EUROPEAN JOURNAL OF PSYCHOTRAUMATOLOGY
2025, VOL. 16, NO. 1, 2499296
https://doi.org/10.1080/20008066.2025.2499296

REVIEW ARTICLE

OPEN ACCESS

Sharing traumatic stress research data: assessing and reducing the risk of re-identification

Nancy Kassam-Adams [a], Kristi Thompson [b], Marit Sijbrandij [c] and Grete Dyb [d]

[a]Center for Injury Research & Prevention, Children's Hospital of Philadelphia, Philadelphia, PA, USA; [b]Western Libraries, Western University, London, Canada; [c]Department of Clinical, Neuro-, and Developmental Psychology, Amsterdam Public Health Institute, WHO Collaborating Center for Research and Dissemination of Psychological Interventions, Vrije Universiteit, Amsterdam, The Netherlands; [d]Norwegian Center for Violence and Traumatic Stress Studies, Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway

ABSTRACT

Background: FAIR Data practices support data sharing and re-use and are essential for advancing science and practice to benefit individuals, families, and communities affected by trauma. In traumatic stress research, as in other health and social science research, ethical, legal, and regulatory frameworks require careful attention to data privacy. Most traumatic stress researchers are aware of basic methods for de-identifying/anonymising datasets that are to be shared. But our field has not generally made use of systematic, data analytic approaches to reduce the risk of re-identification of study participants or disclosure of personal or sensitive information.
Objective: To facilitate safe and ethical data sharing by better preparing traumatic stress researchers to systematically assess and reduce re-identification risk using contemporary data analytic methods.
Method: In two case studies using publicly available trauma research datasets from international, multi-language projects, we applied a systematic approach guided by the Checklist for Reducing Re-Identification Risk in Traumatic Stress Research Data.
Results: For each case study dataset, we identified specific recommended actions to further reduce the risk of re-identification, and we then communicated these recommendations to the original investigators. After implementing the recommended changes, each dataset is judged to be at very low re-identification risk.
Discussion: The particular nature of traumatic stress research, i.e. its content, data, and study designs, can influence the likelihood and potential impact of re-identification or disclosure. The two worked case examples in this paper demonstrate the utility of applying a systematic approach to assess and further mitigate re-identification risk in shared datasets. At each stage of the research data lifecycle, there are research practices and choices relevant to reducing re-identification risk. This paper presents practical tips for research teams to facilitate FAIR data practices while attending to data privacy.

Compartir datos de investigación sobre estrés traumático: evaluar y reducir el riesgo de reidentificación

Antecedentes: Las prácticas de datos FAIR sustentan el intercambio y la reutilización de datos, y son esenciales para el avance de la ciencia y la práctica en beneficio de las personas, familias y comunidades afectadas por el trauma. En la investigación sobre estrés traumático, al igual que en otras investigaciones de ciencias sociales y de la salud, los marcos éticos, legales y regulatorios exigen una cuidadosa atención a la privacidad de los datos. La mayoría de los investigadores en estrés traumático conocen los métodos básicos para desidentificar/anonimizar los conjuntos de datos que se compartirán. Sin embargo, nuestro campo generalmente no ha utilizado enfoques de análisis de datos sistemáticos para reducir el riesgo de reidentificación de los participantes del estudio o la divulgación de información personal o sensible.
Objetivo: Facilitar el intercambio seguro y ético de datos mediante una mejor preparación de los investigadores en estrés traumático para evaluar sistemáticamente y reducir el riesgo de reidentificación mediante métodos contemporáneos de análisis de datos.
Método: En dos estudios de caso que utilizaban conjuntos de datos de investigación provenientes de proyectos internacionales y multilingües sobre trauma, disponibles públicamente, aplicamos un enfoque sistemático guiado por la Lista de Verificación para la Reducción del Riesgo de Reidentificación en Datos de Investigación sobre Estrés Traumático.
Resultados: Para cada conjunto de datos de estudio de caso, identificamos acciones recomendadas específicas para reducir aún más el riesgo de reidentificación y las comunicamos a los investigadores originales. Tras implementar los cambios recomendados, cada conjunto de datos se considera con un riesgo de reidentificación muy bajo.

