1 Reducing bias in secondary data analysis via an Explore and Confirm Analysis Workflow

2 (ECAW): A proposal and survey of observational researchers

3 Robert T. Thibault[1,2], Marton Kovacs[3,4], Tom E. Hardwicke[5], Alexandra Sarafoglou[6],

4 John P. A. Ioannidis[1,7], & Marcus R. Munafò[2,8]

5 [1] Meta-Research Innovation Center at Stanford (METRICS), Stanford University.

6 [2] School of Psychological Science, University of Bristol.

7 [3] Doctoral School of Psychology, ELTE Eotvos Lorand University, Budapest, Hungary

8 [4] Institute of Psychology, ELTE Eotvos Lorand University, Budapest, Hungary

9 [5] Melbourne School of Psychological Sciences, University of Melbourne.

10 [6] Department of Psychology, University of Amsterdam.

11 [7] Meta-Research Innovation Center Berlin (METRIC-B), QUEST Center for Transforming

12 Biomedical Research, Berlin Institute of Health, Charité – Universitätsmedizin Berlin.

13 [8] MRC Integrative Epidemiology Unit at the University of Bristol.

14 Author Note

15 Correspondence concerning this article should be addressed to Robert T. Thibault,

16 Enter postal address here. E-mail: robert.thibault@stanford.edu

17                                              Abstract

18  **Background.** Although preregistration can reduce researcher bias and increase

19  transparency in primary research settings, it is less applicable to secondary data analysis.

20  An alternative method that affords additional protection from researcher bias, which

21  cannot be gained from conventional forms of preregistration alone, is an Explore and

22  Confirm Analysis Workflow (ECAW). In this workflow, a data management organization

23  initially provides access to only a subset of their dataset to researchers who request it. The

24  researchers then prepare an analysis script based on the subset of data, upload the analysis

25  script to a registry, and then receive access to the full dataset. ECAWs aim to achieve

26  similar goals as preregistration, but make access to the full dataset contingent on

27  compliance. The present survey aimed to garner information from the research community

28  where ECAWs could be applied—employing the Avon Longitudinal Study of Parents and

29  Children (ALSPAC) as a case example.

30

31  **Methods.** We emailed a web-based survey to researchers who had previously applied for

32  access to ALSPAC's transgenerational observational dataset.

33

34  **Results**. We received 107 responses, for a 10% response rate. The results suggest that—at

35  least among our sample of respondents—ECAWs hold the potential to serve their intended

36  purpose and appear relatively acceptable. For example, only 10% of respondents disagreed

37  that ALSPAC should run a study on ECAWs (versus 55% who agreed). However, as many

38  as 25% of respondents agreed that they would be less willing to use ALSPAC data if they

39  were required to use an ECAW (versus 44% who disagreed).

40

41  **Conclusion.** Our data and findings provide information for organizations and individuals

42  interested in implementing ECAWs and related interventions.

43

<sup>44</sup> **Preregistration:** https://osf.io/g2fw5 Deviations from the preregistration are outlined in

<sup>45</sup> Supplementary Material A.

<sup>46</sup> *Keywords:* Blind data analysis; preregistration; ALSPAC; meta-research; open

<sup>47</sup> science; Explore and Confirm Analysis Workflow (ECAW)

<sup>48</sup> Word count: 3546

Reducing bias in secondary data analysis via an Explore and Confirm Analysis Workflow

(ECAW): A proposal and survey of observational researchers

## Introduction

Many published research findings are non-reproducible and potentially false or misleading (Camerer et al., 2016; Collaboration, 2015; Errington et al., 2021; Ioannidis, 2005, 2008). These shortcomings may stem from a combination of researcher bias, publication bias, selective reporting of results, incomplete reporting of methods, and other questionable research practices. These issues can lead to research waste, and ineffective or even harmful healthcare and policy interventions (e.g., see Prasad and Cifu (2019)).

Some research disciplines have adopted standards to address these issues. For example, researchers conducting clinical trials regularly register outcome measures before enrolling participants, and they keep both participants and experimenters blind regarding group assignment. In cases where datasets already exist, however, researchers cannot easily adopt these practices. Researchers performing secondary data analyses generally have access to complete datasets without any blinding imposed on them. This may lead to exploration of diverse analyses and selective reporting based on the nature of the results.

### Preregistration in the context of secondary data analyses

Increasing the uptake and effectiveness of preregistration for secondary data analyses may require a different approach than it does for clinical trials. For example, the analytical space for secondary data analyses (e.g., of a longitudinal cohort dataset) is more vast than the analytical space for most clinical trials. Thus, the clinical trial standard of registering outcome measures, but no analysis plan, remains insufficient. Moreover, preregistration of secondary data analyses remains uncommon, and unlike for clinical trials (where the dataset doesn't yet exist), there is no guarantee that researchers performing secondary data analyses haven't accessed the dataset. To extend the benefits of preregistration to

74 secondary data analyses, we propose a workflow that necessitates the preregistration of an

75 executable analysis script before researchers can access the full dataset.

76 **Explore and Confirm Analysis Workflow (ECAW)**

77     In this intervention we propose, a data management organization that controls access

78 to a preexisting data set would first provide access to only a subset of the data a researcher

79 requests. After the researcher prepares an analysis script based on the subset of data and

80 uploads it to a registry—where it will be openly accessible and permanent (e.g., to the

81 Open Science Framework Registry)—the data management organization then provides

82 access to the full dataset (e.g., via a secure data environment) and the researcher can

83 proceed as they wish. The exact implementation (e.g., whether the data management

84 organization runs quality checks on the analysis scripts) would depend on the preferences

85 of the data management organization and the research community using their dataset (see

86 Box 1 for an hypothetical example).

87     We call this process an Explore and Confirm Analysis Workflow (ECAW).

88 Researchers may use the subset of data to generate hypotheses and/or to simply ensure

89 that their intended analysis runs properly. ECAWs may increase the uptake of

90 preregistration and provide assurance that the analyses were developed before the

91 researchers observed the full dataset. Moreover, whereas conventional preregistrations often

92 leave many degrees of freedom regarding exactly how the data will be analyzed (e.g.,

93 Bakker et al. (2020)), ECAWs provide an executable analysis script. Researchers may also

94 find ECAWs more agreeable than conventional preregistration because they allow for

95 exploration. ECAWs would not fully solve publication bias, but could make it easier to

96 detect because the registered analysis script would serve as a form of preregistration.

