Responsible Data Sharing: Identifying and Remediying Possible Re-Identification of Human Participants

Kirsten N. Morehouse¹, Benedek Kurdi², and Brian A. Nosek³, ⁴

¹ Department of Psychology, Harvard University
² Department of Psychology, University of Illinois Urbana–Champaign
³ Department of Psychology, University of Virginia
⁴ Center for Open Science, Charlottesville, Virginia, United States

Open data collected from research participants creates a tension between scholarly values of transparency and sharing, on the one hand, and privacy and security, on the other hand. A common solution is to make data sets anonymous by removing personally identifying information (e.g., names or worker IDs) before sharing. However, ostensibly anonymized data sets may be at risk of re-identification if they include demographic information. In the present article, we provide researchers with broadly applicable guidance and tangible tools so that they can engage in open science practices without jeopardizing participants’ privacy. Specifically, we (a) review current privacy standards, (b) describe computer science data protection frameworks and their adaptability to the social sciences, (c) provide practical guidance for assessing and addressing re-identification risk, (d) introduce two open-source algorithms developed for psychological scientists—MinBlur and MinBlurLite—to increase privacy while maintaining the integrity of open data, and (e) highlight aspects of ethical data sharing that require further attention. Ultimately, the risk of re-identification should not dissuade engagement with open science practices. Instead, technical innovations should be developed and harnessed so that science can be as open as possible to promote transparency and sharing and as closed as necessary to maintain privacy and security.

Public Significance Statement
The rapid and widespread adoption of open science practices has increased the transparency and reproducibility of psychological research. However, it also poses a pressing new challenge: How can researchers protect the anonymity of research participants in open data sets? The present article provides psychological scientists with broadly applicable guidelines and tangible tools so that they can continue to engage in open science practices without jeopardizing the privacy of participants.

Keywords: privacy, open science, data anonymity, re-identification, research integrity

Supplemental materials: https://doi.org/10.1037/amp0001346.supp

Kirsten N. Morehouse ¹

Brian A. Nosek was employed by the nonprofit Center for Open Science that has a mission to increase openness, integrity, and reproducibility of research. Benedek Kurdi is a member of the Scientific Advisory Board of Project Implicit, a 501(c)(3) nonprofit organization and international collaborative of researchers who are interested in implicit social cognition. Code and data related to this project can be obtained from the Open Science Framework at https://osf.io/r9vu/?view_only=a379582125964a7ad7a45621b24492db. Kirsten N. Morehouse played a lead role in conceptualization, formal analysis, investigation, methodology, project administration, software, validation, visualization, and writing—original draft and an equal role in writing—review and editing. Benedek Kurdi played a supporting role in conceptualization, methodology, validation, visualization, and writing—original draft and an equal role in writing—review and editing. Brian A. Nosek played a supporting role in conceptualization, methodology, validation, visualization, and writing—original draft and an equal role in writing—review and editing.

Correspondence concerning this article should be addressed to Kirsten N. Morehouse, Department of Psychology, Harvard University, 33 Kirkland Street, Cambridge, MA 02138, United States. Email: kirsten_morehouse@g.harvard.edu
The open science movement has increased the rate of public data sharing in the social sciences considerably. Services such as Dataverse, Zenodo, and the Open Science Framework (OSF; https://osf.io/) now host millions of openly available data files, many of which originate from research involving human participants. Standard ethical practices for managing data obtained from human participants mandate publicly accessible data to be anonymized by removing personally identifying information such as names and email addresses. However, even with this safeguard in place, standard anonymization techniques may be insufficient to protect individual identities for some kinds of data.

The reason for this threat of privacy is that even a few pieces of demographic information can uniquely identify an individual. Sweeney (2000) demonstrated that just three variables from the 1990 United States Census—gender, postal code, and date of birth—uniquely characterized 87% of the U.S. population. The consequence of this finding is striking; many individuals—like the only woman born on January 18, 1965, living in 02142 (Cambridge, Massachusetts) in April 1990—could easily be identified using this publicly available information (see Koot et al., 2010; Rocher et al., 2021, for similar demonstrations using data from the Netherlands and census data from 89 countries).

Three features make this risk of re-identification especially relevant to the social sciences. First, data sets in the social sciences often contain multiple pieces of demographic or geographic information (e.g., race/ethnicity, level of education, birthplace), referred to as quasi-identifiers. Each additional quasi-identifier multiplicatively increases the uniqueness of the combination of variables characterizing each participant. For example, if a data set includes gender (male, female, other) and birth date, then the number of unique combinations is more than one hundred thousand: three gender response options \( \geq 36,500 \) birth dates (given a 100-year range). If the data set also includes the U.S. postal code of residence (41,692 postal codes), then the number of unique combinations is over 450 million. This multiplicative property is why 87% of the U.S. population is uniquely characterized by only these three variables. And, of course, many openly available data sets contain several quasi-identifiers beyond these three, thus increasing the risk of re-identification even further.

Second, quasi-identifiers collected by social scientists exist in other public data sets. As a result, data sets can be linked to reveal sensitive and protected data. For example, in the United States, gender, postal code, and birth date appear in both voter registration information and medical records. This allowed Sweeney (2001) to re-identify purportedly anonymous health records (for a review of re-identification attacks on health data, see El Emam et al., 2011). Sweeney famously fixed the then-sitting governor of Massachusetts, William Weld, his health records while on the phone with him (Sweeney, 2001). More recently, quasi-identifiers have been used to re-identify Netflix movie ratings (Narayanan & Shmatikov, 2007), taxi trips in New York City (Douriez et al., 2016), and genomic information (Gymrek et al., 2013).

