How to protect privacy in open data

Wicherts, Jelte M.; Klein, Richard A.; Swaans, Sofie H. F.; Maassen, Esther; Stoevenbelt, Andrea H.; Peeters, Victor H. B. T. G.; Jonge, Myrthe de; Rüffer, Franziska

Published in:
Nature Human Behaviour

DOI:
10.1038/s41562-022-01481-w

Publication date:
2022

Document Version
Peer reviewed version

How to protect privacy in open data

Nature Human Behaviour 2022
https://doi.org/10.1038/s41562-022-01481-w

Jelte M. Wicherts[1*], Richard A. Klein[1], Sofie H. F. Swaans[1], Esther Maassen[1], Andrea H. Stoevenbelt[2], Victor H. B. T. G. Peeters[1], Myrthe de Jonge[1], & Franziska Rüffer[1]
[1]Department of Methodology and Statistics, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands
[2]Department Educational Science, University of Groningen, Grote Rozenstraat 3, 9712 TG Groningen, The Netherlands

*j.m.wicherts@tilburguniversity.edu

When sharing research data for verification and reuse, behavioural researchers should protect participants' privacy, particularly when studying sensitive topics. Because personally identifying data remains present in many open psychology datasets, we urge researchers to mend privacy via checks of re-identification risk before sharing data. We offer guidance for sharing responsibly.

The general public, funders, institutions, publishers, and researchers themselves increasingly promote and value the sharing of data alongside published behavioural research. Openly sharing data enables independent verification of results, allows reuse, and might increase scientific impact and rigour[1], but it also poses privacy risks to research participants and survey respondents. Such privacy risks are particularly relevant when the data pertain to potentially sensitive issues such as participants' health, sexual activities or preferences, ethnicity, political opinions, criminal behaviour, professional life, and religious or philosophical beliefs. A key challenge for the behavioural sciences in the era of open science is how to share data responsibly.

Legal and ethical requirements

While professional ethical standards, like ethics code of the American Psychological Association [2], fail to specify how one should protect privacy, most countries now have extensive privacy laws, such as the General Data Protection Regulation (GDPR) in the European Union and the Health Insurance Portability and Accountability Act (HIPAA) for health-related data in the US. We do not expect many behavioural researchers to be familiar with all stipulations in privacy laws such as those in the 261-page GDPR, but do stress their responsibilities to protect participants' privacy, especially in open data.

A simple way to protect privacy is to disallow re-identification of individuals in the data, but how could one determine such identifiability in practice? Given the broad definition of personal data in the GDPR that might even encompass weather data[3], it is practically useful to consider HIPAA's identifiers to assess identifiability of research participants in datasets. The HIPAA identifiers include names, initials, address information (e.g., street address, city, zip code), birthdates, telephone and fax numbers, email addresses, social security or medical record numbers, account numbers, certificate or license numbers, vehicle numbers, Web URLs, Internet Protocol (IP) addresses, finger or voice prints, photographic images, and other characteristics that could uniquely identify someone. Most direct identifiers can be readily detected within datasets, and we are currently working on tools to do so automatically[4]. It is key to exclude direct identifiers from shared data, particularly when the data bears on sensitive issues such as racial or ethnic origin, political preferences, health, mood, religion, philosophical beliefs, professional standing, biological characteristics, or sexual preferences and behaviours.

Reidentification risks are common

But beyond direct identifiers, we urge behavioural researcher to check whether any participant can be uniquely identified given information on the sampling scheme and other (demographic) information in the dataset. For instance, if a study recruited undergraduate students from a given university and the data listed gender, age, ethnicity, or country of birth, it is often not too hard to re-identify some of the participants by cross-referencing the given information, particularly when adding information from other (public) sources.

For instance, some time ago one of us completed a survey about academic life asking for academic specialty, position, and university alongside questions about harassment. Supposed "assurances of anonymity" in the survey meant very little as there is only one full professor in methodology at Tilburg University. In this instance re-identification was not problematic, but re-identification risks are particularly high when studying demographic minorities (e.g., based on ethnicity and gender) that might be more vulnerable and whose exposed sensitive answers might cause real harm. Readers may be surprised at how easily demographic information can identify individuals[5]. For example, 87% of American voters could be identified merely from zip code, gender, and date of birth[6]. It is clear that merely deleting directly identifying variables like names or email addresses is insufficient and that genuine efforts must be made to assess re-identification risks before sharing data.

Unfortunately, many datasets shared alongside articles in psychology still include identifiable information. In a recent study of computational reproducibility that involved over 2000 datasets, we found directly identifying information in around 5% of open datasets[7]. (At the time of writing, this study is a non-peer reviewed preprint.) The direct identifiers in these datasets included IP addresses, dates of birth, initials combined with age, full names, and first names combined with location. An additional 4% of the datasets had no directly identifying information, but still involved some risk of re-identification, because they included detailed demographic variables that could be used together with reported information on the sampling scheme or other information to re-identify some participants. In only one of the cases did we identify an overriding scientific rationale to include such

information in the shared data. Box 1 includes an overview of common issues with reidentification we encountered in the shared data sets.