97     In short, ECAWs give researchers access to enough data to develop an informed

98 analysis, but without compromising the confirmatory nature of a final analysis on the

99 complete dataset.

## Box 1. A hypothetical example of an ECAW in practice.
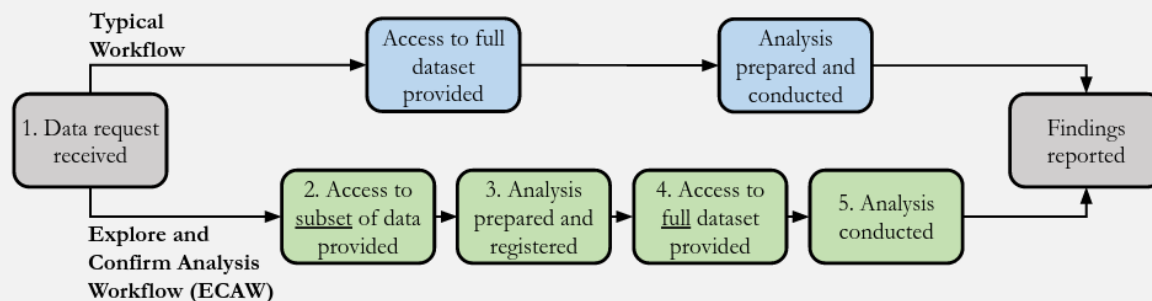


The overarching framework for ECAWs is depicted above. Within this framework, data management organizations would need to pre-specify several details for each step. Below, we provide a hypothetical example for illustrative purposes. The exact implementation of ECAWs would depend on the preferences of each data management organization and the community who uses their dataset.

1. A research team submits to a data management organization (i) a paragraph describing the analyses they want to run, and (ii) a list of the variables they want to analyze.

2. The data management organization provides the research team with access to a subset of the data. For example, if the research team requests data for 15 variables collected from 10,000 participants, the data management organization will provide access to all 15 variables from a random subset of participants—say 1000.

3. The research team prepares an analysis script written in a common programming language (e.g., R or Stata) using the subset of data. They register this analysis script, the output from the script, and the paragraph they sent to the data management organization in Step 1, to www.osf.io/registries as an 'Open-Ended Registration'. The researchers are free to run as many analyses as they would like on the subset of data. However, they should only register the analysis script they plan to use on the complete dataset.

4. The data management organization performs a basic quality check. They ensure that (i) the analysis script, the script output, and the paragraph are registered, and (ii) the output contains a clear number of itemized results (e.g., as clinicaltrials.gov does for outcome measures). If the quality check fails, the data management organization asks the researchers to update their registration until it passes. The research team is then given access to a dataset with all the variables requested for all participants.

5. The researchers can proceed as they wish. They can run the registered analysis script on the complete dataset, make adjustments to the analysis if desired, or not proceed at all. The data management organization does not perform any additional check on what analyses were run. Regardless of what the research team decides, a permanent version of the planned analysis is available on the OSF Registry.

**Examples of ECAW-like workflows**

One study probed the benefits and drawbacks of a workflow similar to ECAWs, but they provided a synthetic dataset rather than a data subset (Sarafoglou, Hoogeveen, and Wagenmakers (2023)). The researchers recruited 120 teams to analyze a single observational dataset and had half the teams preregister a written analysis plan and the other half prepare an analysis plan by drafting an analysis script based on a dataset with shuffled data for the variables of interest (i.e., the analysts were effectively blinded). Based on self-reports from the participants, the researchers found that two workflows were comparable in terms of effort and that teams using blinded data analysis made fewer deviations from their analysis plan.

Some disciplines, such as particle physics, implement blind data analyses regularly (Klein and Roodman (2005); MacCoun and Perlmutter (2015)). A version of the ECAW workflow has also been successfully implemented by eight teams performing secondary data analysis on a dataset managed by the Psychological Science Accelerator group (Forscher et al. (2020)) and is currently being used for Registered Reports based on a large psychology dataset (Schmidt et al. (2023)).

In medicine and clinical epidemiology, a similar workflow is implemented by the software platform OpenSAFELY (www.opensafely.org). This platform provides a dataset of simulated health records from which researchers can develop an analysis script. When ready, a researcher submits their analysis script which is automatically logged and made public in GitHub. The analysis then runs in a Trusted Research Environment (TRE) on data which is stored in the data centers where patients' records already reside (i.e., the data is not copied or moved). This workflow keeps individual health records hidden while also documenting all analyses run on the real data.

**Study objectives**

Here, we present a descriptive and exploratory survey study. We had no hypotheses, but we did have two specific objectives. (1) To gain insights on the opinions and practices of researchers who already use preexisting observational datasets, in regards to the trustworthiness and reproducibility of research. (2) To use these insights to inform future research—including a potential trial of ECAWs with the Avon Longitudinal Study of Parents and Children (ALSPAC)—on how data management organizations can encourage rigorous and reproducible research practices.

# Methods

**Participants**

We sent an email to invite researchers on the mailing list for the UK-based Avon Longitudinal Study of Parents and Children (ALSPAC) to participate in an online survey (see Supplementary Material B). We partnered with ALSPAC because they manage an oft-requested dataset and expressed interest in studying ways to ensure the research stemming from their dataset is rigorous. ALSPAC is "a transgenerational prospective observational study investigating influences on health and development across the life course. It considers multiple genetic, epigenetic, biological, psychological, social and other environmental exposures in relation to a similarly diverse range of health, social and developmental outcomes." (Boyd et al., 2013; Fraser et al., 2013). Thus, our invitation reached researchers that use observational data across the health and social sciences. The survey was open from 10 Oct 2022 to 1 Nov 2022. We sent two reminder emails, exactly one week and two weeks after the original email invitation. The mailing list comprises researchers whose email addresses were present on a proposal to access the ALSPAC dataset and included 1148 email addresses.

