University of BRISTOL

INFORMATION SHEET

MAPS: Mapping the Analytical PathS of a crowdsourced data analysis

Robert Arbon, Katie Drax, Natalie Thurlby, Nic Timpson, Kate Northstone, Kate Robson Brown, Alex Kwong, and Marcus Munafò

You are being invited to take part in a research project. Before you decide, it is important for you to understand why the research is being done and what it would involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or if you would like more information. Take time to decide whether you wish to take part and remember that your participation is voluntary.

Overview

Background

For the public to have faith in the conclusions of scientists it is important that the methods they employ are robust and transparent. This is especially important for controversial topics with major implications for mental health. The public should rightly demand that such findings are not contingent on the beliefs of the scientists, their particular methods, computational quirks or simple accident. This is of particular relevance when total transparency is not possible because the data is sensitive.

This project aims to understand how scientists arrive at answers to controversial questions and whether crowdsourced analysis and data anonymisation techniques can ensure findings are robust and transparent.

Method

We will recruit teams of independent data analysts and look at how they answer the question, *"is computer use during weekdays and weekends at 16 years old associated with depression at 18 years old?",* using the same data. This 'crowdsources' the data analysis because multiple research teams simultaneously and independently explore the same research question using the same data effectively (Silberzahn et al., 2018).

Answers teams provide to this controversial question will be used to examine whether statistical results depend on how they are calculated. To do this, we will use a 'multiverse analysis' whereby reasonable alternatives to choices made by the teams, during the data analysis, are explored and recorded to see how sensitive the results were to the choices made (Steegen et al., 2016).

There is also the possibility that statistical results on who they are calculated by. For example results may vary according to the analyst's statistical expertise, expertise of different research domains or subjective beliefs about research question of interest (Silberzahn et al., 2018). To explore this, we will measure participants' statistical expertise,

domain expertise and subjective beliefs and explore the impacts these three factors have on the results.

The dataset teams will be analysing is from the Avon Longitudinal Study of Parents and Children (ALSPAC). As this is sensitive data we will synthesise an anonymised version which can be opened to the teams. Using anonymised data is controversial as the anonymisation process may erase important features of the data. This project aims to investigate the extent to which the anonymisation process affects the results by comparing results from both the synthetic data and the real, ALSPAC data.

Finally, our project will challenge the teams to come up with interesting ways to visualise the answers to these questions in exchange for a prize. Visualising the teams' answers, along with how robust they are, in a clear, accessible way will be important to help communicate complex results both for this study and in the future.

Project questions

The primary questions to be answered by the project will be:

1. How do teams' results compare to the multiverse of results?
2. How does the variability of teams' results compare the variability of the multiverse of results?
3. How does each team's results compare to all the other teams' results?
4. What are effective ways of visually communicating the above comparisons?

Secondary project questions to be answered will be:

5. Does domain expertise affect the results of teams' analyses?
6. Does statistical expertise affect the results of teams' analyses?
7. Are teams' subjective beliefs about the research question associated with the results of teams' analyses?
8. Does the anonymisation process affect the results of teams' analyses?

Planned timeline

There are eight phases for this project. Participants may be involved in four, namely Phases 3, 4, 7 and 8, and the details of this involvement are provided in the '**What will I have to do?'** section below and the *Team Instructions*. The materials participants will require to complete the project are detailed in Table 1. Some phases will start before the previous phase has been completed for example team registration and data analysis will begin while the project coordinators are promoting the project. Deadlines are subject to extension but we will notify participants in good time. The planned timeline is visualised in Figure 1.

Project start date: 25 March 2019

Phase 1: building datasets, materials and webpage. The project coordinators will build the datasets and materials and webpage for the project. The webpage will be maintained throughout the project and materials will be released at the start of the phase they are required for.

Phase 2: promotion (anticipated deadline: 6 May) . The project coordinators will promote the project both on and offline.

Phase 3: team registration and data analysis (anticipated deadline: 24 June) . Participants will register and complete and submit their analyses (required; see below for more details). There is not a target sample size, instead as many participants as possible will be recruited before the deadline. Upon submission, the project coordinators will run the team's script on the *Real Dataset* and report the results back to the team.

