A practical guide to adopting transparent and reproducible practices in statistically orientated social science research during COVID-19

NATIONAL CENTRE FOR RESEARCH METHODS

The unprecedented nature of the COVID-19 global pandemic has had momentously disruptive effects on contemporary social life. The empirical findings that flow from social science inquiries have important implications for establishing policies and changing practices. The speed at which the pandemic has unfolded has led to a previously unparalleled requirement for rapid results from social science studies. This acceleration has consequences for verifying empirical results, and for building incrementally on research findings.

This guide considers the methodological issues associated with undertaking transparent and reproducible social science research and provides a set of recommendations. The focus of this guide is social science research that employs statistical techniques for the analysis of large-scale and complex datasets (e.g. social surveys, administrative social science data and big data resources); however many of the issues pervade other forms of social science research.

The challenge

The COVID-19 pandemic is an urgent threat to global health, and during the crisis a number of authors in the scientific community have emphasised that research must be a reliable, rigorous and transparent process, especially in the context of a pandemic where research findings need to be rapidly translated into practices[1,2,3,4,5]. In health research it has been recognised that when researchers share data, research code, and software and generally make their work as transparent as possible, it allows other researchers to verify results and to expand upon work, and it enhances public officials' ability to make scientifically informed decisions[6].

There is a parallel requirement for rapid results from social science studies during the COVID-19 pandemic. Similarly, social science research on COVID-19 must be transparent in order that findings can be verified, and that results can be reproduced and incrementally developed. There are both general methodological issues associated with undertaking transparent and reproducible social science research that employs statistical techniques for the analysis of large-scale and complex datasets, and some specific methodological issues associated with research during the COVID-19 pandemic.

The problems

This guide was produced during the COVID-19 pandemic, and reflects the current methodological issues associated with undertaking statistically orientated social science analyses of large-scale data resources.

1. The concept of data sharing is not especially new[7], but its potential importance has been reemphasised during the COVID-19 pandemic[8]. Some social science datasets have open access, but many large-scale social science datasets are only available via an end user license (e.g. from a national data archive), and analytical datasets cannot be shared publicly.

2. Conventional outlets for publishing social science research findings, for example paper-based academic journals, do not provide sufficient space for researchers to provide the exact details of how the analytical dataset was produced. Social science enterprises that undertake statistical analyses of large-scale data resources usually commence with an unprocessed dataset. It is common that a large amount of data enabling work (often called data wrangling) is undertaken to produce the 'analytical dataset' that is required for the research. The data wrangling phase will usually include operations such as organising variables into formats that are suitable for the analyses.

In this phase, it is typically for the data analyst to be guided by theoretical considerations and by practical requirements when selecting appropriate measures and deciding how to operationalise them. Variable selection is not a trivial activity. Research datasets often contain a wide range of variables, and can commonly contain different measures of key analytical concepts such as income, socioeconomic status, and education[9]. Analytical datasets are the combination of the decisions that are made and the actions that are taken during the data wrangling phase; these comprise choosing which cases to include and operationalising and coding measures.

It is infeasible for a third party who is unconnected with the original research to be able to validate an empirical result without access to the analytical datasets. Analytical datasets are too complex to be 'reversed engineered' from the limited information that is routinely provided in conventional published outputs. Many social science journal articles contain the popular statement, often within a footnote, that further information is 'available by request'. In reality, this protocol for gaining more detailed access to research materials is ineffective[10]. During the COVID-19 pandemic the inability to gain rapid access to information on how the analytical dataset was constructed restricts the possibilities for other researchers to verify results and to expand upon work; and it may also diminish public officials' trust in social science results.

3. In practice, the particulars of more comprehensive statistical analyses cannot be deduced from the short methods sections that are contained in most paper-based journal articles, or even from well annotated tables of empirical results. This issue is exacerbated because many contemporary large-scale surveys have complex designs. Detailed information on the survey's characteristics for example sampling, stratification, clustering, and weighting, may be obtained (e.g. in the survey documentation). However, precise information on how these features of the survey were represented in analyses is required to consistently reproduce results.

Handling missing data is another example of when detailed information is essential for duplicating results, and the problem is acute when comprehensive techniques, such as multiple imputation, are employed. There are also more imperceptible analytical situations where detailed information is required, such as when there are technical differences between statistical approaches. One illustration is the variety of possible estimation procedures that can be employed within multilevel modelling. Insufficient detailed information on the analytical procedures presents barriers to other researchers being able to accurately duplicate social science research results in order to verify and build on empirical findings.

Recommendations

We propose the following guidelines for undertaking transparent and reproducible social science research that employs statistical techniques for the analysis of large-scale and complex datasets during the COVID-19 pandemic.

1. If it is legal, and if it is feasible, then publicly share the analytical dataset.

2. Clearly identify the exact version of the unprocessed (or raw) dataset and its origins (i.e. where and when it was obtained) using a persistent identifier such as a digital object identifier (DOI).