**Survey**

The survey contained 6 blocks and is available at https://osf.io/5h7gb. We developed the survey with feedback from the Principal Investigator and Executive Director of ALSPAC (Nicholas Timpson and Kate Northstone). We aimed to include as few questions as possible (to encourage a high response rate) while still garnering substantive information on whether ECAWs are relevant and acceptable to ALSPAC users.

Block 1 assessed the extent to which respondents believe that observational research using preexisting data is trustworthy and reproducible (2 questions). Block 2 asked respondents how often they use practices related to transparency and reducing researcher bias, including preregistration, blinded data analysis, and sharing analysis scripts (5 questions). Between Block 2 and Block 3, the survey described ECAWs. Block 3 assessed the extent to which respondents believe that ECAWs would make observational research using preexisting data more trustworthy and reproducible (2 questions). Block 4 directly asked respondents whether ALSPAC should run a study on ECAWs and whether they would participate (5 questions). Block 5 contained open-ended questions about the perceived benefits and drawbacks of ECAWs, reservations about ECAWs, whether additional incentives might be needed to use ECAW, and suggestions for other research practices or policies that data management organizations could implement to improve research quality (3 questions). Block 6 asked participants how concerned they are about research quality, how many relevant studies they have published, what software they use for data analysis, and for any additional comments (4 questions). All rating-scale questions contained the response option *"I don't understand the question"*. The questions presented in Figure 3 also contained the response option *"Unsure"*.

172 **Analyses**

173 We present the results for closed-ended questions as prevalence rates or counts in
174 Figures 1-3 and Supplementary Material C. When reporting percentages, we collapse
175 together all positive responses on the rating scales (e.g., *"strongly agree"* and *"somewhat*
176 *agree"*) as well as all negative responses. The total number of responses differ among
177 questions due to the missing values and responses of *"I don't understand the question"* and
178 *"Unsure"*. We narratively synthesize the responses to the open-ended questions in Table 1
179 (i.e., we provide a non-systematic qualitative summary).

180 <div align="center">**Results**</div>

181 **Survey completion**

182 Of 1148 emails sent, 1094 went through and 54 bounced. The survey was completed
183 107 times and partially completed 21 times, leading to a response rate of 10% for complete
184 surveys and 2% for incomplete surveys.[1] The median time taken for complete survey
185 responses was 7.20 minutes (IQR: 4.30 to 13.10). This manuscript presents the results for
186 complete surveys. Supplementary Material D presents the results with partially complete
187 surveys included.

---

[1] The ALSPAC mailing list has been maintained as a record of collaborators for many years and is
constantly added to. The ALSPAC team only removes email addresses from their mailing list if someone
explicitly requests this action. Thus, their list likely includes several email addresses that are no longer
monitored. For example, we received one email reply stating that the recipient hasn't been active in
research for 30 years. It is also possible that some researchers have multiple email addresses on the mailing
list (e.g., because they moved institutions). These two factors may have deflated the response rate.

**Participants**

Respondents had published a median of NA (IQR 2 to 25) studies using preexisting observational data (Supplementary Figure C1). They reported using the following programming languages or software packages: R (n = 65), Stata (n = 48), SPSS (n = 17), SAS (n = 15), Python (n = 6), Mplus (n = 3), Bash (n = 2), MATLAB (n = 1), Nextflow (n = 1), and plink2 (n = 1) (Supplementary Table C1)[2]. 61% (62/101) of participants reported being more concerned with research trustworthiness, bias, rigour, and reproducibility compared to what they think of as a typical research who uses preexisting observational data; 6% (6/101) reported being less concerned (Supplementary Figure C2).

**Survey results**

Most respondents agreed that studies that analyze preexisting observational datasets are trustworthy (72%; 77/107) and reproducible (79%; 84/107) (Figure 2, top panel). At the same time, many agreed that a study using an ECAW would be *more* trustworthy (70%; 71/101) and *more* reproducible (69%; 70/102) compared to a typical study using preexisting observational data (Figure 1, bottom panel).

Over half of respondents reported that their studies using preexisting observational data are preregistered never or almost never (35%; 37/107), or sometimes (24%; 26/107) (Figure 2). About half reported sharing their analysis scripts never or almost never (20%; 21/107), or sometimes (31%; 33/107). 75% (80/107) reported that they never or almost never blind the data analyst. Almost all respondents answered that they use both exploratory (91%; 98/107) and confirmatory (85%; 92/107) analyses at least sometimes.

25% (26/102) of respondents agreed (versus 44%; 45/102 who disagreed) that they would be less willing to use ALSPAC data if they were required to use an ECAW (Figure 3). 53% (50/95) agreed (20%; 19/95 disagreed) that they would opt-in if ALSPAC ran a