Third, data sharing is becoming the norm across research disciplines. For example, in the United States, the National Institutes of Health established a strong data-sharing policy beginning in 2023 (National Institutes of Health, 2020) and the Office of Science and Technology Policy issued a memo directing all federal agencies to develop and implement similar policies by 2026 (White House Office of Science & Technology Policy, 2022). Similarly, the European Parliament passed legislation to increase data sharing in the European Union (European Parliament, 2022) and require private sector data to be publicly available (European Parliament, 2023). As such, the amount of data sharing will be increasing considerably during the next decade, which makes the issue of re-identification even more pressing.

Given these considerations and regulations mandating the secure sharing of data (e.g., Health Insurance Portability and Accountability Act in the United States and General Data Protection Regulation in the European Union), managing re-identification risks should be a priority for the social sciences. This article offers guidance about how to reduce such risks while also making science more transparent and reproducible. We (a) review existing de-identification strategies, (b) provide a framework for assessing risk and determining appropriate data-sharing practices, and (c) introduce open-source tools, tailored for the social sciences.

---

1 Quasi-identifiers refer to any information that could exist in multiple data sets. In the Netflix example, movie name and rating information served as quasi-identifiers given that they were shared across data sets. Demographic and geographic information are common quasi-identifiers in social science data sets.
that reduce re-identification risks. In doing so, the present article provides a complete pipeline for assessing re-identification risk, identifying appropriate risk mitigation and data-sharing strategies, and implementing those strategies. It also complements other discussions related to data sharing (Hahnel & Valen, 2020; Jati et al., 2022; Wicherts et al., 2022a; Wilkinson et al., 2016), including intellectual property and data ownership (Dennis et al., 2019), the secure storage of data (Inukollu et al., 2014; Marx, 2013), and informed consent (Meyer, 2018).

Mitigating Re-Identification Risks: Existing Strategies

Researchers often protect participants’ identities by not collecting direct identifiers or by removing them from data sets before sharing. Direct identifiers are attributable to an individual or small set of individuals and therefore eliminate anonymity. They include information such as IP addresses, phone numbers, emails, and home addresses. Direct identifiers are collected by default by some survey platforms and are often collected in longitudinal studies or under other circumstances that require repeated interaction with participants.

Individual researchers and institutional review boards are attentive to the privacy concerns posed by direct identifiers. Worker IDs for crowdsourcing platforms such as Amazon Mechanical Turk (MTurk) qualify as direct identifiers at some institutions (e.g., Harvard University; https://cuhs.harvard.edu/files/cuhs/files/hua_irb_newsletter_may_2021.pdf), and when direct identifiers need to be collected, some institutional review boards conduct a more intensive review process. Moreover, researchers are often instructed to keep direct identifiers confidential and separate from publicly shared data or even from data used internally for analysis. In addition to removing direct identifiers, another common risk mitigation strategy is pseudonymization. Pseudonymization involves replacing direct identifiers with pseudonyms or aliases. For instance, within health care databases, patient names can be replaced with unique identifiers (Neubauer & Heurix, 2011). Despite these safeguards, researchers sometimes inadvertently release identifiable information because of incomplete assessment of data sets prior to sharing (Lease et al., 2013; Wicherts et al., 2022b). This suggests that there is room for improvement on basic standards for data review and security prior to making data openly available.

However, even if implemented fully and correctly, these solutions do not protect against the re-identification risk associated with quasi-identifiers. At the same time, there are methods to identify and reduce the risk of re-identification developed in computer science that can be adapted for use in the social sciences. Sweeney (2002b) provided a framework for data protection called k-anonymity. This framework identifies re-identification risks by calculating the number of individuals in a population that share the same combination of quasi-identifiers, such as {White, woman, born on May 12, 1990}. If any single individual is characterized by a unique combination of quasi-identifiers, then the level of k-anonymity for the data set is $k = 1$. By contrast, if each combination of quasi-identifiers is shared by three or more individuals, then $k = 3$.

The higher the level of k-anonymity, the less concern there is about the risk of a particular participant becoming identifiable. k-anonymity is conservative in focusing on the most unique individual in the data set rather than the average degree of anonymity, but this is appropriate for social science data-sharing circumstances under which privacy is expected for all participants, not just the average participant. Indeed, participants with rare combinations of quasi-identifiers, such as individuals who belong to minority racial/ethnic groups, are typically the most vulnerable to re-identification. This disparate risk highlights the importance of adopting a conservative stance, rather than a focus on the average person in the data set.

This framework can also be used to reduce the risk of re-identification. Researchers can specify a minimum level of k-anonymity such that a data set is deemed to be “de-identified” if each combination of quasi-identifiers is shared by at least $k$ individuals. For example, a researcher could specify that every combination of quasi-identifiers needs to be shared by at least five participants ($k = 5$). This constraint increases privacy because it reduces the risk that any one individual can be directly identified.

To illustrate, imagine Jane Doe who is 35 years old and lives in the postal code 02138. If Jane is the only individual with that combination of quasi-identifiers in the population, then Jane can be identified by linking this information with other data sets that include directly identifying information (e.g., her name, which may appear in voter registration data). By contrast, if five individuals share that combination of quasi-identifiers {woman, 35, 02138}, then identifying Jane becomes more difficult; without any additional
information, there is only a one in five chance that the individual is Jane. Of course, mitigating re-identification risk is not as simple as specifying a minimum level of $k$-anonymity. Adopting a level of $k$-anonymity does not change the underlying population. To reach a given level of $k$-anonymity, researchers must apply techniques to reduce the risk of re-identification. The following sections provide heuristics for assessing re-identification risk, determining appropriate data-sharing strategies, and implementing risk mitigation strategies.

