A survey of registration practices among observational researchers using preexisting datasets

Robert T. Thibault[1,5], Marton Kovacs[2,6], Tom E. Hardwicke[3], Alexandra Sarafoglou[4], John P. A. Ioannidis[4], & Marcus R. Munafò[1,7]

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

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

[3] Melbourne School of Psychological Sciences, University of Melbourne.

[4] Department of Psychology, University of Amsterdam.

[5] School of Psychological Science, University of Bristol.

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

[7] Meta-Research Innovation Center Berlin (METRIC-B), QUEST Center for Transforming Biomedical Research, Berlin Institute of Health, Charité – Universitätsmedizin Berlin.

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

[9] Departments of Medicine, Epidemiology and Population Health, Biomedical Data Science, and Statistics, Stanford University.

Author Note

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

Correspondence concerning this article should be addressed to Robert T. Thibault, Enter postal address here. E-mail: robert.thibault@stanford.edu

Abstract

placeholder for an abstract

Keywords: keywords

Word count: X

A survey of registration practices among observational researchers using preexisting datasets

Introduction

Methods

Results

Participants

We invited the ALSPAC mailing list to participate, which included 1148 email addresses. 54 emails bounced, leaving 1094 emails that went through. The survey was completed 107 times and partially completed 21 times, leading to a response rate of 10% for complete surveys and 2% for incomplete surveys.[1] The median time taken for complete survey responses was 7.22 minutes (IQR: 4.30 to 13.10).

Respondents published a median of NA (IQR 2 to 25) studies using preexisting observational data (Figure S1). They reported using the programming languages 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) (Table S1)[2]. 61% 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 (Figure C2); 6% reported being less concerned.

[1] The ALSPAC mailing list has been active for >30 years and may contain 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. Excluding these email addresses would increase the response rate, but we do not know by how much.

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

Survey results

Most respondents agreed that studies that analyze preexisting observational datasets are trustworthy[3] (72%; 77/107) and reproducible[4] (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 2, 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 2A). About half reported sharing their analysis scripts never or almost never (20%; 21/107), or sometimes (31%; 33/107) (Figure 3). 75% (80/107) reported that they never or almost never blind the data analyst (Figure 3). Almost all respondents answered that they use both exploratory (91%; 98/107) and confirmatory (85%; 92/107) analyses at least sometimes (Figure 3).

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 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

[3] The survey defined trustworthy as: "meaning that the results and conclusions of the publications are valid, reliable, rigorous, and accurate. That they merit trust."

[4] The survey defined reproducible "in the sense that other researchers re-analysing the data with the same research question would produce similar results."

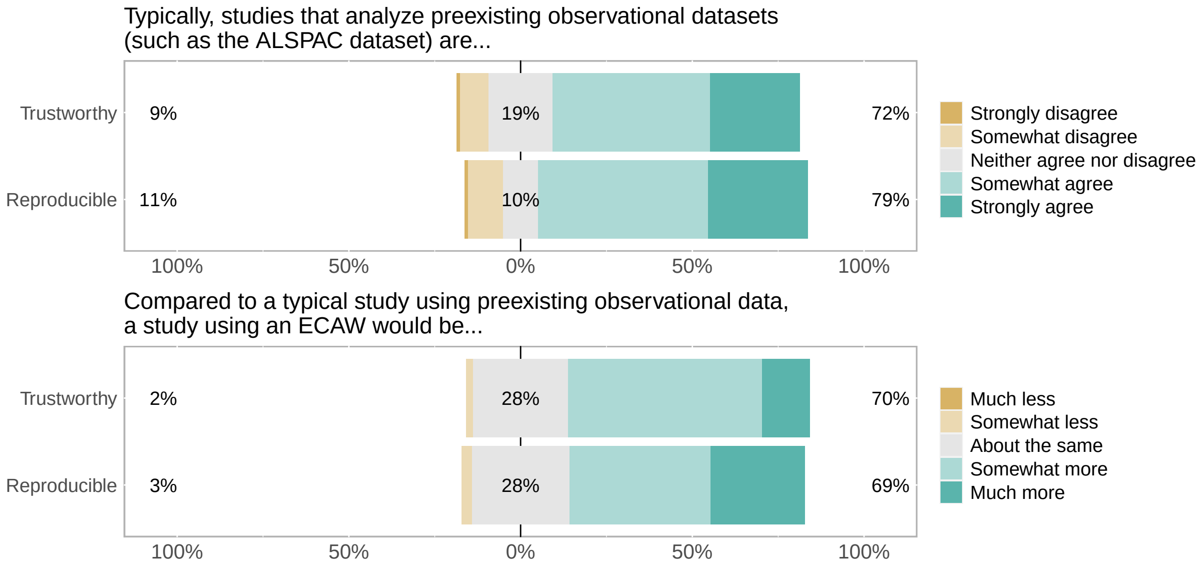

Figure 1. **Responses to the survey questions on trustworthiness and reproducibility of observational research with preexisting data and ECAWs.** The survey defined trustworthy as "meaning that the results and conclusions of the publications are valid, reliable, rigorous, and accurate. That they merit 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. For the bottom panel we excluded the missing responses (n = 2; 2) and responses of "I don't understand the question" (n = 4; 3).

