Privacy Protection in the Era of Open Science

Jelte M. Wicherts[1], Richard A. Klein[1], Sofie H. F. Swaans[1], Esther Maassen[1], Andrea H. Stoevenbelt[1,2], Chris Hartgerink[3], Victor H. B. T. G. Peeters[1], Myrthe de Jonge[1], & Franziska Rüffer[1]

[1] Tilburg University

[2] University of Groningen

[3] Liberate Science GmbH

Author note

Acknowledgements. This research was supported by a Consolidator grant from the European Research Council (ERC grant IMPROVE no. 726361). We thank Sarah Zheng for help in data collection.

Conflict of interest statement: The authors declare that no competing interests exist.

Correspondence concerning this article should be addressed to Jelte M. Wicherts, Tilburg University, Department Methodology and Statistics, PO Box 90153, 5000 LE Tilburg, The Netherlands. E-mail: j.m.wicherts@tilburguniversity.edu

Abstract

Given the many benefits of sharing data, an increasing number of psychological researchers publicly share the data underlying their research via online repositories. While undoubtedly a positive scientific development that enables greater verification and data re-use, it is important to protect the interests and confidentiality of research participants while doing so. This is particularly relevant when studying sensitive topics, for example related to health, religion, politics, and sexual behaviors. We systematically assessed the risk of identification of individual participants in 2,169 psychological datasets shared alongside articles published in *Psychological Science* from 2014 - 2019, *Judgment and Decision Making* from 2011 - 2014, and *PLOS ONE* from 2013 - 2015. Results show that individuals could be readily identified by names, IP addresses, web identifiers (email, MTurk Worker ID), birth dates, or ZIP codes and initials combined with other demographic variables in 114 (5.3%) of the datasets. An additional 94 datasets (4.3%) included (often unnecessary) demographic information that posed some re-identification risk. Moreover, of these datasets with identifying or potentially identifying data, 110 (53%) also contained data considered sensitive according to the GDPR. The majority of cases presenting privacy risks could have been prevented through simple procedures to de-identify datasets without sacrificing valuable information or transparency. We offer practical guidance to improve privacy protections in this transitional period towards greater data openness.

Keywords: Data sharing, open data, privacy, research ethics, open science

Word count: 217

Privacy Protection in the Era of Open Science

The open science movement has forged a new era of science characterized by greater transparency, more comprehensive data use, and increased opportunities for error detection and correction. The general public, science funders, academic institutions, and scientists themselves increasingly promote and value the sharing of data alongside the publication of research in many scientific fields, including psychology. The benefits of open data include increased rigor and verifiability of research, and greater opportunity for re-use of data by other researchers (Nosek et al., 2015; Wicherts, 2013).

At the same time, it is essential to protect the privacy of research participants, particularly in light of the increase in size and breadth of data collections by scientists, governments, (tech) companies, and others. Besides the ethical standards that stipulate care for privacy and confidentiality, most countries now have explicit laws to protect the privacy of their inhabitants. Scientists in the US are often bound by the Health Insurance Portability and Accountability Act (HIPAA) when they collect data involving human participants. Meanwhile, the General Data Protection Regulation (GDPR) that went into effect in 2018 puts regulatory oversight on the use of personal data for researchers based in Europe or researchers who make use of data collected in the European Union. The GDPR, in particular, is quite strict when it comes to the processing of personal data, but appears largely focused on use of (big) data by large tech companies. Although the GDPR allows scientific organizations to submit codes of conduct for formal approval, there is currently little jurisdiction on how psychological researchers handle personal data. Therefore, in the current research landscape much of the responsibility for navigating ethical and legal hurdles associated with sharing data is placed on individual researchers. As “open data by default” is a recent development, it is unclear whether researchers have enough training or knowledge about how to do so while protecting participant privacy.

The present project examines open datasets associated with a broad selection of research articles in psychology to assess privacy risks according to HIPAA and GDPR definitions. We scrutinize open datasets published alongside manuscripts in *Psychological Science*, *Judgment and Decision Making (JDM)*, and *PLOS ONE* with two goals: First, assess the risk of participant re-identification, or the risk that data available in these public datasets could be tracked back to an individual respondent (Meyer, 2018; Walsh et al., 2018). Second, we examined whether the datasets contained "sensitive" data in the context of the European GDPR.