ARTICLE HISTORY
Received 13 January 2025
Revised 15 April 2025
Accepted 22 April 2025

KEYWORDS
FAIR data; data sharing; de-identification; anonymisation; re-identification risk

PALABRAS CLAVE
Datos FAIR; intercambio de datos; desidentificación; anonimización; riesgo de reidentificación

HIGHLIGHTS

- Data sharing and re-use help to advance science and practice, benefiting individuals, families, and communities impacted by trauma.
- Safe and ethical data sharing requires that datasets be effectively de-identified or anonymized to protect participant privacy.
- This paper presents and demonstrates a systematic approach and contemporary analytic methods to assess and reduce the risk of re-identification of participants when traumatic stress research data are shared.

CONTACT Nancy Kassam-Adams adamsn@chop.edu Center for Injury Research & Prevention, Children's Hospital of Philadelphia, 2716 South Street, Philadelphia, PA 19146, USA

Supplemental data for this article can be accessed online at https://doi.org/10.1080/20008066.2025.2499296

Discusión: La naturaleza particular de la investigación sobre estrés traumático, es decir, su contenido, datos y diseños de estudio, puede influenciar la probabilidad e impacto potencial de la reidentificación o divulgación. Los dos ejemplos de casos trabajados en este artículo demuestran la utilidad de aplicar un enfoque sistemático para evaluar y mitigar aún más el riesgo de reidentificación en conjuntos de datos compartidos. En cada etapa del ciclo de vida de los datos de investigación, existen prácticas y opciones de investigación relevantes para reducir el riesgo de reidentificación. Este artículo presenta consejos prácticos a los equipos de investigación para facilitar las prácticas de datos FAIR, respetando al mismo tiempo la privacidad de los datos.