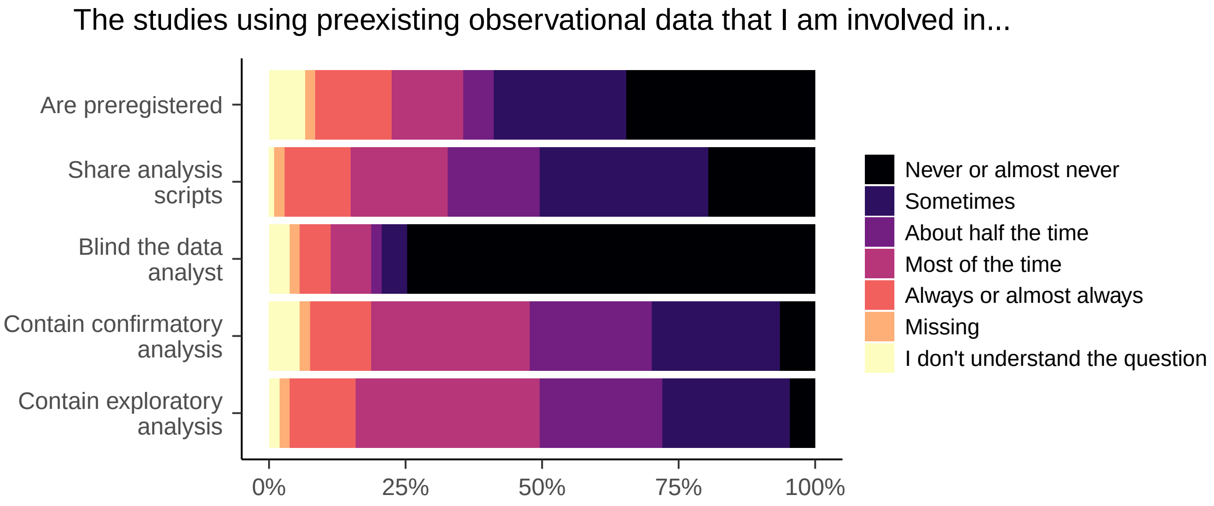

Figure 2. **Responses to survey questions about the research practices of participants.**

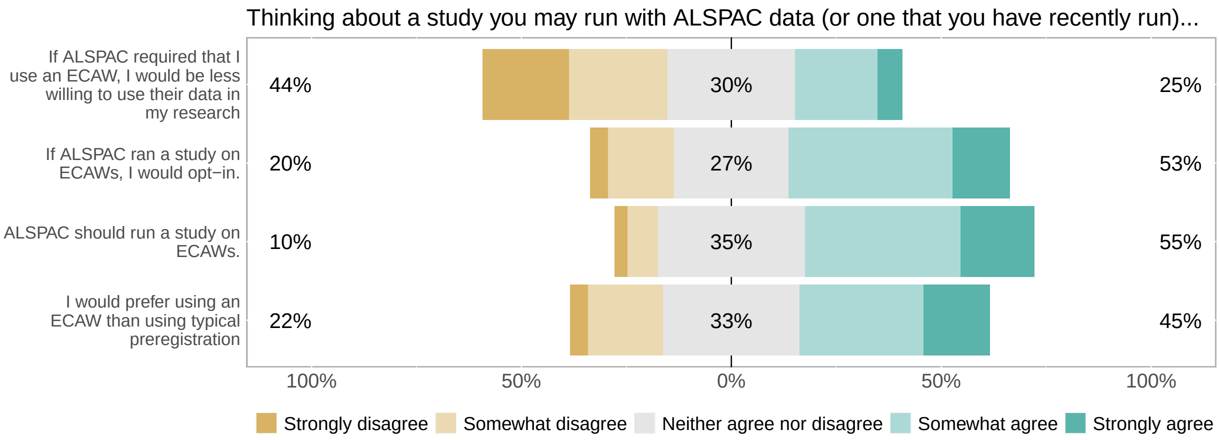

Figure 3. **Responses to survey questions about using ECAWs.** The 4 questions we excluded missing values (n = 3; 3; 3; 3), responses of "I don't understand the question" (n = 0; 4; 1; 1), and responses of "Unsure" (n = 2; 5; 6; 8).