---

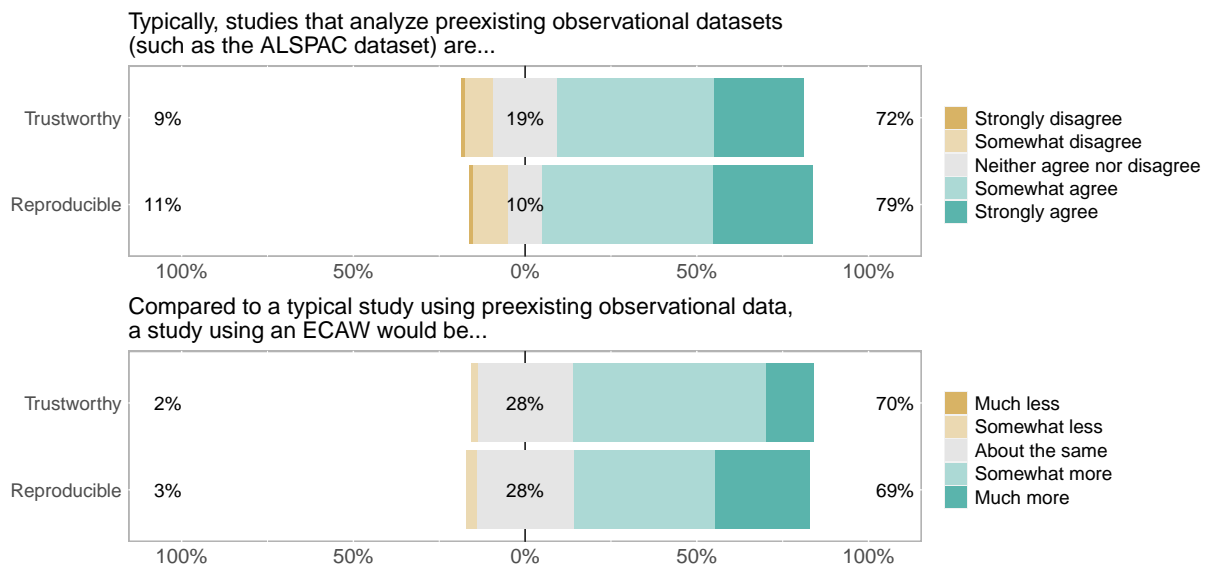[2] Participants could select multiple responses to this survey question.

**Figure 1. Responses to the survey questions on trustworthiness and reproducibility of observational research with preexisting data and ECAWs.** The survey defined trustworthy as "meaning that the results and conclusions of the publications are valid, reliable, rigorous, and accurate. That they merit trust". The survey defined reproducible "in the sense that other researchers re-analysing the data with the same research question would produce similar results." For each item, the number to the left of the data bar indicates the combined percentage for the responses depicted in any shade of brown/orange. The number in the center of the data bar (gray) indicates the percentage of neutral responses. The number to the right of the data bar indicates the combined percentage for the responses depicted in any shade of green. The bar charts in the top panel had no missing responses or selection of the option *"I don't understand the question"*. The bar charts in the bottom panel excluded missing responses (n = 2; 2) and responses of *"I don't understand the question"* (n = 4; 3).

The studies using preexisting observational data that I am involved in...
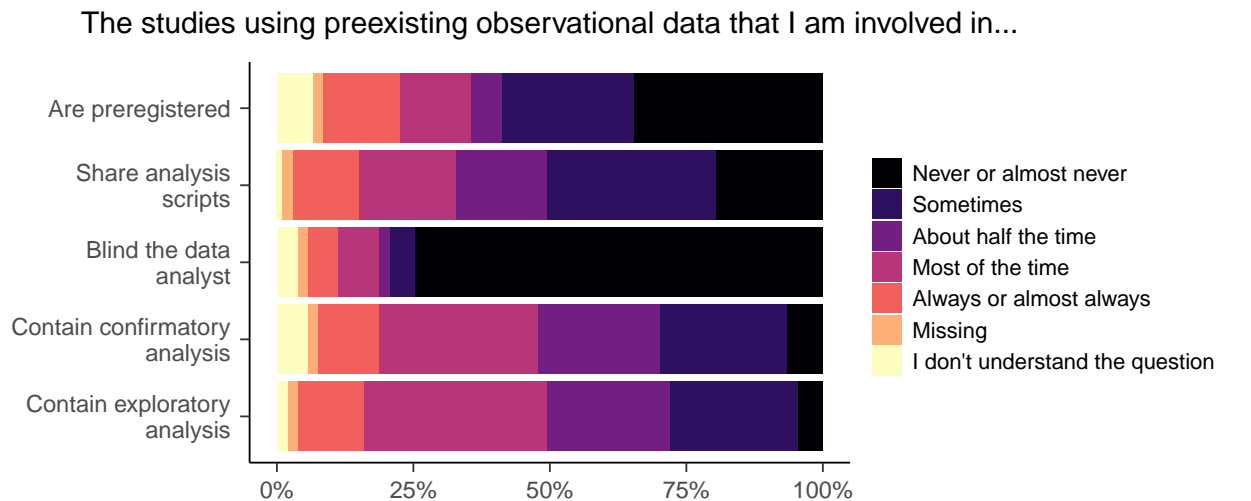


**Figure 2. Responses to survey questions about the research practices of participants.**

study on ECAWs. 55% (53/97) agreed (10%; 10/97 disagreed) that ALSPAC should run a study on ECAWs. 45% (43/95) agreed (22%; 21/95 disagreed) that they would prefer using an ECAW than using typical preregistration.

**Table 1. Recurring topics in responses to the open-ended survey questions.** The survey included 4 open-ended questions with broad prompts regarding running a study on ECAWs, benefits and drawbacks of ECAWs, related research practices, and general comments. These questions received a total of (92) responses from (55) unique respondents. A complete list of responses are viewable in the open data [LINK]. We synthesized the responses to open-ended questions into the 9 topics on the left side of this table. We divide these into three sections: (i) concerns about the acceptability of ECAWs,

---

[3] Another four respondents agreed with at least the first and second question, which appear contradictory. We did not preregister these considerations. More careful wording of these questions could have circumvented the ambiguity in interpreting these seemingly contradictory responses.