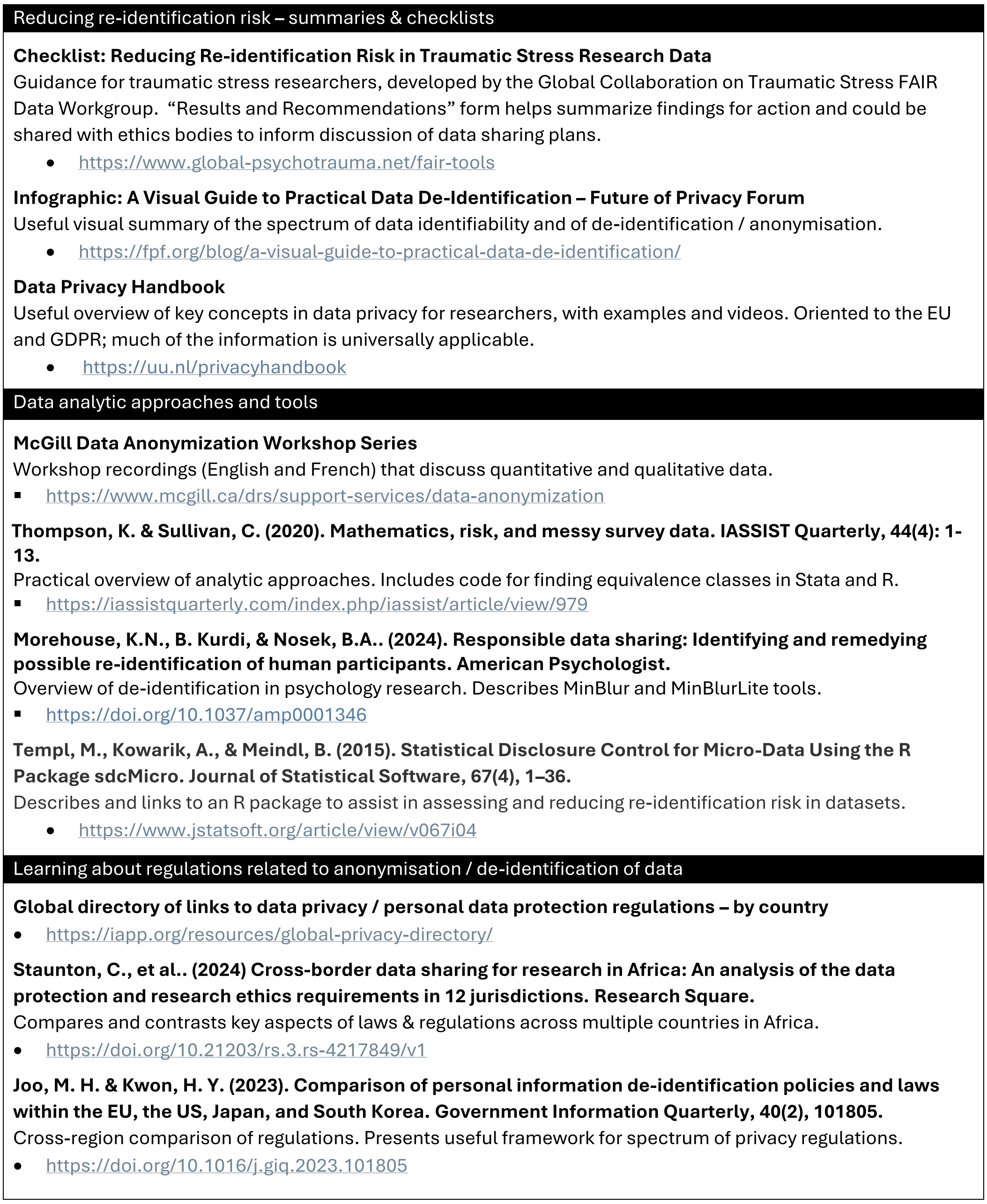
1. Introduction

Traumatic stress research serves important societal goals, addressing the impact of a wide range of types of trauma, from disasters to childhood maltreatment. Worldwide, studies have examined the mental health consequences of trauma, the effects of interventions to mitigate these consequences, as well as pathways to resilience and adaptation in trauma-exposed individuals and communities (Galatzer-Levy et al., 2018; Keyan et al., 2024). Data sharing and re-use are essential for advancing science and practice to benefit those affected by trauma, yet traumatic stress research data often remain unused by anyone other than the original research team (Kassam-Adams & Olff, 2020). Researchers are increasingly aware of the FAIR Data principles, i.e. that research data should be Findable, Accessible, Interoperable, and Re-usable (Sadeh et al., 2023; Wilkinson et al., 2016), of the broader movement towards more open, transparent science, and of growing mandates from research funders to share data. Both FAIR Data and open science principles recognize the need for additional care when sharing data collected from or about human participants, i.e. making these data 'as open as possible, as closed as necessary' (Hodson et al., 2018).

Nearly all mental health research involves human participants and promises confidentiality of the information collected. Sharing these data requires careful attention to data privacy and to the risk of re-identifying study participants or disclosing personal or sensitive information about individuals. Traumatic stress research has several features that provide additional impetus for lowering these risks: the content of information collected (e.g. nature of trauma exposure) may be subject to stigma or adverse social or legal consequences; and in some cases, exact dates or named events could reveal that participants are from a specific (identifiable) small group of people, even without other identifiers.

One challenge in achieving safe and ethical data sharing is that many legal and regulatory frameworks address data privacy but provide limited guidance on how to address privacy concerns while still sharing data. Most of these regulations incorporate the concepts that (a) datasets that include personal, identifiable information can be modified in some way, such that the data may then be considered anonymous or de-identified, and (b) in this new state these data are not subject to the same regulations that govern personal, identifiable data (Joo & Kwon, 2023; Mondschein & Monda, 2019). Towards that end, some regulations specify variables or information that cannot be included in a dataset that is de-identified / anonymous. One example is a United States (US) rule specifying 18 types of identifiers that must be removed from de-identified health information (US Department of Health and Human Services, 2012). Others, like the European Union (EU)'s General Data Protection Regulation (GDPR) state that the regulations apply only to 'personal data', thus when data are completely rendered anonymous, the GDPR does not apply and these data can generally be shared (European Parliament, 2016). Unfortunately, EU GDPR regulations provide only limited guidance for how to achieve 'anonymised' datasets and do not provide clear metrics for assessing anonymity of data. See Figure 1 for resources on regulations around the world.

The task of de-identifying or anonymising data can be complex, in that the 'anonymity or otherwise of data is a function of both the data and their context' (Elliot et al., 2020, p. 9). Several frameworks have been proposed to think holistically about data privacy and anonymity, including, for social science research, the 'Five Safes' framework (safe projects, people, data, settings, outputs) (Ritchie, 2017) and the UK Data Archive Anonymisation Decision Making Framework (Elliot et al., 2020). There are also systematic approaches, including contemporary data analytic approaches, to assessing and reducing re-identification risk in shared data (Morehouse et al., 2024; Sweeney, 2002; Thompson & Sullivan, 2020) However, traumatic stress researchers (like most health and social science researchers) are largely unfamiliar with, and have not received training in, these approaches. And, to our knowledge,[1] only one prior publication has described systematic approaches to anonymising and sharing trauma-related research data, in this case, qualitative data (Campbell et al., 2023).