**Heuristics for Assessing Re-Identification Risk**

There is no universally applicable standard for quantifying re-identification risks. This heterogeneity arises because perceived risk and risk tolerance vary across research contexts and participant populations. As an example, a researcher may have a higher risk tolerance for a data set that includes adult participants than a data set that includes minors or individuals from marginalized communities. Nevertheless, there are heuristics that can alert researchers to potential risks and prompt closer examination of whether and how to enact additional data protection steps. Risk assessment involves considering both (a) the potential for re-identification and (b) the consequences of re-identification.

**Assessing Re-Identification Risk**

In the previous section, $k$-anonymity was introduced to quantify re-identification risks. Here, we outline the features of data sets, variables, and metadata that are associated with a higher risk of re-identification (see Figure 1). For example, the presence of multiple quasi-identifiers is associated with higher risk because each additional quasi-identifier multiplicatively increases the distinctiveness of the set of variables that define each participant (see Supplemental Material for a demonstration). These features are offered to provide researchers with an intuition about the level of risk without calculating $k$-anonymity.

As a starting rule of thumb, if the data set has $0–1$ of the higher risk conditions present, then the data set may be low risk; if $2–3$ or more conditions are present, the data set may be moderate risk; and if four or more conditions are present, the data set may be high risk. These heuristics could be too conservative or too liberal for clarifying actual risk in a data set. As such, they are not intended to be sufficient for decision making; they are merely intended to initiate the review of re-identification risk.

**Assessing the Consequences of Re-Identification**

In addition to the potential for re-identification, risk assessment involves examining the anticipated consequences of re-identification. For instance, the consequences of re-identification are more severe if a data set includes information about participants’ mental health than if it includes reaction times to mundane stimuli in an experiment about visual processing. As a rule of thumb, three levels of sensitivity are provided: negligible, substantive, and prohibitive. The decision-maker should have a lower tolerance for re-identification risk if the sensitivity of the data set is substantive or prohibitive.

A data set of negligible sensitivity is one that would not cause harm or only nominal harm if the identity of research participants was to be revealed. The hypothetical experiment on processing low-level visual features of an image is an example. A data set of substantive sensitivity is one that could cause discomfort if released. For example, many individuals would not want their purportedly anonymous responses about their emotion regulation techniques, life satisfaction, or policy preferences (e.g., support for immigration) to be revealed publicly, even if such disclosure did not have highly negative repercussions.

A data set of prohibitive sensitivity is one that contains highly sensitive, privileged, or confidential data. Some data, like health information, are widely understood as being highly sensitive and prohibited from being both identifiable and publicly shared in some countries (e.g., Health Insurance Portability and Accountability Act in the United States and General Data Protection Regulation in the European Union). However, there are many other kinds of data that might present risks for individuals if they were to be identified. Revealing data such as financial records and self-reported risky behavior could have negative external implications for participants.

---

2 The likelihood that U.S. residents can be uniquely identified by their [gender, postal code, date of birth] can be quantified at https://aboutmyinfo.org/identity. For a demonstration that includes additional quasi-identifiers and quantifies uniqueness outside of the United States, see https://cpg.doc.ic.ac.uk/observatory/take-the-quiz.
Moreover, some data may be prohibitively sensitive in some cultures or contexts but only nominally or substantively sensitive in others. A participant might reveal identities, express political views, or hold religious convictions that are controversial or stigmatized in certain cultures or contexts but widely accepted in others. For example, sexual orientation is a legally protected class in some countries but punishable by death in others (Carroll, 2016; Pew Research Center, 2023). In general, if the data collected can cause personal (e.g., legal repercussions), social (e.g., ostracization by family or friends), financial (e.g., identity theft), or professional (e.g., loss of job) harm, then the data should be treated with caution. Given that there may be sensitivities that are unanticipated and understood only after the risk is realized, it is prudent to presume higher sensitivity than is apparent so that a greater emphasis is placed on reducing re-identification risk.

After considering both re-identification risk and the potential consequences of re-identification, there will be some circumstances for which the risks are so minimal that the data can be shared publicly without further preparation. There will also be some circumstances for which the risks are so high that there are no conditions under which the full data can be shared publicly. In the latter case, we recommend that researchers either (a) share only portions of the data set that are undoubtedly anonymized or (b) archive the data in a manner that maintains strong security and, if appropriate, offers mechanisms for sharing by permission. For example, data can be stored in a protected repository in which users must request access to the data after providing evidence of ethical review and agreeing to confidentiality terms (“controlled access with approval”). This practice is already followed by organizations such as the American National Election Studies that require users to comply with terms and conditions for the ethical use of the data and expressly forbid re-identification.

The remainder of this article provides guidance for data sets that fall between these two extremes—circumstances for which providing open data is possible but only after ensuring that re-identification risk is reduced sufficiently. Specifically, the following sections include a guide for determining appropriate data-sharing practices and de-identification techniques and introduce two open-source algorithms that can increase the protection of open data sets.