[4] We intended for these questions to ask about ALSPAC running a trial on ECAWs. However, due to the ambiguity around the terms "study", some respondents may have interpreted this as a survey, focus group, feasibility, or pilot type of study.
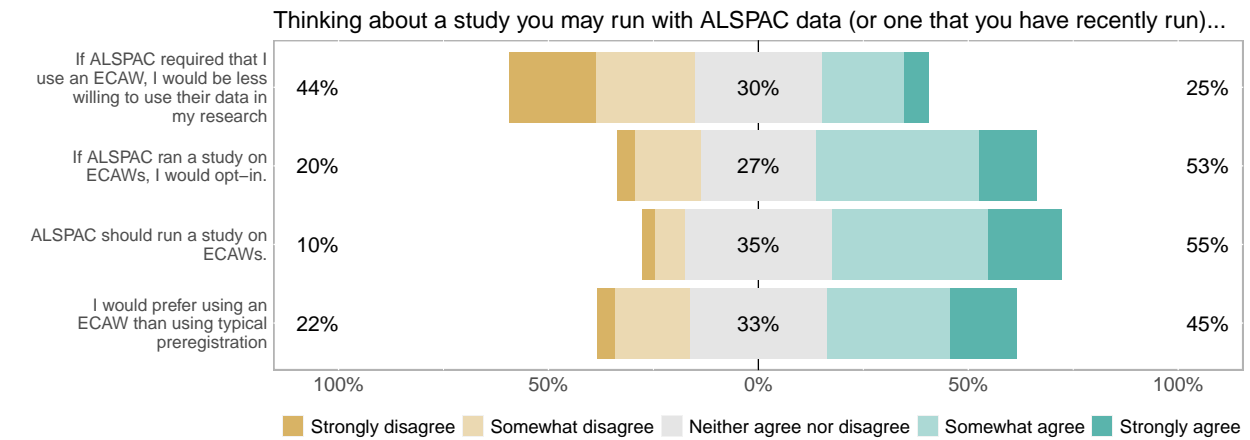
**Figure 3. Responses to survey questions about using ECAWs.** These bar charts exclude missing values ($n = 3$; 3; 3; 3; respectively from top to bottom), responses of *"I don't understand the question"* ($n = 0$; 4; 1; 1), and responses of *"Unsure"* ($n = 2$; 5; 6; 8). Agreement with the first question may be slightly inflated due to the format of the questions in this block. Respondents with a highly positive inclination towards ECAWs would be expected to disagree with the first question, but agree with the next three questions. Four respondents agreed with all four statements, suggesting they may have glazed over the word "less" in the first question [3]. Interpreting responses to the second and third question come with a degree of ambiguity as the survey did not specify what was meant by the term "study" [4].

(ii) concerns that ECAWs will not have their intended impact, and (iii) alternative interventions that may achieve similar goals as conventional preregistration and ECAWs. On the right side of the table, we provide our reflection on each topic.

**Exploratory analyses**

Results can be explored interactively at [TEMPORARY LINK], or by running the code available at [INSERT DATA LINK]. Based on the sample size of 107 participants and a lack of visually striking differences when exploring subsets of respondents, we do not report further on exploratory analyses.

| Respondents' comments[1] | Our reflection |
| --- | --- |
| **Concerns about acceptability** | |
| 1. ECAWs would take too much time. | Compared to a study that is not preregistered, a study using an ECAW could very well take more time—especially at the beginning of the research process when preparing an analysis script. However, the advantages may outweigh the time commitment and further research could help understand this potential tradeoff. One survey suggests that psychology researchers believe that preregistration leads to longer project duration, but also higher research quality (Sarafoglou et al., 2022). Moreover, compared to conventional preregistration, some evidence suggests that an ECAW-like workflow does not take more time (Sarafoglou et al., 2023).[2] |
| 2. The subset of data would be insufficient for certain applications (e.g., for genome-wide association studies, rare outcomes, questions regarding small sub-samples, and imputation). | Our survey suggested providing access to 10% of the dataset in the first ECAW stage. To develop an analysis for certain research questions, this subset of data may be insufficient. In these cases, a larger subset could be provided (e.g., 50%) and the final analysis could be run on the other 50% of the data, rather than the full dataset. Alternatively, the data management organization could create and provide access to a full-sized synthetic dataset. |
| 3. Researchers will often have to change their analysis after seeing the full dataset. | ECAWs do not stop researchers from changing their analysis. ECAWs make these changes transparent. This allows others to better assess risk of bias and calibrate their confidence in the study findings. As compared to conventional preregistration, some evidence suggests that an ECAW-like workflow leads to fewer deviations from the analysis plan (Sarafoglou et al., 2023). |
| 4. Researchers sometimes reuse the same dataset which they will already have full access to. | ECAWs do not stop researchers from reusing the same dataset for new analyses. Although, as ECAWs were described in our survey, a research team could only use an ECAW once per dataset (because they will be exposed to the full dataset). Data management organizations could overcome this shortcoming by providing a secure data environment that publicly logs all analyses run, but never exposes the real dataset. Alternatively, they could provide access to only the subset of variables needed for a specific analysis. This would allow researchers to run another ECAW on the same larger dataset—if it focuses on variables not used in their previous ECAW. |
| **Concerns about impact** | |
| 5. Researchers can still p-hack / data-dredge when using ECAWs. | As is the case for conventional preregistration, ECAWs discourage, but do not necessarily prevent p-hacking and data-dredging. ECAWs increase transparency by making these questionable research practices detectable. |
| 6. Researchers may give findings from the subset too much weight (e.g., by not performing an analysis on the full dataset because it was null in the subset, or vice versa). | We hope that the instructions provided when implementing ECAWs make it clear that the subset of data is provided to help write an analysis script, but not to help decide which research questions to ask. |
| **Alternative interventions** | |
| 7. Data management organizations should require final analysis scripts to be shared (so that results are reproducible, and easy for other researchers to build on). | This policy could increase computational reproducibility. However, it may be difficult for data management organizations to implement and to ensure compliance. Whereas data management organizations can withhold access to the full dataset until an analysis script is registered, their influence is less direct at the publication stage. |
| 8. Data management organizations should maintain a repository that outlines each research project conducted using their dataset (to reduce duplication and facilitate replication). | Such a repository could be maintained in parallel with the use of ECAWs. It could also be used as a lighter touch intervention that may achieve some of the benefits we presume ECAWs would entail. |
| 9. Data management organizations could make a synthetic version of their dataset openly available. That dataset could be used for ECAWs. | Several initiatives have employed synthetic datasets (e.g., OpenSAFELY, Sarafoglou et al., 2023, Schmidt et al., 2023). On the one hand, creating a synthetic dataset requires more technical knowledge as compared to creating a data subset, and synthetic datasets can obscure the relationships between variables, which some researchers may dislike. On the other hand, they reduce the risk of participant re-identification and provide a full-sized dataset. All considered, synthetic datasets present a reasonable alternative to data subsets. |

[1] The text in the Respondents' comments column is synthesized from multiple responses and/or paraphrased. They are not verbatim responses.