Reducing re-identification risk – summaries & checklists

Checklist: Reducing Re-identification Risk in Traumatic Stress Research Data
Guidance for traumatic stress researchers, developed by the Global Collaboration on Traumatic Stress FAIR Data Workgroup. "Results and Recommendations" form helps summarize findings for action and could be shared with ethics bodies to inform discussion of data sharing plans.

- https://www.global-psychotrauma.net/fair-tools

Infographic: A Visual Guide to Practical Data De-Identification – Future of Privacy Forum
Useful visual summary of the spectrum of data identifiability and of de-identification / anonymisation.

- https://fpf.org/blog/a-visual-guide-to-practical-data-de-identification/

Data Privacy Handbook
Useful overview of key concepts in data privacy for researchers, with examples and videos. Oriented to the EU and GDPR; much of the information is universally applicable.

- https://uu.nl/privacyhandbook

Data analytic approaches and tools

McGill Data Anonymization Workshop Series
Workshop recordings (English and French) that discuss quantitative and qualitative data.

- https://www.mcgill.ca/drs/support-services/data-anonymization

Thompson, K. & Sullivan, C. (2020). Mathematics, risk, and messy survey data. IASSIST Quarterly, 44(4): 1-13.
Practical overview of analytic approaches. Includes code for finding equivalence classes in Stata and R.

- https://iassistquarterly.com/index.php/iassist/article/view/979

Morehouse, K.N., B. Kurdi, & Nosek, B.A.. (2024). Responsible data sharing: Identifying and remedying possible re-identification of human participants. American Psychologist.
Overview of de-identification in psychology research. Describes MinBlur and MinBlurLite tools.

- https://doi.org/10.1037/amp0001346

Templ, M., Kowarik, A., & Meindl, B. (2015). Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro. Journal of Statistical Software, 67(4), 1–36.
Describes and links to an R package to assist in assessing and reducing re-identification risk in datasets.

- https://www.jstatsoft.org/article/view/v067i04

Learning about regulations related to anonymisation / de-identification of data

Global directory of links to data privacy / personal data protection regulations – by country

- https://iapp.org/resources/global-privacy-directory/

Staunton, C., et al.. (2024) Cross-border data sharing for research in Africa: An analysis of the data protection and research ethics requirements in 12 jurisdictions. Research Square.
Compares and contrasts key aspects of laws & regulations across multiple countries in Africa.

- https://doi.org/10.21203/rs.3.rs-4217849/v1

Joo, M. H. & Kwon, H. Y. (2023). Comparison of personal information de-identification policies and laws within the EU, the US, Japan, and South Korea. Government Information Quarterly, 40(2), 101805.
Cross-region comparison of regulations. Presents useful framework for spectrum of privacy regulations.

- https://doi.org/10.1016/j.giq.2023.101805

Figure 1. Resources and tools.

The objective of this paper is to serve as a tutorial and resource – to facilitate safe and ethical data sharing by better preparing traumatic stress researchers to systematically assess and reduce re-identification risk. We first introduce key concepts related to re-identification risk. We describe data analytic risk assessment approaches, and how these can be applied to traumatic stress research data. We then demonstrate the application of these approaches to assess and mitigate re-identification risk via two case studies of publicly available datasets from projects of the Global Collaboration on Traumatic Stress. Finally, we provide practical recommendations for researchers – key steps across the research data lifecycle to address re-identification risk and allow appropriate data sharing and re-use.

In this paper we focus on quantitative data derived from surveys, interviews, and questionnaires or extracted from health or administrative records. We recognize the importance of addressing re-identification risk in other types of traumatic stress research data (e.g. genomic data [Bonomi et al., 2020]; qualitative or narrative data [Campbell et al., 2023]), however, these entail additional considerations beyond the scope of this introductory paper.