### Figure 1

**Features of Data Sets, Variables, and Metadata That Influence Risk of Re-Identification**

<table>
<thead>
<tr>
<th>Features of Datasets</th>
<th>Risk increases as the number of quasi-identifiers increases</th>
<th>Why? Each additional quasi-identifier multiplicatively increases the uniqueness of the combination of variables categorizing each subject</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple Quasi-Identifiers</td>
<td>Risk increases when other data with the same quasi-identifiers from the same population may exist</td>
<td>Why? Having shared quasi-identifiers increases the potential of joining a dataset with others that have sensitive data and other variables that increase re-identifiability</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Features of Variables</th>
<th>Risk increases as the number of levels within a quasi-identifier increases</th>
<th>Example: Postal code (higher risk) versus binary gender (lower risk)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of levels</td>
<td>Risk increases when certain levels of the quasi-identifier are low-frequency</td>
<td>Example: Gender when non-binary gender options are included (higher risk) versus binary gender (lower risk)</td>
</tr>
<tr>
<td>Low-frequency incidence</td>
<td>Risk increases when reliability of a quasi-identifier is perfect or near perfect</td>
<td>Example: Political party membership (higher risk) versus political ideology (lower risk)</td>
</tr>
<tr>
<td>Reliability of measurement</td>
<td>Risk increases when responses to quasi-identifiers cannot change over time</td>
<td>Examples: Nation of birth (higher risk) versus nation of residence (lower risk); Vote choice in 2020 election (higher risk) versus political party membership (lower risk)</td>
</tr>
</tbody>
</table>

| Features of Metadata | Risk increase when there are metadata that shrink the population from which the sample is drawn. If these metadata are known, then they should be incorporated into re-identification risk assessment. | Examples: Time period of data collection, location of data collection, context of data collection, population from which the sample was drawn, and whether the dataset is linked to other data about the participants (e.g., longitudinal study, merged with administrative data) |

**Note.** See the online article for the color version of this figure.
A Practical Guide to Determining Appropriate Data-Sharing Practices

After researchers have determined that a data set may be at risk of re-identification, the next step is to determine what data-sharing practices are most appropriate. The present section provides a decision tree and companion online application to guide these determinations (https://responsible-data-sharing.shinyapps.io/MinBlur).

The decision tree consists of four questions (see Figure 2). The first two questions draw from previously discussed concepts and are integrated into the decision tree to enable the online application to function as a stand-alone resource. This free online application allows researchers to walk through the decision tree using their own data set, apply de-identification techniques, and determine whether the techniques applied sufficiently reduced the re-identification risk while maintaining the usefulness of the data set. Researchers who answer “yes” to at least the first two questions are encouraged to use this tool. Tutorials for common data set types are also available on the OSF (https://osf.io/c7q8h).

The first question is “Does the data set include quasi-identifiers or sensitive information?” As outlined above, quasi-identifiers are pieces of information that could exist in multiple data sets. The most common class of quasi-identifiers collected by social scientists include demographic (e.g., gender, age, race/ethnicity) or geographic (e.g., postal code) information, but behavioral information such as browsing history, social media interactions, and product preferences could also serve as quasi-identifiers if they are present in multiple data sets. Sensitive information includes any information with the potential to cause personal, social, financial, or professional harm. If the data set does not include either, then the data set may be shared publicly, without any additional modifications.

However, if the data set does include quasi-identifiers or sensitive information, then researchers should proceed to the second question: “Is the re-identification risk of the data set moderate to high?” The prior section suggests first counting the number of risk-enhancing features of the data set, variables, and metadata. In most cases, the presence of 0 or 1 risk-enhancing features will be low risk for re-identification, but it is possible under rare circumstances that even a single risk factor is associated with nontrivial risk. If two or more risk-enhancing features are present, the re-identification risk deserves further scrutiny. In many cases, that assessment will reveal that risk is low. For example, a data set may include gender assigned at birth, which is a highly reliable measure and not transient (see Figure 1). This means that the data set has two risk-enhancing features for one quasi-identifier. But, if there are no other risk-enhancing features, then the data set would be very low risk for sharing. If a more in-depth risk assessment reveals that re-identification risk is moderate to high, then the researcher should proceed to the third question.

Figure 2

Decision Tree for Determining Appropriate Data Practices

Note. Q = question. See the online article for the color version of this figure.
The third question is “Can de-identification techniques reduce the risk of data sharing?” De-identification techniques can be applied to data sets with one or more of the following properties: (a) granular geographic or date information (e.g., postal code, birth date), (b) quasi-identifiers that are not critical to the research question, and (c) quasi-identifiers that can be recoded without reducing their usefulness (e.g., converting birth date to age). For many cases, the straightforward act of converting a variable such as birth date or postal code to a more general or transient variable like age or state is sufficient to substantially reduce risk. If de-identification techniques cannot render the data set low risk, then the data set should not be shared openly. Instead, more secure alternatives, such as controlled access with approval, data enclaves in which users can conduct analyses without accessing the data, and open sharing of data-adjacent materials (e.g., code, codebooks, metadata, descriptive summaries), should be considered.

If these techniques or algorithms can render the data set low risk, then researchers should proceed to the fourth and final question: “Does the lower risk version of the data set maintain the values of transparency and sharing?” That is, would researchers be able to make sufficiently similar inferences from the lower risk version of the data set than the original data set? If so, then sharing the lower risk version of the data set advances the values of openness and maintaining participant privacy simultaneously. If the data are no longer useful after reducing the identification risk, then researchers should either (a) adjust the de-identification techniques applied and then reassess the utility of the data set or (b) consider less open archiving and sharing alternatives such as controlled access with approval.