How to minizine reidentification risks

The benefits of open data for science and society are abundantly clear, but we urge behavioural researchers sharing data to minimize the risk to participant privacy[8,9]. Often, simple measures could help restore anonymity (see Box 2). For starters, it is frequently not needed to collect direct identifiers. But if identifying data on participants are needed for linkage and analyses during data collection (e.g., in extensive longitudinal and multi-source studies), researchers should protect privacy diligently. Specifically, one could use pseudonymization and keep the keys that relate the personal identifiers and randomized participant identifiers separately in a secure location and manner (e.g., using encryption). Yet, for many behavioural studies, such identifying information is not needed to begin with and should certainly not be shared openly. Rich qualitative (text) data or recordings should be gleaned for identifying information and only shared fully under strict conditions of confidentially.

Indirect identifiers related to demographics are often needed for the analyses, but should be handled such that they cannot be used to reidentify participants. Researchers should carefully consider whether singular or cross-referenced information could identify some of the participants. One way to do is to ask a colleague or one of the co-authors to perform meticulous check of the re-identifiability of participants before sharing data. Subsequently, one could delete easily identifiable data from the datasets before sharing, create new larger categories of (demographic) variables such that a few or single cases do not stand out, or shuffle some data if needed to ensure privacy or use differential privacy methods to do so.

Researchers who plan to share should always obtain informed consent for sharing data and clearly indicate to participants how privacy will be protected using safe storage with encryption for sensitive personally identifiable information. In their effort to promote scientific rigor through transparency, behavioural researchers have a real responsibility, not merely an ethical but increasingly also a legal one, to protect the privacy of research participants.

References

1 Wicherts, J. M. Science revolves around the data. *Journal of Open Psychology Data* **1**, e1, doi:10.5334/jopd.e1 (2013).

2 American Psychological Association. *Ethical Principles of Psychologists and Code of Conduct*. (American Psychological Association, 2017).

3 Purtova, N. The law of everything. Broad concept of personal data and future of EU data protection law. *Law, Innovation and Technology* **10**, 40-81, doi:10.1080/17579961.2018.1452176 (2018).

4 Wicherts, J. M., Klein, R. A. & Hartgerink, C. H. J. Automatic Detection of Identifiers in Open Data (ADIODA). *ResearchEquals.com*, doi:10.53962/qj4h-2j1n (2022).

5 El Emam, K., Jonker, E., Arbuckle, L. & Malin, B. A systematic review of re-identification attacks on health data. *PLoS One* **6**, e28071, doi:10.1371/journal.pone.0028071 (2011).

6 Sweeney, L. *Simple Demographics Often Identify People Uniquely*, <https://dataprivacylab.org/projects/identifiability/paper1.pdf> (2000).

7 Wicherts, J. M. *et al.* Privacy Protection in the Era of Open Science. doi:10.31234/osf.io/ybzu9 (2022).

8 Meyer, M. N. Practical Tips for Ethical Data Sharing. *Advances in Methods and Practices in Psychological Science* **1**, 131-144, doi:10.1177/2515245917747656 (2018).

9 Walsh, C. G. *et al.* Enabling Open-Science Initiatives in Clinical Psychology and Psychiatry Without Sacrificing Patients' Privacy: Current Practices and Future Challenges. *Advances in Methods and Practices in Psychological Science* **1**, 104-114, doi:10.1177/2515245917749652 (2018).

Competing interests

The authors declare no competing interests.

Acknowledgments

Chris Hartgerink (Liberate Science GmbH) also contributed to this work and was originally included as a co-author. He opted to withdraw from the author list because this piece could not be published under an open access license. This work was supported by a Consolidator Grant (726361 IMPROVE) of the European Research Council (ERC) and an Open Science Fund grant from the Dutch Research Council (NWO 203.001.155). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Box 1. *Common reasons for increased risks of re-identification in psychology datasets*

- Presence of direct identifiers (e.g., MTurk Worker IDs, full names) that allow connecting with other demographic data and geographic region.
- Sharing of signed informed consent forms with full names and dates linking to the larger dataset
- Recording of IP addresses and approximate geographical coordinates alongside demographics such as age, ethnicity, gender, relationship status, and profession that allow for triangulation of some participants.
- Inclusion of initials or first names of participants alongside information about the sampling scheme (e.g., psychology freshmen in a given university cohort)
- Exact dates of birth or alongside other demographic information and sampling scheme
- Inclusion of potentially triangulating familial and relationship data (relationship status, number of siblings, parental professions)
- Demographic or other information that is outlying or uncommon in the sample (e.g., a middle-aged participant in a student sample, a unique country of birth in an organizational sample)

Box 2. *Measures to lower identification risks*

- Avoid collecting direct identifiers that are not necessary for the study
- Remove from shared data indirect identifiers that can be linked for triangulation (birth dates, zip codes, geographic information, initials) or ensure that they cannot be cross-referenced (by de-linking files).
- Delete or aggregate in higher-order groupings outlying demographics or other uncommon characteristics that could lead to identification.
- As a rule of thumb, ensure that each demographic cross-tabulation includes at least five participants.
- Create new larger categories of (demographic) variables such that a few or single cases do not stand out
- Shuffle some data if needed to ensure privacy or use differential privacy methods to do so
- Carefully consider rich qualitative (text) data or recordings for identifying information, and safely store these
- Use pseudonymization, encryption, and safe storage when direct identification is truly needed.
- Clearly indicate to participants how privacy in data is to protected and always obtain informed consent for sharing data
- Ask a trusted colleague or co-author to consider privacy risks before opening up the data