1.1. Key concepts and definitions related to re-identification risk

1.1.1. Anonymisation/de-identification

Anonymising or de-identifying a research dataset refers to the process of changing its elements in order to protect the identities of research participants and avoid linking participant identity to specific information in the dataset regarding individuals. (In this paper, we will use 'de-identify' and 'anonymise' interchangeably.) This process as undertaken by researchers is often guided by governmental or ethical rules or regulations. It may include removing a defined list of identifiers or protected data elements. In some definitions (e.g. in the EU GDPR), anonymisation is distinguished from 'pseudonymisation'. Pseudonymous data contains coded information allowing researchers to link data to specific individuals, e.g. a pseudonym known to the researchers, a participant ID linked to a master list of participant names. In this definition, anonymous data are data that have been irreversibly stripped of such information. In the EU, there is debate on whether, in order for data to be considered anonymous when shared, the data controller (original researcher) is allowed to retain a securely stored copy of a master list or whether the controller must also delete this list (International Association of Privacy Professionals (IAPP), 2023).

1.1.2. Re-identification

Re-identification refers to the potential for a dataset, even after anonymisation / de-identification, to be used to gain information about study participants. This could include discovering whether a specific individual is included in the dataset ('membership disclosure')/ discovering the identity of specific dataset case ('identity disclosure'), as well as discovering potentially sensitive information about specific individual participants ('attribute disclosure'; [Walsh et al., 2018]). In the trauma field, attribute disclosure could include the fact that a person has experienced a specific trauma or has a specific mental health diagnosis. In some cases, membership disclosure equals attribute disclosure, e.g. knowing that a person is included in a sample of survivors of childhood maltreatment.

The risk of re-identifying participants or disclosing sensitive data related to individuals is a function of many factors, including the content of the dataset but also how data are collected, stored, managed, shared or otherwise made accessible (Sensitive Data Expert Group, 2020). Re-identification may occur inadvertently, with no ill intentions, or via the efforts of intruders or 'bad actors' who purposefully set out to discover information that should not be available to them. The measures we undertake should aim to protect against both. Trauma researchers should bear in mind that we might not be the only source of information regarding a well-known traumatic event. Often the media and social media publish extensive information about experiences of survivors or bereaved loved ones; this information might be linked to research reports and increase the risk of re-identification.

1.1.3. Identifiers

Identifiers are pieces of information (variables) in the dataset that could be used to discover the identity of research participants. The most obvious are 'direct identifiers', variables that place participants at immediate risk of re-identification, e.g. full or partial names, physical or email addresses, phone numbers, and some exact dates. 'Indirect identifiers' or 'quasi-identifiers' are variables that, while not directly identifying individuals, might be used alone or in combination with other data to infer participant identity. For example, in trauma research datasets, the exact date of a traumatic event in combination with a participant's age, gender, profession, or city of residence, might uniquely identify an individual. Open-text responses, often collected in the context of studies that primarily record quantitative data, might unintentionally include direct or indirect identifiers.

1.1.4. Equivalence classes

Within a dataset, equivalence classes, or sets of 'data twins', are cases for which all of a specified set of quasi-identifiers are identical. An equivalence class may be of any size. An equivalence class of one, i.e. the only case in the sample with these exact identifiers, is called 'sample unique'. For example, the only 45-year-old male bus driver in a sample of disaster survivors would be a sample unique case based on age, sex, and profession.

1.1.5. Sample relative to population

A dataset that is a complete sample of a small, known population is very difficult to de-identify unless the number of demographic and attribute variables is trivial. It is important to define the size and identifiability of the population of interest, and how closely the study sample comes to including the complete population. Sampling a subset of the population creates uncertainty – we do not know that any given individual is in the dataset at all. Sampling can also protect against attribute disclosure, since individuals in the dataset are likely to have 'data twins' outside the dataset whose identities or information (e.g. potential responses to sensitive questions) are unknown. In other words, we cannot know if the reported finding refers to Steven (a 45-year-old bus driver in the sample) or to David (a 45-year-old bus driver who is a disaster survivor but not a study participant).

1.1.6. Sampling frame and re-identification frame

The theoretical target population for a study (e.g. adults seeking treatment for PTSD) can be distinguished from the sampling frame – a concrete / non-theoretical list of people who may be targeted for inclusion, e.g. the patient list at a clinic where participants were recruited. (In some cases, such as widely shared online surveys, there may be no sampling frame.) In considering re-identification risk, it is useful to think of the 're-identification frame' for a dataset – a list of individuals whose data may be included that could feasibly be constructed by an outside person (whether legitimate data user or malicious intruder).