MinBlur and MinBlurLite: Open-Source Algorithms to Identify and Remedy Risks to Privacy

Computer scientists have developed several techniques to increase the protection of data sets against the risk of re-identification. However, many of these techniques rely on opaque statistical methods that are difficult to interpret (e.g., recurrent neural networks; Dernoncourt et al., 2017; Liu et al., 2017) or on algorithms that apply arbitrary partitioning processes to quasi-identifiers (e.g., optimal k-anonymization; Bayardo & Agrawal, 2005). For example, if a data set has many participants under 40, the algorithm might produce a new age variable with the following levels, which are both highly unequal in their span and difficult to interpret substantively: <18–19>, <20–21>, <22–25>, <26–29>, <30–40>, and <41–79>.

Indeed, these techniques are often ill-suited for use by social scientists who seek to preserve the usefulness of demographic information in open data. The minimal blurring algorithms, MinBlur and MinBlurLite, are introduced here as innovations in algorithms that apply conceptual frameworks like k-anonymity to be fit-for-purpose for data sharing in social and behavioral research. However, they are compatible with any data set with quasi-identifiers, regardless of field, topic, population, or size. These open-source algorithms apply intuitive de-identification techniques (see Figure 3; Sweeney, 2002a) while maximizing four criteria: (a) data retention, (b) flexibility, (c) transparency, and (d) data fidelity.

First, data retention refers to the practice of preserving as much data as possible in the to-be-shared data set. MinBlur and MinBlurLite enable researchers to maximize data retention by blurring (removing) values for specific rows (participants) rather than blurring entire columns (variables). For example, rather than removing every participant’s age from the data set, these algorithms would blur only rare ages, as these values have a relatively high potential to re-identify individuals. Blurring occurs iteratively to maximize the number of retained values.

Second, flexibility refers to the researcher’s ability to change the algorithms’ inputs, depending on their research question. That is, users can apply domain knowledge to flexibly select which quasi-identifiers should be blurred and in what order. This ensures that information most central to the research is prioritized for retention.

For example, imagine a data set with five quasi-identifiers: QI = {race, state, gender, age, political ideology}. If a researcher is interested in exploring the effect of age on a memory task, then participants’ age will be more informative than their race or political ideology. As such, race and other less relevant variables will be blurred first, whereas age information will be maximally preserved. This flexibility is crucial to social scientists who investigate demographic and geographic variation. If MinBlurLite or MinBlur was to randomly blur columns or individual values, then the resulting output might satisfy the desired level of k-anonymity but distort the data set for purposes of demonstrating reproducibility or extensions of its primary purpose.

Third, transparency refers to the ease with which users can implement and interpret MinBlur and MinBlurLite. Both algorithms require users to specify only three inputs, and the output of both algorithms is intuitive, making the de-identification process highly transparent. Users can easily ascertain (a) which values have been blurred and (b) which rows (if any) have been suppressed. This enables users to comprehend the adjustments to the data and determine whether it satisfies their needs. If the initial decisions result

3 While blurring removes specific columns (i.e., variables), suppression removes entire rows (i.e., individuals). Specifically, certain rows may be withheld from a data set if they are relatively rare and therefore at greater risk of re-identification. However, this practice tends to suppress data from nonmajority or disadvantaged individuals (Christ et al., 2022); as such, researchers should review suppressed output to ensure that data protection algorithms have not systematically removed certain demographic subgroups from the data set.
in undesirable adjustments to the data set, then users can update the inputs and rerun the algorithm.

Fourth, data fidelity denotes the level of similarity between the raw data set and the de-identified data set. If the data and statistical inferences drawn from the de-identified data set align with the raw data set, then data fidelity is high. Given that demographic and geographic variation are often of primary or secondary interest, this criterion is crucial for social scientists. Sharing data sets where key quasi-identifiers are substantially redacted or distorted decreases, or even fully eliminates, their usefulness and accuracy. MinBlur and MinBlurLite allow users to specify the blurring order to prioritize preserving the most relevant information.

Determining the true re-identification risk of a data set requires determining (a) the identifiable population from which the sample is drawn for which (b) there are other data that could be joined to the data set to enhance the identifiability of individuals. For example, if we knew that a sample of 520 adults was collected in July 2023 and included only sitting U.K. Members of Parliament (MPs), then the relevant population is U.K. MPs in 2023. Given that other public data of that population’s gender, age, and postcode of residence exist, all, or almost all, individuals in the data set could be re-identified easily. Similarly, if an article reports that the sample consisted of Harvard undergraduates in an introductory psychology course from Fall 2023, then quasi-identifying information (e.g., race/ethnicity, nationality) can pose an outsized risk of re-identification if there are other data sources available identifying Harvard undergraduates in 2023 because the effective population is relatively small.

On the other hand, if we knew that the sample of 520 U.K. adults was collected sometime in the 2010s from a public website with thousands of visitors daily, then there may be no identifiable population narrower than the U.K. population for which there are other data that could be joined to the data set. In that case, the k-anonymity of interest would be very high because the risk would be in reference to U.K. population data and with the added ambiguity of the date of data collection over a decade.

In most cases, social science data are samples of a much larger population, which means that the actual re-identification risk is a feature of variability in the population, not the sample. Thus, the minimum blurring algorithm (MinBlur) attempts
to approximate true re-identification risk by allowing users to provide demographic information about the population. Other reports (e.g., Ye et al., 2017) provide detailed information about how to create this population-level information. This information is customarily referred to as “joint distributions” because it captures probability distributions for two or more variables, such as gender and race/ethnicity in a population. For example, a joint distribution over gender, postal code, and birth date would specify the probability that a U.S. resident would be characterized by any given combination over the three variables given known population characteristics.