Methods

The present report combines multiple efforts to assess data sensitivity and re-identification risk across three journals: *Psychological Science*, *PLOS ONE*, and *Judgment and Decision Making*. These three journals are prominent in their respective areas, serve different psychological subdisciplines, and show relatively high rates of data sharing (Federer et al., 2018; Kidwell et al., 2016; Nuijten et al., 2017). Because these are independent efforts combined into a single report, the years sampled sometimes differs between journals. However, the overall methodology was always quite similar: We identified articles with associated open data, and then had coders systematically assess whether the datasets contained sensitive or identifying information.

We examined all 430 articles published in *Psychological Science* that were granted open data badges from 2014, the year badges were introduced, through 2019. We excluded articles that used non-human participants, meta-analyses, and articles reporting analyses of secondary data. From the remaining articles, we were able to locate and access data from 398 (96%), yielding 1,443 datasets. Of these, 104 datasets were not feasible to code (e.g., undecipherable data labels or too many data files) leaving a total of 1,303 datasets that we assessed from *Psychological*

Science. In addition to assessing data availability, re-identification risk, and data sensitivity, we also recorded the nature of the sensitive data.

To cover additional topics and collect data from multiple journals with various procedures, we additionally assessed identification risk and data sensitivity in articles from *Judgment and Decision Making* from 2011 – 2014, and a sampling of *PLOS ONE* articles from 2013 – 2015. Both of these journals have policies mandating open data when possible (Federer et al., 2018; Nuijten et al., 2017). We again excluded articles that used non-human participants, meta-analyses, and articles reporting analyses of secondary data, and considered only articles with open data that we were able to access. This yielded 166 articles and 436 datasets from *Judgment and Decision Making*, and 215 articles and 430 datasets from *PLOS ONE*. Combined, we report below the results from examining 749 articles and 2,169 datasets with available open data. Given the wide sub-disciplinary coverage of *PLOS ONE* and *Psychological Science* and the sampled period, we consider our sample to be broad and roughly representative of datasets shared by early adopters of data sharing in psychology.

Determining identifiability

Given the fairly broad definition of personal data in the GDPR (Purtova, 2018), we used the identifiers from the US Health Insurance Portability and Accountability Act (HIPAA) to assess identifiability of research participants or respondents in the 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 number, certificate or license number, vehicle numbers, Web URLs, Internet Protocol (IP) addresses, finger or voice prints, photographic images, and other characteristics that could uniquely identify an individual.

For the latter somewhat broadly defined identifier, we employed the following rule: could at least one participant be uniquely identified given the information on the sampling scheme and some demographic information? For instance, if a study recruited undergraduate students from a given university and the data listed gender, age, ethnic origin or race, duration of stay in a country, or country of birth, it would not be too hard to re-identify some of the participants by crossing the given information. We scored such datasets as posing a medium risk of re-identification. Readers may be surprised at how easily this sort of information can identify individuals (El Emam, Jonker, Arbuckle, & Malin, 2011; Sweeney, 2002). For example, 87% of American voters could be identified merely from zip code, gender, and date of birth (Sweeney, 2000). For articles with high re-identification risk (e.g., direct identifiers according to HIPAA), we alerted corresponding authors to these risks via email.

Determining sensitivity

We also scored each dataset for whether it contained any sensitive data. Here we used the definition of sensitive data given in the GDPR, which includes information on racial or ethnic origin, political opinions, religious or philosophical beliefs, genetic or biometric measures, trade union membership, criminal convictions, health related data, and someone's sexual activities or sexual orientation. For the *Psychological Science* articles from 2014 – 2019 (but not for articles in the other two journals), we categorized these variables into the following categories: (1) race and/or ethnic origin, (2) political opinions and/or political affiliation, (3) religious beliefs and/or religious affiliation, (4) sexual preference and/or sexual behaviors, and (5) health and/or biological data. The latter category included any information on disease, psychopathology, measures of substances, and mood measures bearing on psychopathology. We also classified datasets as "potentially sensitive" if they contained data not specifically highlighted as sensitive in the GDPR, but that, by our judgement, could reasonably be considered sensitive.