[2] Sarafoglou et al.'s study used a pre-cleaned dataset. We are not aware of data that elucidates the time difference between ECAW-like workflows and conventional preregistration when starting from raw data.

## Discussion

<sub>230</sub>

<sup>231</sup> The survey results suggest that a trial of ECAWs with ALSPAC could be feasible for
<sup>232</sup> at least three reasons. First, ECAWs are possible because most respondents reported
<sup>233</sup> performing confirmatory analyses with analysis scripts written using programming
<sup>234</sup> languages. Second, ECAWs are relevant because many respondents reported limited use of
<sup>235</sup> other methods to improve rigour and reproducibility—including preregistration, sharing
<sup>236</sup> analysis scripts, and blinding analysts. Moreover, although participants generally agreed
<sup>237</sup> that findings from observational research using pre-existing data are trustworthy and
<sup>238</sup> reproducible, they also believed that ECAWs would make research findings more
<sup>239</sup> trustworthy and more reproducible. Third, ECAWs appear relatively acceptable. For
<sup>240</sup> example, only 10% of respondents disagreed that ALSPAC should run a study on ECAWs,
<sup>241</sup> and 25% agreed that they would be less likely to use the ALSPAC dataset if they were
<sup>242</sup> required to use an ECAW.

<sup>243</sup> The open-ended responses revealed interest in policies and interventions with similar
<sup>244</sup> goals to ECAWs. These include requirements for the sharing of final analysis scripts, a
<sup>245</sup> database of all ongoing studies that use a particular dataset, and openly available synthetic
<sup>246</sup> datasets.

### Limitations

<sup>247</sup>

<sup>248</sup> Given the response rate of 10%, our results represent the opinions of a select set of
<sup>249</sup> researchers who may be more interested and involved in reproducible research practices.
<sup>250</sup> Indeed, respondents themselves believed that they were more concerned about
<sup>251</sup> reproducibility than other researchers in their field (Supplementary Figure C2). ECAWs
<sup>252</sup> may be less acceptable among non-responders and non-response may be a mark of lack of
<sup>253</sup> interest in the concept. Nonetheless, the absolute number of responders suggests that there
<sup>254</sup> is an audience of researchers who might be interested to pursue this approach.

The survey results are best understood as the initial thoughts of participants when introduced to the concept of ECAWs. The median response time was 7.2 minutes (IQR: 4.3 to 13.1) for a survey that included over 20 questions and a 500-word description of ECAWs. Thus, it is unlikely that respondents spent much time reflecting on ECAWs and their implications.

We invited researchers from the ALSPAC mailing list which includes researchers across the fields of health and social sciences, but we did not record the specific disciplines in which the respondents were active. Thus, while ECAWs may be more relevant to some disciplines, the present survey does delve into the idiosyncrasies among disciplines.

Data management organizations could vary widely in their implementation of ECAWs. For example, they could perform checks on the analysis scripts to ensure they run and they could require commented analysis scripts with a clear indication of the primary outcomes. Our survey results do not elucidate the best implementation of ECAWs.

**Recommendations**

Some data management organizations looking to implement ECAWs will need to consider a balance between updating their data access workflow and maintaining their user numbers, engagement, and funding. This consideration will depend heavily on the structure of the data management organization. An organization managing data that is routinely collected somewhat regardless of its potential for use in research (e.g., electronic health records) may have leeway to test new workflows even if they impact user numbers. Other organizations—such as ALSPAC—coordinate ongoing data collection efforts that compound the value of their dataset and their continued operation remains contingent on funding cycles and data access fees. Even if ECAWs led to an increase in user numbers in the long-term, a temporary decrease could preclude the cost recovery systems on which their staff rely. In a more extreme case, the likelihood of another successful funding cycle

could be impacted and compromise the project's continuation. Funders and institutions interested in supporting these types of initiatives could alleviate concerns by providing targeted funding for testing these interventions and offering contingency funds to maintain cost recovery systems.

With these considerations in mind, an organization like ALSPAC may benefit from first leveraging the substantial number of respondents who would opt-in to a study on ECAWs. Trialing ECAWs with this user group would allow organizations to collect data that may support more widespread implementation, including project completion time, study quality, and researcher satisfaction when using ECAWs. They could refine the ECAW pipeline with minimal concern about user numbers. In the situation that a data management organization is already considering implementing policies on preregistration, they may benefit from considering alternate workflows such as ECAWs, which many respondents deemed preferable to conventional preregistration. The open-ended responses to our survey also suggest some confusion around the purpose of ECAWs and the process of using them. A clear-cut module provided by data management organizations that explains the ECAW concept alongside step-by-step instructions could help address researchers' concerns preemptively and help them adopt this workflow.

The stakes are lower in cases where a static final dataset already exists and concerns about funding are absent. For example, researchers with an interest in rigorous analyses and who control access to a dataset have already employed ECAW-like workflows (e.g., Forscher et al. (2020); Schmidt et al. (2023)). Concerns about a reduction in user numbers and engagement may also be less relevant for datasets containing unique data. For example, a researcher trying to answer a question about health and development with ALSPAC data, may also be able to answer that question with another cohort dataset. However, a researcher trying to answer a question about the population of a specific country may need access to that country's census data, regardless of the workflow required by that data management organization. A final consideration is that user numbers and engagement may

307 increase if researchers feel that ECAWs increase the trustworthiness of their findings and

308 others come to associate research from datasets using ECAWs as more open and rigorous.

## Conclusion

310 In this manuscript, we outlined a research workflow—ECAW—which necessitates

311 certain open science practices and can be implemented by data management organizations.

312 Responses to our survey provide information for organizations interested in developing and

313 testing ECAWs and interventions with related goals.

314 ────────────────

## Ethical approval

316 Ethical approval for the study was obtained from the ALSPAC Ethics and Law

317 Committee and the Faculty of Life Sciences Research Ethics Committee at the University

318 of Bristol (approval code: 12260).