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

Exploratory analyses

Discussion

Acknowledgements

References

Appendix

Supplementary materials C

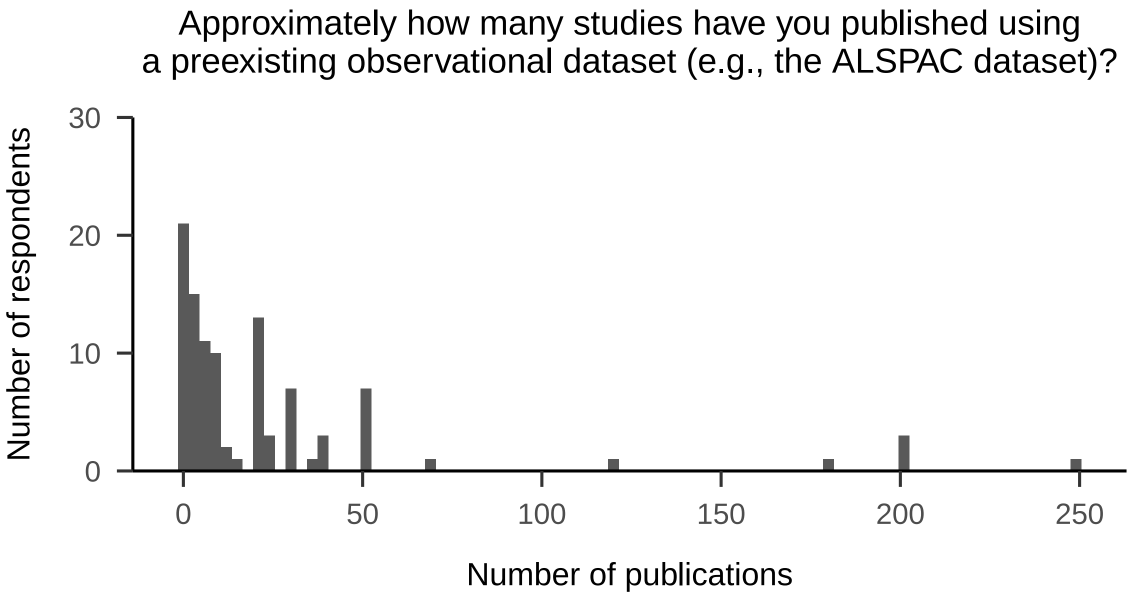

Figure A1. Caption goes here...

Table A1

What programming language or software do you use for your analyses of preexisting observational data?

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

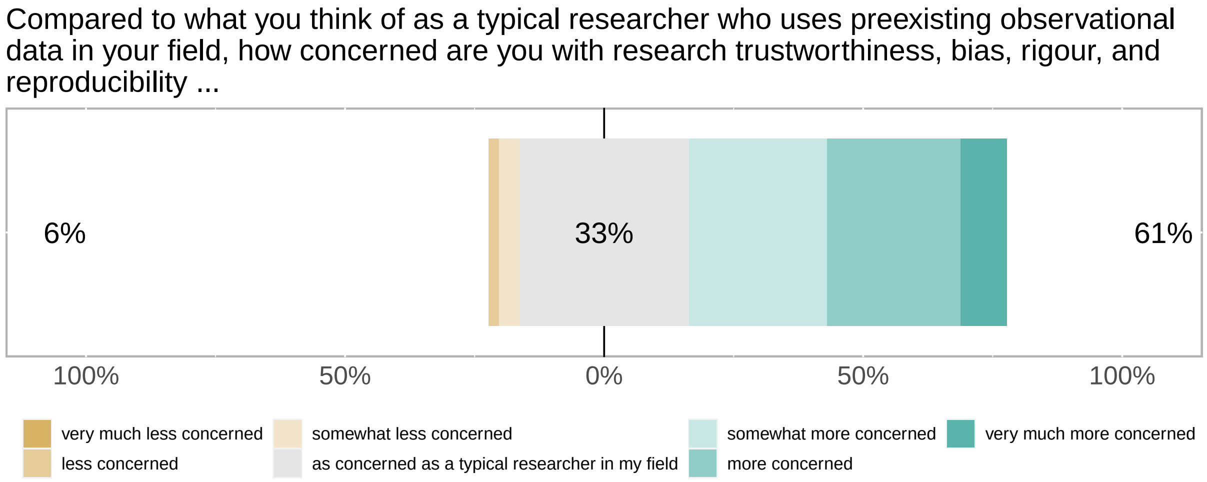

Figure A2. Caption goes here...