Although we combine data from separate coding efforts, the general method was essentially identical.[1] An initial coder checked each dataset for re-identification risk and data sensitivity following the same HIPAA and GDPR definitions described earlier. A second coder verified each coding or indicated (dis)agreement. In cases of disagreement, the two coders discussed with each other and tried to reach an agreement. All cases with initial disagreement (even if agreement was reached after discussion) were reviewed and ultimately ruled on by one of the senior researchers (either the first or second author). For the *PLOS ONE* and *JDM* samples, we recorded the number of disagreements, and overall, the two assessors reached agreement 95% of the time. In addition, to assess the intercoder-reliability of our consensus-based coding scheme, we randomly sampled 50 articles (N = 148 datasets) and asked three independent coders to assess them for identification risk and data sensitivity. All three coders discovered datasets with re-identification risks (mean = 22.3, range = 15 - 27) and datasets with sensitive data (mean = 37, range = 29 - 40). Because the coding scheme used in the larger sample of datasets involved multiple assessors examining each repository and reaching consensus in case of discrepancies, we compared the "majority vote" in the independently coded sample to our actual ratings.[2] Using weighted κ with equal spacing weights (Cicchetti & Allison, 1971) there was substantial

[1] The coding deviated slightly for *Psychological Science* datasets from 2014 – 2016. In this case, the last author coded all datasets for data availability, identification risk, and sensitivity. The first author subsequently considered all auxiliary information shared alongside the articles, checked the coding, and documented the types of variables that would render a dataset identifiable and sensitive. The second and third authors subsequently verified all codes, followed by a final check by the first author.

[2] More specifically, we used the median value for this comparison in case all three raters disagreed, but this occurred extremely rarely.

agreement between the original consensus coding and the majority vote drawn from the three independent coders on identification risk ($\kappa = 0.63$, $p < .001$) and high agreement on sensitivity ($\kappa = 0.83$, $p < .001$). We also examined the agreement between the three individual independent coders with Fleiss' κ and found low-moderate agreement for identification risk ($\kappa = 0.39$, $p < .001$), and moderate agreement for sensitivity ($\kappa = 0.58$, $p < .001$). This indicates that single coders sometimes overlook some identification risks (or sensitivity indicators) in the often large and complex and sometimes poorly described datasets and repositories. This strengthens our trust in our consensus coding scheme using multiple coders but also highlights the possibility that our scoring might have missed particular identifiers or sensitive data.

Results

In total, we assessed 2,169 datasets for identification risks and data sensitivity. Because this coding results in a dataset that is, itself, a privacy risk (e.g., directly identifying public datasets with identifying and/or sensitive information), we are unable to share the raw data.

Sensitive data

Across all included years and journals, 550 of the 2,169 datasets (25%) contained sensitive or potentially sensitive information as defined by the GDPR. We further classified the types of sensitive data found in the 2014 – 2019 *Psychological Science* subset (389 datasets containing sensitive data). Datasets were categorized as including sensitive data because they included information on race and/or ethnic origin (267; 69% of studies with some sensitive data), political preferences (119; 30%), health, mood, or biological variables (85; 22%), religion or philosophical beliefs (60; 15%), and sexual preference and/or sexual behaviors (27; 7%). Many of the datasets (77; 20%) were sensitive for multiple reasons (e.g., included information on both religious and political beliefs). In total, 153 datasets (39%) were categorized as sensitive only because of the

inclusion of race/ethnic origin, which is commonly done in studies ran in the US. If we, instead, consider only the remaining (non-race/ethnicity) categories as being sensitive, 236 of the datasets (18% of the *Psychological Science* subsample) would include sensitive information.

Identification Risk

Concerning privacy of research participants, across all years and journals, we found that 114 of the datasets (5.3%) posed a direct risk of identification (e.g., initials and location data, IP addresses, etc.), while an additional 94 (4.3%) contained potentially identifying data by crossing demographics. Of these identifiable or potentially identifiable datasets, 53% (110 cases) involved sensitive data.