1.1.7. Generalization and suppression

Generalization is when more granular values in a variable are combined to create broader categories that contain more cases, e.g. recoding age from years to decade categories, or 'top-coding' age to group all those over 80 in one category. Suppression occurs when selected cases with unusual combinations of quasi-identifiers are deleted or masked because they pose greater risk of re-identification. This may mean deleting cases or suppressing the values of one or more variables on these cases. Each of these techniques involves losing information and potential analytic value and should be employed balancing this loss with the potential to reduce the likelihood or impact of re-identification.

1.1.8. Penetration testing

In this context, penetration testing is the attempt to identify weaknesses in anonymisation by locating unusual cases or small groups of cases that could be identifiable to someone attempting to determine participant identities or attributes. It involves finding rare combinations of quasi-identifiers and examining them to determine if they are likely to pose risk of a person being distinguishable in the general population, e.g. an 85-year-old participant who lives in a small town and has an uncommon profession.

1.2. Data analytic methods to assess risk of re-identification

Although not familiar to most trauma researchers, the larger data privacy world has developed several analytical approaches that help assess the risk of re-identifying individual participants and guide efforts to ameliorate that risk (Morehouse et al., 2024; Research Data Management Support et al., 2024; Sweeney, 2002; Thompson & Sullivan, 2020) Here we briefly describe the most common of these ('k-anonymity') and suggest how this approach may be applied by traumatic stress researchers. Note that every assessment of re-identification risk is dependent on the specific set of variables and cases included in a dataset; whenever variables or cases are added, removed, or modified, the risk assessment can change.

The concept behind k-anonymity is relatively straightforward: to protect against re-identification risk, it should not be possible to distinguish individual cases (or very small numbers of cases) within a dataset that have specific combinations of quasi-identifiers (e.g. demographic factors) that could be matched to external data. Rare combinations in the data are more likely to correspond to rare and possibly recognizable individuals in the population. For example, if we can identify that just a few cases in the dataset match a certain individual's age, gender, and profession and we have reason to suspect that person is in the dataset, we might be able connect their identity to included sensitive data.

In these analyses, k is an integer set by the researcher. Demonstrating k-anonymity for a selected value of k and a specified set of quasi-identifiers means that each case in the dataset cannot be distinguished from $k-1$ other cases; thus, if $k = 5$, then each case has at least 4 'data twins' with the same values for these quasi-identifiers (Thompson & Sullivan, 2020). Using our earlier example of a sample of disaster survivors, if $k = 5$ for age, gender, and profession, then every case in the dataset has at least 4 'data twins' with the same combination of values for these factors. If there is a participant who is a 45-year-old male bus driver, then there are at least 4 others with matching identifiers. If one of these reported a history of sexual abuse, then even if their identities were known, it would not be possible to know which of the 5 or more 45-year-old male bus drivers was the individual with the history of abuse (attribute disclosure). Investigators consider what value of k is acceptable for each set of quasi-identifiers, taking into account their sample, population, and the potential impact of identity or attribute disclosure. A minimum of $k = 3$ is commonly suggested, i.e. the set of quasi-identifiers is shared by at least three cases / participants. A number of factors impact the ability to achieve a desired (minimum) value of k, including the absolute number of cases in the dataset, the number of quasi-identifiers for which k-anonymity is being evaluated, and the diversity of values for those identifiers within this dataset.

An additional wrinkle is that achieving an appropriate level of k-anonymity may not protect against attribute disclosure, if all of the cases in an equivalence class have the same sensitive data attributes, e.g. if there are a large number of 45-year-old male bus drivers in the sample but all provided the same response regarding a history of sexual abuse. In data privacy analytics, the concept of l-diversity helps to quantify this. A dataset is 'l-diverse' (for a value of l selected by the researcher) when the cases in each equivalence class have at least l different values for each sensitive variable of concern (Research Data Management Support et al., 2024; Thompson & Sullivan, 2020). However, when data come from only a small sample of a large re-identification frame population, then all cases in the dataset are likely to correspond to many non-study-participants (external to the dataset) with the same quasi-identifiers, whose attributes (e.g.

behaviour, trauma history) are unknown. This can provide adequate protection against attribute disclosure.