Phase 4: multiverse analysis (anticipated deadline 22 July): The project coordinators will perform the multiverse analysis on as many team analyses as possible within the allotted time. The selection will be random and any remaining team analyses will be performed after Phase 7.

Phase 5: data visualisation challenge (anticipated deadline: 23 August). Interested participants will produce visualisations of the *Project Dataset* (optional; see below for more details).

Phase 6: data visualisation challenge judgement (anticipated deadline: 6 September). A panel, which will include the project coordinators, Marcus Munafò and Kate Robson Brown, will judge the data visualisations.

Phase 7: showcase event (anticipated date: mid-September). Interested participants attend an event in Bristol arranged by the project coordinators (optional; see below for details).

Phase 8: manuscript writing (anticipated deadline: 30 September). The project coordinators will write up a manuscript reporting the results from this project. Interested participants will discuss the project and review any resulting manuscript (optional; see below for details).

Table 1. Required materials

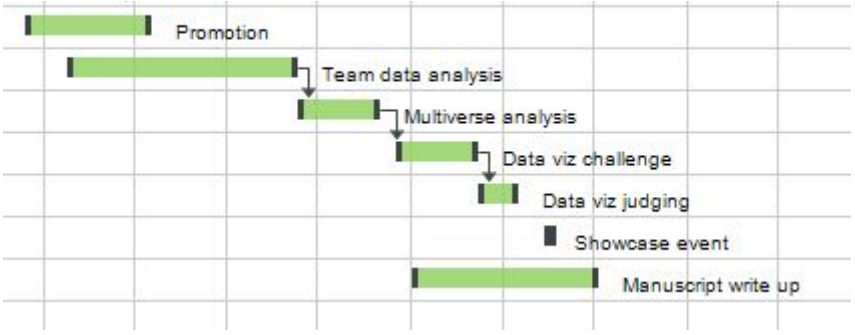
Name	Description	Access
Real Dataset	The dataset used to generate the *Synthetic Dataset.* It is based on a dataset from the ALSPAC used in a previous analysis.[1]	N/A. This data is only available from ALSPAC to bona fide researchers.
Synthetic Dataset	The dataset to be analysed by participants. It will be synthesised from the *Real Dataset*.	By email upon completion of a *Team Registration Form.* Once registration is closed it will be made openly available.

[1] Khouja, J. N., Munafò, M. R., Tilling, K., Wiles, N. J., Joinson, C., Etchells, P. J., ... Cornish, R. P. (2019). Is screen time associated with anxiety or depression in young people? Results from a UK birth cohort. *BMC Public Health*, *19*(1), 82. https://doi.org/10.1186/s12889-018-6321-9

Project Dataset	The data generated by this project i.e. the participants' questionnaire results, models and multiverse analyses.	Project Open Science Framework page (https://osf.io/9qke2/)
*Information Sheet**	This document which summarises the project and provides links to the other materials.	
*Team Instructions**	Step-by-step guide to the project. Expands on the **What will I have to do?** section in this document.	
Team Registration Form	Form to complete to be included as a participant in project.	
*Final Analysis Report**	Form participants use to submit their data analyses of the *Synthetic Dataset.*	
Real and Synthetic Dataset Description	A comparison of the *Real* and *Synthetic Datasets.*	
Data Dictionary	A data dictionary describing the variables in the *Real* and *Synthetic Datasets.*	
Data Visualisation Report	Form participants use to submit their visualisation attempts of the *Project Dataset.*	
Project Dataset Description	A data dictionary for the *Project Dataset* and a description of the multiverse analysis.	

* Informed by Silberzahn et al.'s supplementary materials.

Figure 1. Project gantt chart



The data

The data for this project is from the Avon Longitudinal study of Parents and Children (ALSPAC, also known as Children of the 90s). For more information about the study visit http://www.bristol.ac.uk/alspac/**.**

As the ALSPAC data is highly sensitive we cannot open the data to the public without ensuring the data is anonymised and non-identifiable. To achieve this an anonymised dataset will be generated using the R package SynthPop (Nowok, Raab & Dibben, 2016) on the ALSPAC data, creating the *Synthetic Dataset* and the *Real Dataset*. Participants will have access to the *Synthetic Dataset* but not the *Real Dataset*.