We discuss a subset of 311 *Psychological Science* articles (from 2014 - 2016; representing 24% of all datasets coded from *Psychological Science*) in more depth to provide an overview of typical causes of re-identification risk. In this subset, 23/311 (7.4%) datasets contained directly identifying information. We discuss these cases in Table 1, where we randomly shuffled some of the information without loss of relevant information to protect the privacy of the researchers and their participants. The main identifiers in these open datasets were IP addresses combined with other identifying information (11 datasets), date of birth combined with location and other identifying information (7 datasets), initials combined with age and other identifying information (2 datasets), full name and date of birth (2 datasets), and first name combined with location and other identifying information (1 dataset). In none of these cases did we identify an overriding scientific rationale to include such information in the shared data. We identified 31 additional datasets (10%) in which there was no direct identifying information, but that we still considered to have some risk of re-identification. We discuss these cases in Table 2, again shuffling some potentially identifying information to protect researcher and participant privacy. Typically, these datasets included overly 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 these cases did we identify a substantive reason to include this information, while in other cases this information appeared to have been included without any clear rationale. Tables 1 and 2 also provide guidance on how to render these datasets more private with minimal loss of information.

Discussion

The benefits of open data for science and society are abundantly clear, but there are also risks associated with the sharing of data from participants or survey respondents, some of which are embedded in legal frameworks for privacy protection. Here, we systematically assessed a large number of open psychological datasets on risk of identification of participants. We found that there was a concrete risk of identification of participants in 5.3% of the datasets and some risk of identification of participants in another 4.3% of the datasets. Of the 208 datasets with concrete or possible risk of identifying individuals, 110 (53%) included some sensitive information as defined by the GDPR. This clearly contrasts with the 1,961 datasets that were sufficiently anonymized, of which 440 (22%) were sensitive; $\chi^2(1) = 90.50$, $p < .001$. This shows that datasets that include identifiable information also often include sensitive information. Although this would suggest that there is little privacy risk in about 90-95% of psychological studies from which data are openly shared by early adopters of open data practices, the cases with privacy issues appear to be mostly due to inattention by researchers to the issue. Most often there appeared to be little reason to collect the identifying information in the first place, let alone share them alongside the more relevant parts of the data that are useful for verification or later re-use of the data in subsequent research.

Arguably, we focused here on the low hanging fruit for anyone who would want to re-identify research participants in the shared datasets, namely easy identifiers such as birthdates

and IP addresses combined with other demographic information. In our assessment of medium risk of re-identification, our approach also entailed checking whether we would find it easy to identify some of the participants. For ethical (and legal) reasons, we did not attempt to actually do this. Nor did we use more advanced digital means to help us re-identify any of the participants, but we are well aware of the technical options to do so. Such algorithms often work with more concrete variables such as demographic variables or IP addresses, but they could readily handle other (psychological) variables too. Psychological researchers should be aware of the risks that these algorithms pose to the privacy of their research participants. We do note, however, that the amount of information collected in the context of psychological research that could potentially be misused to identify someone is dwarfed by the enormous systematic data collection that tech giants currently have in place for monitoring digital communication and use of the web.

The open science movement will no doubt continue to improve psychological science, but we need to be vigilant to new risks such as the privacy concerns raised by the present results. Importantly, these privacy concerns should not be cause to return to the days of keeping data private, or operating under a "share upon request" model in which many datasets were ultimately lost or otherwise unavailable (Wicherts, Borsboom, Kats, & Molenaar, 2006). Open data policies appear effective in increasing the proportion of datasets that are available to other researchers and re-usable, although there is still room for improvement (Hardwicke et al., 2018). Many datasets can be openly shared with little risk, and researchers are increasingly finding creative solutions to share data from more sensitive designs (Bishop, 2009; Gilmore, Kennedy, & Adolph, 2018; Joel, Eastwick, & Finkel, 2018; Levenstein & Lyle, 2018).