## Acknowledgements

320 We thank Nicholas Timpson and Kate Northstone for invaluable input throughout

321 this research project. We thank the ALSPAC Executive for their collaboration and

322 willingness to run this study. We thank all the participants for their responses.

## Funding

## Contributions

**Conceptualization:** Robert T. Thibault and Marcus R. Munafò. **Data curation:** Marton Kovacs. **Formal analysis:** Robert T. Thibault and Marton Kovacs. **Funding acquisition:** Robert T. Thibault and John P. A. Ioannidis. **Investigation:** Robert T. Thibault. **Methodology:** Robert T. Thibault, Tom E. Hardwicke, Alexandra Sarafoglou, John P. A. Ioannidis, and Marcus R. Munafò. **Project administration:** Robert T. Thibault. **Resources:** Robert T. Thibault. **Software:** Marton Kovacs. **Supervision:** Robert T. Thibault and Marcus R. Munafò. **Validation:** Robert T. Thibault and Marton Kovacs. **Visualization:** Robert T. Thibault and Marton Kovacs. **Writing - original draft:** Robert T. Thibault. **Writing - review & editing:** Robert T. Thibault, Marton Kovacs, Tom E. Hardwicke, Alexandra Sarafoglou, John P. A. Ioannidis, and Marcus R. Munafò.

## Competing interests

All authors declare no conflict of interest.

## References

Bakker, M., Veldkamp, C. L., Assen, M. A. van, Crompvoets, E. A., Ong, H. H., Nosek, B. A., … Wicherts, J. M. (2020). Ensuring the quality and specificity of preregistrations. *PLoS Biology*, *18*(12), e3000937.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., … Chan, T. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436.

Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology. *Elife*, *10*, e71601.

Forscher, P. S., Noyce, A., DeBruine, L. M., Jones, B., Flake, J. K., Coles, N. A., & Chartier, C. R. (2020). *Incentivizing discovery through the PSA001 secondary analysis challenge.* https://psysciacc.org/2020/09/14/incentivizing-discovery-through-the-psa001-secondary-analysis-challenge/.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124.

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology*, 640–648.

Klein, J. R., & Roodman, A. (2005). Blind analysis in nuclear and particle physics. *Annu. Rev. Nucl. Part. Sci.*, *55*, 141–163.

MacCoun, R., & Perlmutter, S. (2015). Blind analysis: Hide results to seek the truth. *Nature*, *526*(7572), 187–189.

Prasad, V. K., & Cifu, A. S. (2019). *Ending medical reversal: Improving outcomes, saving lives.* Johns Hopkins University Press.

Sarafoglou, A., Hoogeveen, S., & Wagenmakers, E.-J. (2023). Comparing analysis blinding

with preregistration in the many-analysts religion project. *Advances in Methods and Practices in Psychological Science*, *6*(1), 25152459221128319.

Schmidt, K., Jones, A., Szabelska, A., Ebersole, C. R., Hawkins, C. B., Graham, J., & Nosek, B. A. (2023). *The ideology 2.0 study and dataset.* https://osf.io/2483h/wiki/home/; OSF.

Appendix A

Supplementary Material A. Deviations from the preregistration.

1. We rephrased our second objective to be more accurate. The preregistration reads "2. To use these insights to inform future research on how data management organizations can encourage rigorous and reproducible research practices (survey Blocks 3-6). This objective includes assessing and refining potential interventions—such as ECAWs—and assessing their acceptability. ". The manuscript reads "(2) To use these insights to inform future research—including a potential trial of ECAWs with the Avon Longitudinal Study of Parents and Children (ALSPAC)—on how data management organizations can encourage rigorous and reproducible research practices."

2. Marton Kovacs was added as a contributor during data collection. This led to a few of the projected contributor roles outlined in the preregistration to be different from the final contributor roles outlined in the manuscript.

3. The preregistration stated that: "We will tabulate descriptive summary statistics for all the survey questions." Instead of tabulating the results, the manuscript presents this data in figures. We feel that the figures are easier to digest as compared to tabulated data.

4. The preregistration stated that: "we will present results that include responses from participants who did not complete the entire survey, alongside the associated response rate for each question.". The results from all participants are presented in Supplementary Material D. However, instead of reporting the associated response rate for each question, we simply state that "The survey was completed 107 times and partially completed 27 times, leading to a response rate of 10

5. The preregistration and survey use the term "typical preregistration". We changed

<sub>402</sub> this to "conventional preregistration" in the manuscript because we believe it is the

<sub>403</sub> more appropriate term.

<sub>404</sub> 6. We performed a brief visual inspection of the data presented in spreadsheet format.

<sub>405</sub> Three participants clicked through to the end of the survey—so they were coded as

<sub>406</sub> completing the survey—however, they only responded to a few of the first questions.

<sub>407</sub> One other participant appears to have provided non-sincere responses in that all their

<sub>408</sub> responses were neutral or skipped and they completed the survey in 146 seconds. We

<sub>409</sub> retained these four participants as partially completed responses (results in

<sub>410</sub> Supplementary Material D), but removed them from the main dataset. These

<sub>411</sub> decisions were not preregistered.

<sub>412</sub> 7. We did not preregister which results we would present in the abstract. We decided to

<sub>413</sub> report the results with the highest and lowest percentage about the acceptability of

<sub>414</sub> ECAWs: "For example, only 10

Appendix B

Supplementary Material B. Invitation email

**Original email**

Dear ALSPAC Data User,

We are working with Dr Robert Thibault, a postdoctoral scholar at Stanford University and the University of Bristol, to support his work on scientific rigour and reproducibility in observational studies.

To assist him in his research we would be very grateful if you would consider completing a short survey. The purpose of this is to understand researcher's practices and thoughts regarding the rigour and reproducibility of observational research that uses pre-existing datasets (such as the ALSPAC resource). Results from this survey may be used in the future to inform initiatives for ALSPAC to better serve our users and to maximise the quality of the research using ALSPAC.