The *Synthetic* and *Real Datasets* contain a sample of 15,445 and over 80 variables. A detailed description of the variables included is in *Data Dictionary.*

Project end-goals

The findings from this research project may be published in an appropriate scientific journal (and made available open access), and/or presented at an appropriate meeting. Anonymised project data will be collected and held by the project coordinators. All the data, *except for the Real Dataset*, will be made available for sharing via an online data repository.

Why have I been invited?

You have been invited because you have enquired about the project or have asked to receive further information following reading the summary version described in a project advertisement.

Do I have to take part?

It is up to you to decide whether or not to take part. If you do decide to take part you would be given this information sheet to keep and be asked to sign a consent form prior to any further procedures. If you decide to take part you are still free to withdraw at any time and without giving a reason. A decision to withdraw at any time, or a decision not to take part, would not affect your future or be held against you in any way.

Am I eligible to take part?

Please read all of these criteria very carefully and contact the researcher if you have any doubts regarding your eligibility. If you do not meet the criteria your data will not be used and you will not be included as an author.

In order to take part you should;

- Be aged 18 or above
- Perform your final analysis (see below) in either R or Python and report it in English.

You would **not** be able to take part in the project if you;

- Register your interest in the project outside of the stated time frame.

Expenses and reimbursement

You would receive no financial reimbursement for your participation in the project and your participation is voluntary.

What will I have to do?

For a step-by-step guide on what analysts will be doing please see the *Team Instructions.* A summary of the actions detailed in this document are summarised below. If you wish to be included as an author on a resulting manuscript you must complete the **required** actions relating to Phase 3. All other actions are optional.

1. **Registration, data analysis and submission (required; part of Phase 3)**
    a. Register your team with the project coordinators. Registration is complete once all team members have submitted a *Team Registration Form.*
    b. Analyse the *Synthetic Dataset* to answer the research question "*is computer use during weekdays and weekends at 16 years old associated with depression at 18 years old?".*
    c. Submit your data analysis to the project coordinators using the *Final Analysis Report*
2. **Data visualisation challenge (optional; part of Phase 5)**
    a. Compete in the data visualisation challenge. You may attempt to visualise the results from the project and submit your attempts to the project coordinators for judging.
3. **Showcase event (optional; part of Phase 7)**
    a. Attend the showcase event. This will involve talks and a prize giving for the winners of the Data Visualisation Challenge.
4. **Project discussion (optional; part of Phase 8)**
    a. While the manuscript is being written up participants will be invited to discuss the methods and results of the project with the project coordinators and other participants.
    b. Review a first draft of the manuscript

What are the possible benefits of taking part?

If you complete the project you will included as an author on the resulting paper if you wish to be. You will also be invited to attend a showcase event in Bristol which will present the results of the project.

Authorship is earned by completing and submitting a reproducible analysis within the stated time frame. This refers to the submission of a satisfactory *Final Analysis Report*, as judged by Robert Arbon, Katie Drax and Marcus Munafò. If an analyst does not provide a satisfactory *Final Analysis Report* their analysis strategy will not be run on the *Real Dataset* and they will not be entitled to authorship.

What if there is a problem?

Any complaint about the way you have been dealt with during the project or any possible harm you might suffer would be addressed.

We anticipate no potential harms from taking part in this research project. If you are harmed there are no special compensation arrangements. If you are harmed due to someone's negligence, then you may have grounds for legal action but you may have to pay for it. Regardless of this, if you wish to complain or have any concerns about any aspect of the way you've been approached or treated during the course of this project , please contact the project coordinators (maps-project@bristol.ac.uk).

Will my taking part in this project be kept confidential?

Yes. Your identity and personal information that could identify you (e.g., name, email address, date of birth) will be kept securely by the project team and will not be shared publicly or with other research groups. On occasion this information may be made available to university research staff and government bodies which monitor whether research studies are performed properly. However, this will only be in the context of monitoring and this information will not be used to contact you or to make your participation in this project known.

This research project will adhere to General Data Protection Regulations (GDPR). We will be using information from you in order to undertake this project and will act as the data controller for this. This means that we are responsible for looking after your information and using it properly. We will keep identifiable information about you (name, email address) until one year after the project, but this will not be shared or be part of your project data. If you agree to take part in the project, data about you will be processed for a task in the public interest, which is consistent with the University Charter for scientific research.

What would happen to the results of the research project?