In short, our recommendations to ensure that openly shared data pose little risk to participant privacy are (1) always seek informed consent for sharing data,[3] (2) pre-register the study with a clear description of the variables needed to conduct the study, (3) do not collect identifiable information unless needed for the study or its logistics, (4) use safe storage with encryption for sensitive personally identifiable information, (5) delete easily identifiable data from the datasets, (6) create new larger categories of (demographic) variables such that a few or single cases do not stand out, (7) shuffle some data if needed to ensure privacy or use differential privacy methods to do so, and (8) ask a trusted colleague or co-author to consider the privacy risks before opening up the data. Researchers have a real responsibility, not merely an ethical but increasingly also a legal one, to protect the privacy of research participants.

[3] Also note that research participants do not appear to be dissuaded by open data provisions in consent forms as long as anonymity is guaranteed, and open data provisions do not appear to affect subsequent responding by participants (Cummings, Zagrodney, & Day, 2015; Eberlen, Nicaise, Leveaux, Mora, & Klein, 2019).

References

Bishop, L. (2009). Ethical Sharing and Reuse of Qualitative Data. *Australian Journal of Social Issues*, *44*(3), 255–272. https://doi.org/10.1002/j.1839-4655.2009.tb00145.x

Cicchetti, D. V., & Allison, T. (1971). A new procedure for assessing reliability of scoring EEG sleep recordings. *American Journal of EEG Technology*, *11*(3), 101–110.

Cummings, J. A., Zagrodney, J. M., & Day, T. E. (2015). Impact of Open Data Policies on Consent to Participate in Human Subjects Research: Discrepancies between Participant Action and Reported Concerns. *PLoS ONE*, *10*(5). https://doi.org/10.1371/journal.pone.0125208

Eberlen, J. C., Nicaise, E., Leveaux, S., Mora, Y. Lã., & Klein, O. (2019). Psychometrics Anonymous: Does a Transparent Data Sharing Policy Affect Data Collection? *Psychologica Belgica*, *59*(1), 373–392. https://doi.org/10.5334/pb.503

El Emam, K., Jonker, E., Arbuckle, L., & Malin, B. (2011). A Systematic Review of Re-Identification Attacks on Health Data. *PLoS ONE*, *6*(12), e28071. https://doi.org/10.1371/journal.pone.0028071

Federer, L. M., Belter, C. W., Joubert, D. J., Livinski, A., Lu, Y.-L., Snyders, L. N., & Thompson, H. (2018). Data sharing in PLOS ONE: An analysis of Data Availability Statements. *PLOS ONE*, *13*(5), e0194768. https://doi.org/10.1371/journal.pone.0194768

Gilmore, R. O., Kennedy, J. L., & Adolph, K. E. (2018). Practical Solutions for Sharing Data and Materials from Psychological Research. *Advances in Methods and Practices in Psychological Science*, *1*(1), 121–130. https://doi.org/10.1177/2515245917746500

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., … Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: Evaluating

the impact of a mandatory open data policy at the journal Cognition. *Royal Society Open Science*, *5*(8). https://doi.org/10.1098/rsos.180448

Joel, S., Eastwick, P. W., & Finkel, E. J. (2018). Open Sharing of Data on Close Relationships and Other Sensitive Social Psychological Topics: Challenges, Tools, and Future Directions. *Advances in Methods and Practices in Psychological Science*, *1*(1), 86–94. https://doi.org/10.1177/2515245917744281

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., … Nos, B. A. (2016). Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency. *PLOS Biology*, *14*(5), e1002456. https://doi.org/10.1371/journal.pbio.1002456

Levenstein, M. C., & Lyle, J. A. (2018). Data: Sharing Is Caring. *Advances in Methods and Practices in Psychological Science*, *1*(1), 95–103. https://doi.org/10.1177/2515245918758319

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

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … Yarkoni, T. (2015). SCIENTIFIC STANDARDS. Promoting an open research culture. *Science (New York, N.Y.)*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Nuijten, M. B., Borghuis, J., Veldkamp, C. L. S., Dominguez-Alvarez, L., van Assen, M. A. L. M., & Wicherts, J. M. (2017). Journal Data Sharing Policies and Statistical Reporting Inconsistencies in Psychology. *Collabra: Psychology*, *3*(1), 31. https://doi.org/10.1525/collabra.102