1.3. Applying these approaches to traumatic stress research data

Applying the k-anonymity approach to traumatic stress research has the potential advantage of offering a clear and unambiguous answer to the question of whether a dataset has been sufficiently anonymised – if an individual cannot be singled out within a dataset, it follows that they cannot be specifically identified within the population the dataset is drawn from. The reverse, however, is not true; a person who can be isolated in a dataset may have co-equivalents who are not in the dataset (Perry & Zenk-Möltgen, 2024). Thus, utilizing the k-anonymity rule on its own greatly overestimates re-identification risk in sample data. This is demonstrated by Thompson and Sullivan who found that in one survey, k-anonymity overestimated re-identification risk by a factor of 370 (Thompson & Sullivan, 2020). Another consideration is that many traumatic stress datasets include relatively small samples (in absolute numbers), limiting the ability to demonstrate that a dataset has met the 'k-rule'. Even in smaller samples, researchers may find it useful to assess k-anonymity for one or more sets of quasi-identifiers as part of the process of mitigating risk. Using this approach to find small equivalence classes can help identify cases that require further investigation. Cases with a k of 1 or 2 should be individually inspected for unusual combinations or extreme values likely to pose re-identification risk. For example, a person living in Mexico who responded to a survey in Finnish may be recognizable, where Mexican residents and Finnish speakers in general would not be at risk in an international survey even if they were unique in the data. Results of these analyses can point to areas for improvement, e.g. identifying variables where generalization or suppression would reduce risk while retaining the analytical value of the data.

Assessing the risk of identity and attribute disclosure requires that we consider the sample (cases in the dataset) relative to the re-identification frame population. This varies widely in traumatic stress research, depending on the type(s) of trauma exposure(s) of interest and the research questions addressed. For example, a trauma study initiated in the aftermath of a disaster or violent incident might attempt to recruit every individual directly impacted. If the impacted group is known or can be inferred, then this may come close to being a complete sample of a population. When a limited population is impacted, recruiting a nearly complete sample would be good for the validity of study findings, but might at the same time create challenges for reducing the risk of identity or attribute disclosure. In contrast, a study of the adult impact of child sexual abuse might invite any interested adult to respond to an online announcement. Since the broader population of sexual abuse survivors is not an identifiable group, the re-identification frame would be all adults who have internet access and speak the language(s) in which the survey is offered. Many trauma studies fall between these extremes. To estimate re-identification risk as it relates to the sample and re-identification frame, we can think (in orders of magnitude) about the ratio of the re-identification frame population to the cases in the dataset. We propose that 1000:1 or greater (dataset contains less than 0.1% of the re-identification frame population) might be considered as 'very low risk', that a ratio of 100:1 (dataset includes 1%) is 'low risk', that ratios between 100:1 and 10:1 be considered as 'medium risk', and that a ratio of 10:1 or less (dataset contains more than 10%) be considered as 'high risk'. Datasets that include a complete sample of the re-identification frame population are at 'very high risk'.

2. Method

To demonstrate the application of these methods, we undertook two case studies using publicly available trauma research datasets. Each dataset comes from an international, multi-language project of the Global Collaboration on Traumatic Stress (GCTS), and (somewhat unusually for trauma research datasets) each is shared and openly available. Case Study 1 uses a dataset from a GCTS international survey of 222 traumatic stress researchers examining their views and practices regarding data sharing and re-use (Prakash et al., 2023a; Prakash et al., 2023b). Case Study 2 uses a dataset from the GCTS Global Psychotrauma Screen (GPS) project in which over 10,000 adults worldwide reported their responses to stressful events (Haering et al., 2024; Hoeboer & Olff, 2024; Olff et al., 2021)

Our approach draws on the concepts and analytical approaches described above, utilizing the Checklist for Reducing Re-Identification Risk in Traumatic Stress Research Data developed by the GCTS FAIR Data Workgroup (available online and as a Supplement to this paper) (Global Collaboration on Traumatic Stress FAIR Data Workgroup, 2024). The Checklist helps researchers determine whether and how traumatic stress research data can be made accessible, and how to mitigate risks to make it safer to share the data appropriately. Targeting traumatic stress research data, it supplements existing broader frameworks in health and social sciences (Elliot et al., 2020; Ritchie, 2017). Guided by steps outlined in the Checklist, for each case study we: (a) listed quasi-identifiers, identifying those of greatest potential concern (alone or in combination) for re-identification risk; (b) listed any items of concern for harm / stigma and characterized

Table 1. Case Study 1: results and recommendations.