MinBlur can be accessed freely using the R scripts provided on the OSF (https://osf.io/dcx89/). Users provide a cleaned data set (see Figure 4 for suggested preprocessing guidelines) and relevant population-level demographics (i.e., joint distributions), and specify three inputs: (a) the requested level of _k_-anonymity (_k_), that is, the minimum number of participants sharing the same combination of quasi-identifiers; (b) quasi-identifiable variables; and (c) quasi-identifiable variables to blur. MinBlur assumes that some variables are more informative (i.e., relevant to a particular research question) than others. As such, MinBlur blurs quasi-identifiers in a user-specified order of least informative (blurred first) to most informative (blurred last).

MinBlur operates in two stages, visualized in Figure 5. First, MinBlur calculates _K_, the population-level _k_-anonymity. That is, based on the joint distributions provided by the user, MinBlur identifies the combinations of quasi-identifiers, for example, {woman, 35, 02138}, in the population that satisfy the requested level of _k_-anonymity. Rows in the sampled data set that match these combinations are subsetted into the “de-identified” data set. This approach maximizes the amount of data retained by limiting the scope of de-identification techniques. If a researcher determines that a data set is sufficiently protected when every quasi-identifier is shared by at least _K_ individuals in the population, then de-identification techniques need to be applied only to the rows corresponding to quasi-identifiers that are not _K_-anonymous.

The remaining rows that do not adhere to the user’s requested level of _k_-anonymity are subsetted into a secondary data set. In the second stage, MinBlur takes the secondary data set, blurs the first (and least informative) column, and then recomputes data set-level _k_-anonymity. The rows that newly satisfy the user’s requested level of _k_-anonymity are appended to the de-identified (_K_-anonymous) data set. MinBlur will repeat this process on the remaining rows until (a) all rows are _k_-anonymous or (b) all eligible columns have been blurred. The final output is the de-identified data set and the secondary data set, which includes all the suppressed rows.

The key advantage of MinBlur is its sensitivity to the demographic characteristics of the target population. However, there are occasions in which it is not feasible to use the population for risk assessment. Notably, calculating joint

### Figure 4
**Recommended Data Preprocessing Practices for De-Identified Data Sets**

<table>
<thead>
<tr>
<th>Practice</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Remove direct identifiers.</td>
<td>Remove all direct identifiers from the dataset. This includes variables such as IP addresses, crowdsourcing worker IDs, and email addresses.</td>
</tr>
<tr>
<td>2. Recode variables with over 5 levels.</td>
<td>When possible. For example, consider recoding a 14-level education variable such that similar values or values that do not tend to differentiate participants now belong to the same level (e.g., ‘Master’s degree’, ‘PhD’, ‘MD’, and ‘MBA’ could be collapsed into a single ‘Professional or Graduate Degree’ level).</td>
</tr>
<tr>
<td>3. When recoding is not possible, consider releasing a more aggregate variable.</td>
<td>For example, instead of releasing postal code information (a smaller unit), county or state information (larger units) could be released.</td>
</tr>
<tr>
<td>4. Remove sensitive information and irrelevant information that could identify participants.</td>
<td>For example, in a student sample, the date of participation could be used to deduce the semester of participation.</td>
</tr>
<tr>
<td>5. Consider what information may be accessible as metadata about the dataset.</td>
<td>If reporting or metadata makes clear the year and location of a data collection, then some demographics may be known about the data without being in the dataset. For example, if a report indicates that the data collection occurred using first year students at University of Virginia in 2018, then the population from which the sample is drawn is much smaller than if date and location information were not in the dataset.</td>
</tr>
</tbody>
</table>

*Note.* Users should complete relevant data preprocessing practices before using MinBlur or MinBlurLite. MBA = master of business administration. See the online article for the color version of this figure.
distributions for multiple variables can be difficult in practice. For example, it may be possible to determine the number of 30-year-old men in the United States (two quasi-identifiers) but difficult to determine the number of 30-year-old White men in the United States who were born in New Jersey, have an advanced degree, and are Jewish (five quasi-identifiers). As the number of quasi-identifiers in a data set increases, the availability of population data to compute the corresponding joint distributions declines.

MinBlurLite is introduced as a secondary algorithm for circumstances in which it is not feasible or necessary to use MinBlur. MinBlurLite can be accessed via the online application introduced above (https://responsible-data-sharing.shinyapps.io/MinBlur) or offline via the R scripts provided on the OSF (https://osf.io/dcx89). From a technical perspective, MinBlurLite is identical to MinBlur except that it calculates $k$-anonymity based on sample characteristics rather than population characteristics (see Figure 5). As such, researchers may decide to use MinBlurLite when (a) it is difficult to calculate joint distributions or the population is unknown; (b) metadata shrinks the effective population (e.g., data collected from psychology undergraduates at Harvard University in September 2023); (c) the population and sample are similar sizes (e.g., if 520 of 659 U.K. MPs participate); or (d) the quasi-identifiers produce sufficiently unique combinations that it is unlikely for these combinations to appear frequently in the population, for example, {White, woman, date of birth: January 2, 1998, 02141, master’s degree}.

Unlike MinBlur, MinBlurLite provides a solution for estimating $k$-anonymity within the data set itself. This means that MinBlurLite treats the data set as if the sample and population were identical for the purposes of estimating $k$-anonymity. In many cases, this provides a moderate-to-extremely conservative assessment of risk. That is, the level of risk indicated by the algorithm may be significantly higher than the actual risk in the population because the presence of a unique combination of quasi-identifiers in the data set need not imply that the combination of quasi-identifiers is unique in the entire population. For example, there may be only one 56-year-old woman from Tennessee with a PhD in the data set, but many individuals with the same combination of demographic variables in the population. For highly sensitive data, some conservatism may be reasonable given that external data can be linked to the data and reduce the level of $k$-anonymity. However, with less sensitive data, users can request lower levels of $k$-anonymity from MinBlurLite to provide a less conservative estimate.