During the project, we will anonymise any data you provide either as an individual or team member. This means we give the data a unique identification number and your personal information (e.g., name, date of birth, email address) is removed, so that you cannot be identified by this information.

When the project has been completed, we would analyse the project data we have collected and report the findings. This would be reported in an appropriate scientific journal or presented at a scientific meeting. As your project data are anonymised, it would not be possible to identify you by name from the project data. However you will be identifiable as a participant in the project if you wish to be included as an author on the resulting paper.

Your rights to access your information are limited, as we need to manage your information in specific ways in order for the research to be reliable and accurate. To safeguard your rights, we will use the minimum personally-identifiable information possible.

At the end of the project your data would become “open data”. This means that it would be stored in an online database so that it is publicly available.

What is open data?

Open data means that project data are made available, free of charge, to anyone interested in the research, or who wishes to conduct their own analysis of the data. We would therefore have no control over how these data are used. However, all data would be anonymised

before being made available and therefore there would be no way to identify you from the project data.

Why open data?

Sharing research data and findings is considered best scientific practice and is a requirement of many funding bodies and scientific journals. As a large proportion of research is publicly funded, the outcomes of the research should be made publicly available. Sharing data helps to maximise the impact of investment through wider use, and encourages new avenues of research.

Can I withdraw my project data after I have participated in the project?

Yes. If you decide that you do not want your data to be used you can contact the project team and request that your data are withdrawn. You can do this up to one year after the project ends or up until the point the data are shared as "open data" (whichever comes first). At this point links between your identity and your anonymised data set would be destroyed, and therefore we would no longer be able to withdraw your data as we would no longer be able to identity which data set is yours.

Who is organising and funding the research?

Katie Drax and Robert Arbon are the guarantors and coordinators of the project. Marcus Munafò will supervise the project. All other project personnel (i.e. the authors of this information sheet) contributed to the development of the project design.

This work is being funded by the John Climax Benevolent Fund. The funder will support the conduct of the work by paying KD's stipend. The funder has and will have no input on any aspect of the project, such the protocol design, data collection, data analysis nor interpretation or publication of results.

Who has reviewed the project?

Ethics approval for the project was obtained from the Faculty of Science Research Ethics Committee at the University of Bristol (approval code: 21011980583). The project was approved by the ALSPAC Executive Committee (proposal: B3246; https://proposals.epi.bristol.ac.uk/?q=node/129851). Full details of ethics committee approval references for the ALSPAC study can be found online (http://www.bristol.ac.uk/alspac/researchers/research-ethics/).

.

Who can I contact for further information?

Please contact Robert Arbon and Katie Drax (maps-project@bristol.ac.uk).

Project personnel

Robert Arbon
School of Chemistry
Cantock's Close
Bristol, BS8 1TS
robert.arbon@bristol.ac.uk

Katie Drax
School of Psychological Science
5 Priory Road
Bristol, BS8 1TU
katie.drax@bristol.ac.uk

Natalie Thurlby
Jean Golding Institute
Royal Fort House,
Bristol, BS8 1UH
natalie.thurlby@bristol.ac.uk

Nic Timpson
Bristol Medical School
Oakfield House
Bristol, BS8 2BN
n.j.timpson@bristol.ac.uk

Kate Northstone
Bristol Medical School
Oakfield House
Bristol, BS8 2BN
kate.northstone@bristol.ac.uk

Kate Robson Brown
Jean Golding Institute
Royal Fort House,
Bristol, BS8 1UH
kate.robson-brown@bristol.ac.uk

Alex Kwong
School of Geographical Sciences
Oakfield House
Bristol, BS8 2BN
alex.kwong@bristol.ac.uk

Marcus Munafò
School of Psychological Science
12a Priory Road
Bristol, BS8 1TU
marcus.munafo@bristol.ac.uk

References

Nowok, B., Raab, G. M., & Dibben, C. (2016). synthpop: Bespoke Creation of Synthetic Data in R. https://doi.org/10.18637/jss.v074.i11

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., … Nosek, B. A. (2018). Many Analysts, One Data Set: Making Transparent How Variations in

Analytic Choices Affect Results. Advances in Methods and Practices in Psychological Science. https://doi.org/10.1177/2515245917747646

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. Perspectives on Psychological Science, 11(5), 702–712. https://doi.org/10.1177/1745691616658637