Purtova, N. (2018). The law of everything. Broad concept of personal data and future of EU data protection law. *Law, Innovation and Technology*, *10*(1), 40–81. https://doi.org/10.1080/17579961.2018.1452176

Sweeney, L. (2000). Simple Demographics Often Identify People Uniquely. *Carnegie Mellon University, Data Privacy Working Paper 3*. Retrieved from https://dataprivacylab.org/projects/identifiability/paper1.pdf

Sweeney, L. (2002). K-ANONYMITY: A MODEL FOR PROTECTING PRIVACY. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(05), 557–570. https://doi.org/10.1142/S0218488502001648

Walsh, C. G., Xia, W., Li, M., Denny, J. C., Harris, P. A., & Malin, B. A. (2018). 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*(1), 104–114. https://doi.org/10.1177/2515245917749652

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

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*(7), 726–728. https://doi.org/10.1037/0003-066X.61.7.726

Table 1. *Examples of common reasons for datasets to be labeled high risk of re-identification. Note: Inspired by causes in actual cases, but these examples are fabricated to avoid identifying any specific dataset.*

High-risk example 1. Data collected in-lab with first names of participants, while university and department could be inferred based on article text. Deleting column with first names would successfully anonymize the dataset.

High-risk example 2. Data collected online retained IP address and approximate geographical coordinates combined with demographic information such as age, ethnicity, gender, relationship status, and profession. Deleting the IP address and coordinates would render the potential population large enough such that the data would be anonymous. If geographic location was important to the research question, the coordinates could be replaced with a larger geographic region to accomplish the same goal.

High-risk example 3. Anonymized data files, but left in repository were scanned signed consent forms with full names and dates. These names could be resolved to rows in the dataset by matching participation date and demographics (age, ethnicity). Removing these consent forms would render the repository non-identifiable.

High-risk example 4. In-lab data collection where the location could be inferred from the article text and included date of birth and other demographics (ethnicity, gender, sexual preference, annual income). Removing date of birth or converting to a less specific variable (e.g., age in years) and ensuring no specific outliers in demographics (e.g., extremely high income) would render this dataset anonymous.

High-risk example 5. Online data collection retaining MTurk Worker ID as well as common demographics (age, ethnicity, gender) and geographic region. Dataset could be anonymized by removing the MTurk Worker ID.

Table 2. *Examples of common reasons for datasets to be labeled medium risk of re-identification. Note: Inspired by causes in actual cases, but these examples are fabricated to avoid identifying any specific dataset.*

Medium risk example 1. Dataset including extensive demographic information (age, gender, ethnicity, political preference, profession) as well as geographic region of data collection and exact date and time of participation. Removing some demographic variables or classifying them into broader categories, and setting an upper limit to the age variable, would sufficiently anonymize this dataset.

Medium risk example 2. In-lab data collection at specified university including common demographics (age, gender) as well as extensive free-response text prompts in which participants sometimes disclosed quite specific information (i.e., niche affiliations, family member who works as X at Y company). Detailed free response texts such as these are not necessarily identifying, but should be removed or vetted to ensure no identifying information is disclosed when combined with other information in the dataset.

Medium risk example 3. In-lab data collection at specified university with a narrow sampling frame (i.e. 1st year psychology students) with common demographics (age, gender, ethnicity) but free-response or low-frequency categories within those demographics (i.e., identifying as transgender, or reporting uncommon ethnic background for that region). In these cases, there is a risk that only a few individuals within the population match the demographics and could be identified. It may be necessary to remove those responses to ensure no individual could be triangulated based on the demographics.

Medium risk example 4. Online data collection within specific region with extensive demographics such as age, gender, ethnicity, native language, educational history, and income. A case like this would be more borderline, but we consider that there is some risk that such

extensive demographics could triangulate an individual participant. Removing or recoding variables to broader categories to eliminate outlier responses (e.g., rare native languages) would eliminate any risk without sacrificing informational value.

Medium risk example 5. In-lab data collection with common demographics (age, gender, major) as well as potentially triangulating familial and relationship data (relationship status, number of siblings, parental professions). In this case the researcher could ensure anonymity by removing unused demographic variables or recoding into higher-order groupings (i.e., only sharing the classification of parental professions used in analysis, not the raw responses).