3. Use established data analysis tools (e.g. Stata, SPSS, R or SAS) because using an esoteric statistical analysis software or programing language will not aid reproducibility.

4. Clearly record which data analytical tools are used including the version, and all the libraries, dependencies and plugins that are used.

5. Construct a data dictionary in a clear and literate[11] format that can easily be understood by someone unconnected with the original project.

6. Write down all of the research code (e.g. the Stata code or SPSS syntax)[12] for how the analytical data were prepared for analysis, in a clear and literate format that can easily be understood by someone unconnected with the project.

7. Write down all of the research code for all of the analyses undertaken, and not just the analyses that are presented in the published work, in a clear and literate format that can easily be understood by someone unconnected with the project.

8. Use a specialist platform such as Open Science Framework (OSF)[13] where research code can be shared alongside further project related materials such as conference presentations and preprints.

9. Create a Research Object (RO) which is an artefact that packages up research outputs (e.g. data, metadata, code, results, documentation, and academic papers)[14].

10. Ensure that Research Objects (RO) are produced under the FAIR principles, this means that they should be **F**indable, **A**ccessible, **I**nteroperable and **R**eusable[15].

Useful resources

Connelly, Roxanne, Vernon Gayle, and Chris Playford. Transparent and Reproducible Data Analysis. SAGE Publications Limited, 2020. https://methods.sagepub.com/foundations/transparent-and-reproducible-data-analysis

Playford, Christopher J., Vernon Gayle, Roxanne Connelly, and Alasdair JG Gray. "Administrative social science data: The challenge of reproducible research." Big Data & Society 3, no. 2 (2016): 2053951716684143. https://journals.sagepub.com/doi/pdf/10.1177/2053951716684143

Reproducible Social Research NCRM Online Resource by Vernon Gayle https://www.ncrm.ac.uk/resources/online/all/?id=20732

Research Object examples for ‘Parental Social Class and Filial School Level Educational Outcomes in Contemporary Britain’ ESRC SDAI PROJECT ES/R004978/1 https://osf.io/vgfnr/

References

1. Lonni Besançon et al., "Open Science Saves Lives: Lessons from the Covid-19 Pandemic," BioRxiv (2020).

2. Thierry Gustot, "Quality and Reproducibility During the Covid-19 Pandemic," JHEP Reports: Innovation in Hepatology 2, no. 4 (2020).

3. https://www.natureindex.com/news-blog/covid-nineteen-coronavirus-data-sharing-scientific-research-publishing (accessed 12.03.21).

4. C Michael Barton et al., "Call for Transparency of Covid-19 Models," Science 368, no. 6490 (2020).

5. Mohammad S Jalali, Catherine DiGennaro, and Devi Sridhar, "Transparency Assessment of Covid-19 Models," The Lancet Global Health 8, no. 12 (2020).

6. Josh Q Sumner et al., "Reproducibility and Reporting Practices in Covid-19 Preprint Manuscripts," medRxiv (2020).

7. Stephen E Fienberg, Margaret E Martin, and Miron L Straf, Sharing Research Data (National Academy Press, 1985).

8. Rafael S Rios, Kenneth I Zheng, and Ming-Hua Zheng, "Data Sharing During Covid-19 Pandemic: What to Take Away," Expert Review of Gastroenterology & Hepatology 14, no. 12 (2020).

9. Roxanne Connelly, Vernon Gayle, and Paul S. Lambert, "Modelling Key Variables in Social Science Research: Introduction to the Special Section," Methodological Innovations 9 (2016).

10.https://orgtheory.wordpress.com/2015/08/11/sociologists-need-to-be-better-at-replication-a-guest-post-by-cristobal-young/ accessed 12.03.21.

11. Donald E Knuth, Literate Programming (Addison-Wesley Publishing Company, 1992).

12. For an extended discussion of the role of research code see Victoria Stodden, Peixuan Guo, and Zhaokun Ma, "Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals," PloS one 8, no. 6 (2013).

13. https://osf.io/ (accessed 12.03.21).

14. Sean Bechhofer et al., "Why Linked Data Is Not Enough for Scientists," Future Generation Computer Systems 29, no. 2 (2013).

15. Mark D Wilkinson et al., "The Fair Guiding Principles for Scientific Data Management and Stewardship," Scientific data 3 (2016).

This guide was produced in 2021 by Vernon Gayle, in collaboration with Roxanne Connelly and Christopher Playford and draws on work undertaken as part of ESRC Project ES/R004978/1 'Parental Social Class and Filial School Level Educational Outcomes in Contemporary Britain'.

National Centre for Research Methods
Social Sciences
University of Southampton
Southampton, SO17 1BJ
United Kingdom.

Web	www.ncrm.ac.uk
Email	info@ncrm.ac.uk
Tel	+44 23 8059 4539
Twitter	@NCRMUK