To illustrate the implementation of MinBlurLite, consider an example data set where 98.4% of the data were uniquely characterized by just three quasi-identifiers: birth date, postal code, and gender. A researcher using MinBlurLite to de-identify the data set might specify that (a) the requested level of $k$-anonymity is 3 and (b) the following five variables are quasi-identifiers: gender, age, postal code, race/ethnicity, and education. Additionally, imagine that this study was primarily concerned with differences across race/ethnicity...
and secondarily concerned with differences across gender. As such, the user might (c) preserve these two quasi-identifiers (i.e., race/ethnicity, gender) and instead blur quasi-identifiers that are not of primary or secondary interest but that unnecessarily increase the re-identification risk of the data set (i.e., education, age, and postal code).4

With these three inputs, MinBlurLite is able to increase the share of k-3 anonymous example data from 0% to 100% without suppressing any rows (i.e., without fully removing any single participant from the data; see Supplemental Material). This is important for two reasons. First, it demonstrates how an entire data set, rather than only a subset of the already k-anonymous rows, can be shared while protecting the anonymity of participants. Second, because the most informative quasi-identifiers were not blurred, MinBlurLite can increase k-anonymity while retaining the ability for future researchers to investigate questions central to the original research.

Notably, if other variables are of specific interest to future researchers, then the same steps can be repeated with a different prioritization of variables and the data set reshared with the researcher. However, users should take care to ensure that other variables in the data set cannot be used to link different versions of the data set. For example, if unique participant identifiers or time stamps remain consistent across versions of the data set, then this information can be used to link multiple versions of the data set to each other. To prevent against this, we recommend (a) replacing static participant identifiers with random strings (generated for each data set), (b) generalizing or removing highly unique values (e.g., time stamps that include granular data such as seconds), and (c) requesting that secondary data sets are kept private.

MinBlur and MinBlurLite have four key advantages over related data privacy algorithms.

First, they are flexible and sensitive to domain knowledge. Unlike other algorithms that apply partitioning models to recode data, for example, Mondrian (LeFevre et al., 2006), or add statistical noise to increase privacy, for example, Incognito (LeFevre et al., 2005), researchers have full control over (a) when and how variables are generalized or recoded, (b) which variables are blurred, and (c) the blurring order.5 These parameters allow researchers to test different generalization and blurring strategies to achieve both de-identification and data reusability.

Second, MinBlurLite and MinBlur are fully transparent. That is, the de-identification techniques employed are directly visible in the output. Users can make informed decisions regarding which iteration of the algorithm to adopt.

Third, this transparency allows users to audit the algorithms. MinBlurLite returns both a data set that adheres to the user’s requested level of k-anonymity and a data set that includes the rows that would be suppressed from the de-identified data set. Users can (a) compute the percentage of data retained in the k-anonymous data set, (b) examine which levels of a given variable were blurred, (c) compare the demographic composition of the full data set to the demographic composition of the k-anonymous data set, and (d) explore demographic and geographic variability on the outcome variables.

Fourth, MinBlur and MinBlurLite are open source, and their inputs are easy to specify. Unlike other algorithms that are only specified visually or with pseudocode (i.e., a description of the algorithm that uses both natural language and programming-like constructs), for MinBlur and MinBlurLite, documented scripts and examples are publicly available (https://osf.io/rz9vu/). In fact, MinBlurLite is integrated within a website (https://responsible-data-sharing.shinyapps.io/MinBlur) that can be used without any coding experience. Similarly, the inputs needed to use these algorithms rely on the types of background knowledge that empirical researchers are likely to possess (e.g., which variables are substantively more important to the research question than others). By contrast, algorithms such as Datafly (Sweeney, 1997) require users to provide a parameter that captures the likelihood that each quasi-identifier will be used for linking, which may be difficult for users to specify, especially if they do not have extensive preexisting knowledge of re-identification.

Despite these advantages, MinBlur and MinBlurLite cannot guarantee participant anonymity or fully prevent against adversarial attacks. For example, a blurred value may be imputed using known correlations between variables, such as the correlation between age and education. MinBlur is also prone to homogeneity attacks. If every White 20-year-old man in the sample provided the same answer to a sensitive question, then the data are ostensibly released for every White 20-year-old man in the sample (Machanavajjhala et al., 2007). Additionally, this algorithm is not equipped to prevent background knowledge attacks, that is, the use of prior knowledge about an individual to identify them in a data set (Kifer & Gehrke, 2006; Machanavajjhala et al., 2007; Martin et al., 2007). For example, if you know that your colleague participated in a recent departmental survey, it may be possible to identify them using their demographic information.

General Discussion

Social scientists often confront a tension between two legitimate and important desires: the desire to be transparent,
share data, and conduct reproducible research, on the one hand, and the desire to protect the privacy of research participants, on the other hand. The rise of open science practices does not reduce the importance of maintaining security and privacy. In fact, by switching the default from closed to open, the need for social scientists to be more knowledgeable and skillful at managing privacy and security increases because there are more occasions to violate them inadvertently. In this article, we have provided a framework and open-source tools for social scientists to improve their data management practices and thereby meet the objectives of openness and privacy simultaneously.

