2378023118817378.

29. Parker TH, Fraser H, Nakagawa S, Fidler F, Gould E, Gould E, et al. Evolutionary Ecology Data. 2020 Mar 18 [cited 2021 Sep 28]; Available from: https://osf.io/34fzc/

30. Chambers CD. Registered reports: a new publishing initiative at Cortex. Cortex. 2013;49(3):609–10.

31. Gøtzsche PC. Blinding during data analysis and writing of manuscripts. Control Clin Trials. 1996 Aug;17(4):285–90; discussion 290-293.

32. Boettiger C. An introduction to Docker for reproducible research. ACM SIGOPS Oper Syst Rev. 2015 Jan 20;49(1):71–9.

33. Nüst D, Sochat V, Marwick B, Eglen SJ, Head T, Hirst T, et al. Ten simple rules for writing Dockerfiles for reproducible data science. PLOS Comput Biol. 2020 Nov 10;16(11):e1008316.

34. Liu DM, Salganik MJ. Successes and Struggles with Computational Reproducibility: Lessons from the Fragile Families Challenge. Socius. 2019 Jan 1;5:2378023119849803.

35. Elmenreich W, Moll P, Theuermann S, Lux M. Making simulation results reproducible—Survey, guidelines, and examples based on Gradle and Docker. PeerJ Comput Sci. 2019 Dec 9;5:e240.

36. Hurlin C, Pérignon C. Reproducibility Certification in Economics Research. HEC Paris Research Paper No. FIN-2019-1345. 2019 Jul 12.

37. Pérignon C, Gadouche K, Hurlin C, Silberman R, Debonnel E. Certify reproducibility with confidential data. Science. 2019 Jul 12;365(6449):127–8.

38. Holcombe AO, Kovacs M, Aust F, Aczel B. Documenting contributions to scholarly articles using CRediT and tenzing. PLOS ONE. 2020 Dec 31;15(12):e0244611.

39. Botvinik-Nezer R, Holzmeister F, Camerer CF, Dreber A, Huber J, Johannesson M, et al. Variability in the analysis of a single neuroimaging dataset by many teams [Internet]. 2019 Nov [cited 2021 Sep 27] p. 843193. Available from: https://www.biorxiv.org/content/10.1101/843193v1

40. Olsson-Collentine A, Aert RCM van, Bakker M, Wicherts J. Preprint - Meta-Analyzing the Multiverse: A Peek Under the Hood of Selective Reporting [Internet]. PsyArXiv; 2021 [cited 2021 Sep 27]. Available from: https://psyarxiv.com/43yae/

41. Simonsohn U, Simmons JP, Nelson LD. Specification curve analysis. Nat Hum Behav. 2020;4(11):1208–14.

42. Liu Y, Kale A, Althoff T, Heer J. Boba: Authoring and visualizing multiverse analyses. IEEE Trans Vis Comput Graph. 2020;27(2):1753–63.

43. Dragicevic P, Jansen Y, Sarma A, Kay M, Chevalier F. Increasing the transparency of research papers with explorable multiverse analyses. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 2019. p. 1–15.

44. Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: a tutorial (with comments by M. Clyde, David Draper and EI George, and a rejoinder by the authors. Stat Sci. 1999;14(4):382–417.

45. Richardson S, Gilks WR. A Bayesian approach to measurement error problems in epidemiology using conditional independence models. Am J Epidemiol. 1993;138(6):430–42.