The survey has 21 questions across six sections. It will be open until November 1 at this link: https://bristolexppsych.eu.qualtrics.com/jfe/form/SV_3mVgk4lzXx4kp02. It will take approximately 10-20 minutes to complete. The data you provide is completely anonymous and further information is available on the consent form at the start of the survey.

If you have any questions or comments, please send them directly to Robert (robert.thibault@bristol.ac.uk).

Kind regards,

The ALSPAC Executive

**First follow-up email (sent after 1 week)**

Dear ALSPAC data user,

Thankyou very much to those of you who have already completed the survey below. There is still time to complete the survey (see details in the original email below), which closes of November 1st.

Please send any comments or queries directly to Robert (robert.thibault@bristol.ac.uk).

Many thanks,

The ALSPAC Executive

————————————————————

**Second follow-up email (sent after 2 weeks)**

Dear ALSPAC data user,

Thank you very much to those of you who have already completed the survey below. This will be our final email inviting you to complete the survey (see details in the original email below), which closes of November 1st.

The median time to complete the survey has been 8 minutes.

Please send any comments or queries directly to Robert (robert.thibault@bristol.ac.uk). Many thanks, The ALSPAC Executive
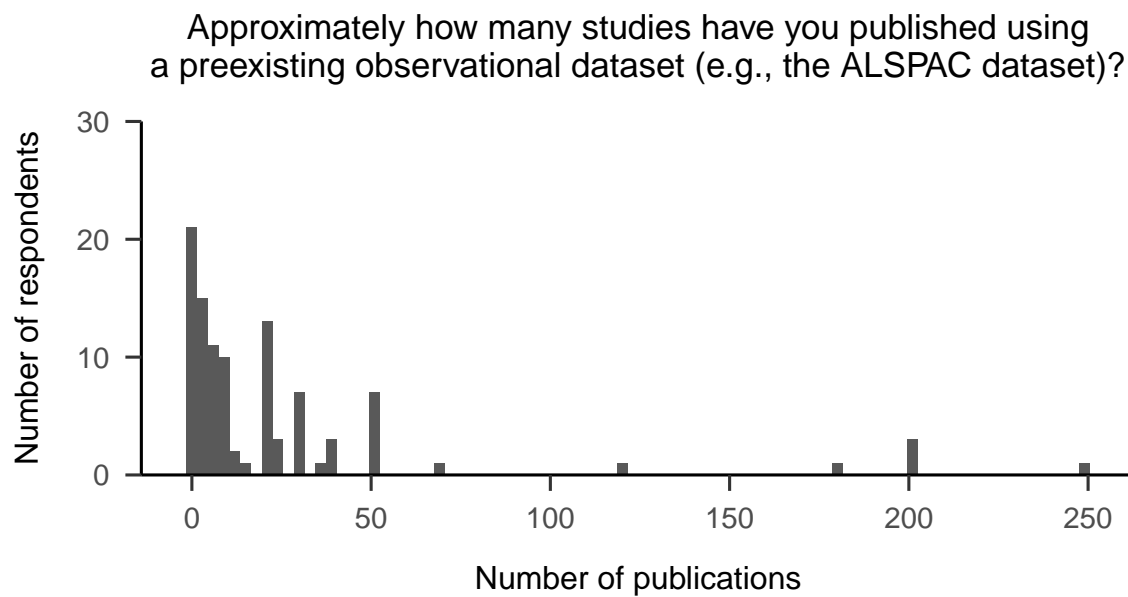
454 **Supplementary Material C. Participant Characteristics**

### Approximately how many studies have you published using a preexisting observational dataset (e.g., the ALSPAC dataset)?

Table B1

*Table C1. "What programming language or software do you use for your analyses of preexisting observational data? (you may select multiple answers)"*

| Programming language | N | Percentage of respondents |
|---|---|---|
| R | 65 | 61 |
| Stata | 48 | 45 |
| SPSS | 17 | 16 |
| SAS | 15 | 14 |
| Python | 6 | 6 |
| Missing | 4 | 4 |
| Mplus | 3 | 3 |
| Bash | 2 | 2 |
| MATLAB | 1 | 1 |
| Nextflow | 1 | 1 |
| plink2 | 1 | 1 |

Compared to what you think of as a typical researcher who uses preexisting observational data in your field, how concerned are you with research trustworthiness, bias, rigour, and reproducibility ...
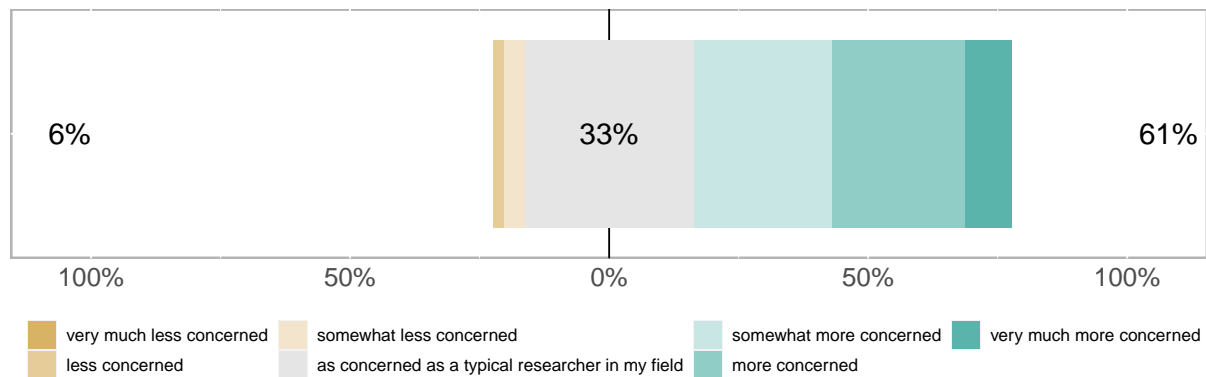


Figure C1. Responses to survey question about number of relevant publications

455 Supplementary Material D. Figures and Tables including participants who

456 partially complete the survey