Part of the solution to resolving the apparent tension between openness and security is to recognize that open versus closed data are not a dichotomy. For data sets with some risk of re-identification, there are practical steps that remove just enough data to minimize the risk of re-identification while simultaneously meeting the objectives of transparency, sharing, and demonstrating reproducibility. The decision-making workflow and tools presented here offer guidance for determining whether the use of such de-identification techniques is possible without undermining the value of data sharing. In some cases, it will not be possible to openly share a de-identified data set that is useful to other researchers. Even then, there are established options other than keeping the data set entirely closed. Such options include storing data in protected archives and sharing only when researchers meet standards for ethical sharing and data protection or using data envelopes to which researchers can submit queries to analyze the data but never gain access to the raw data itself.

As readers will have already recognized, technology and guided workflows for responsible data management and sharing cannot replace thinking and good judgment. Social scientists will confront scenarios that are not anticipated or addressed by the solutions offered here. When the existing solutions do not properly address the risks for a particular data set, it is best to put the existing solutions aside and focus on addressing the risks. The transition from closed to open science will uncover novel scenarios that the existing policy and practice landscape is not prepared to handle. Publicizing and working through these cases will help social scientists better navigate ethical data management and sharing.

Open Questions

For readers who are specifically interested in advancing policies and technologies for ethical data management practices, there are a variety of issues associated with this work that require additional attention. First, what level of $k$-anonymity can best protect participants’ identities while preserving the usefulness of open data sets? At present, $k = 5$ is often used as a benchmark (Ito et al., 2020; Wicherts et al., 2022a). Future work should explore the trade-offs between data fidelity and level of $k$-anonymity—a question that is crucial for social scientists who seek to maximize the usefulness of open data—and provide specific guidelines for $k$-anonymity (e.g., varying standards for data sets with different levels of sensitivity).

Second, to what extent should potential future data reuse influence de-identification decisions? For example, consider a data set that collected education and religion information in a routine demographic questionnaire but did not use these quasi-identifiers in any analyses. On the one hand, removing this information increases the privacy of a data set. On the other hand, removing this information limits the potential reuse of this data for alternative research objectives. For instance, future researchers cannot analyze the effect of religion on the dependent variable using the publicly released data set. Widespread adoption of de-identification practices in the social sciences will require more concrete guidance about how to navigate this trade-off. Nevertheless, intermediate solutions include (a) using algorithms that retain but deprioritize ancillary variables, such as MinBlur or MinBlurLite, when making de-identification decisions, and (b) releasing a codebook of all collected variables, including those that were removed from the open data set, and providing controlled access to a version of the data set that includes removed variables.

Third, do participants understand the risks of re-identification and, if so, does the perceived risk influence their willingness to provide accurate demographic and geographic information? Understanding participants’ perception and tolerance of risks to privacy is important because participants who provide informed consent to data sharing presumably do so under the assumption that their data are and will remain anonymous. For example, anecdotal evidence suggests that crowdsourcing workers who are concerned about data privacy report giving false data to protect their anonymity (Kandappu et al., 2015). As re-identifiability concerns become more widely shared, developing more specific data privacy guidelines for consent procedures will provide reassurance to participants about the risk management of their responses and encourage confidence that they can respond forthrightly.

Fourth, what role do and should participants’ preferences play in data release? The discussion thus far has primarily considered the creation of open data sets that protect participant privacy. However, there is a growing movement to consider providers of personal information—participants, in the case of research studies—as data owners (General Data Protection Regulation, 2016; Papadopoulou et al., 2015; Scassa, 2018). This movement has gained traction because researchers and participants may have competing interests. Researchers may seek to provide data that are de-identified but maximally complete and publicly accessible so that their analyses can be reproduced, and other researchers can use it easily. By contrast, in some cases, participants may value data fidelity less and therefore hold a more conservative stance toward data sharing. Future work should (a) explore this trade-off and consider features such as participant population
when determining how participant preferences could impact data release<sup>6</sup> and (b) articulate potential distinctions in data ownership rights, particularly between data generated for noncommercial research purposes and data that may be sold or possess other economic value.

Fifth, how can social scientists integrate data management and sharing practices into their workflow without taking on undue levels of new burdens? One challenge is that journal policies for data sharing come at the end of the research life cycle, often long after the researchers have been working with their data. Returning to the data and preparing it for sharing at a later date is highly burdensome and prone to error. As the funding policy landscape moves toward open data as the default, new opportunities will emerge to integrate data management and sharing practices into the daily workflow for researchers. Giving researchers tools to add metadata and evaluate risks during planning and collecting data could dramatically reduce the burden for researchers and increase the quality of decision making and data sets that are ultimately shared. Services like MinBlur and MinBlurLite could be integrated with platforms such as the OSF that support the research life cycle to address that burden. Moreover, journals and professional societies could play a more active role in de-identification so that individual researchers are not tasked with making unilateral decisions. For example, journals could provide data privacy guidelines akin to statistical reporting criteria or incorporate questions about re-identification risk in the review process.

**Conclusion**

An aphorism about ethical decision making is that it is rarely about right versus wrong, but rather about the conflict between two rights or between two wrongs. Few researchers disagree that sharing data and protecting participant privacy are both desirable. Challenges arise when those desirable outcomes pull the researcher toward different decisions. In the present article, we introduced some strategies and solutions to help resolve some of the apparent conflicts between data sharing and participant privacy and to help advance both of these important goals.

<sup>6</sup>An intermediate solution may involve more transparent consent forms, so that users can decline to participate if future data usage plans violate their privacy preferences. Meyer (2018) provided further guidance for the language of consent forms.

**References**