Theoretical study population	Researchers of traumatic stress
Potential harm	Participants were promised confidentiality. Researchers might be mildly discomfited to have their data-sharing views known. **Relative risk of potential harm IF re-identified = Low**
Potential to construct a re-identification frame	It would not be possible to construct a complete list of potential participants as the survey was shared on several mailing lists and on social media. Likely participants would include members of several scientific associations and their students, estimated at around 10,000 individuals worldwide. This number of possible participants will be used as our re-identification frame.
Relationship of survey sample to re-identification frame	222 | 10,000. Given each potential survey unique will have ~45 co-equivalents outside the dataset, attribute re-identification in the absence of identity disclosure is not a concern. **Relative risk based on sample proportion = Medium**
Direct identifiers	Not collected in original study
Free text with personally identifiable information	Removed in shared dataset
Select demographic quasi-identifiers of concern	Variables of potential concern: • Age group – 6 categories • Gender – 2 categories • World region – 7 categories • Survey language – 5 categories • Discipline – 7 categories plus other
Initial inspection for k-anonymity	Achieving k-anonymity is very difficult with a sample size this small (*N* = 222). The set of quasi-identifiers listed as initial targets of concern gives 7*5*6*2*7 = 2940 possible combinations of characteristics that the 222 participants could have. In practice 68 participants are sample unique and 44 are pairs (data twins). These data cannot be proven to meet k-anonymity.
Penetration testing – Univariate and bi-variate checks of select quasi-identifiers for outliers and small groups	Inspection of the sample uniques and pairs (cases with k = 0 or 1) identified several unusual combinations of language and region that could lead to inference about the participant's specific location. For example, French respondents from North America are likely from the province of Quebec in Canada. Some disciplines are unusual within the dataset. Psychology and Psychiatry are very common; each of the remaining 5 disciplines had less than 5 cases each. Combining an uncommon discipline with the less common other groupings could increase someone's confidence that they had correctly re-identified a participant. A check for outliers across remaining demographic variables revealed a few participants who reported they had been conducting research in the field for more than 50 years. These unusually high values may be associated with unusually high participant age. Based on the distribution of cases, numbers above 40 were considered outliers. **Relative risk: Some high-risk combinations found**
Recommendations	Delete 'Survey Language' as this is not likely of analytic value for re-use. Group Discipline into 'Psychiatry', 'Psychology' and 'Other'. Top-code 'years conducting research' at 40. Note: An added complication given the nature of this study is that participants may know each other and be likely to access these shared data for re-use, leading to the possibility of inadvertent recognition. Thus, we aim to be conservative regarding unusual combinations. **Assessment: Risk can be substantially reduced if recommendations implemented.**

level of harm if disclosed; (c) Estimated the nature and size of the re-identification frame and characterized relative risk based on relationship of sample to re-identification frame; (d) Conducted an initial inspection for k-anonymity (using Stata Statistical Software: Release 18) and conducted univariate and bivariate checks for outliers and small groups, and then made recommendations for modifications of the dataset to reduce risk. We summarized our findings for each case study using the Checklist's 'Results and Recommendations' form (Tables 1 and 2).

3. Results

3.1. Case study 1: international survey of trauma researchers regarding data sharing and re-use

In this online survey, quasi-identifiers in the shared dataset included: age group, gender, region, discipline, survey language, and years spent conducting research. Several quasi-identifiers collected in the study had been generalized before the data were shared, e.g. age (collected in years) was grouped into decade ranges. Some quasi-identifiers had been partially suppressed before sharing to mask less common responses (in the sample or population), e.g. those not identifying their gender as male or female.

See Table 1 for results and recommendations. Relevant findings include uncommon professional disciplines, cases with unusually high number of years in the field, and several unusual language-region combinations that could increase re-identification risk. We identified specific recommendations to modify these variables to reduce risk.

Outcome of recommendations: We informed the original research team of these recommendations; the changes were implemented in a new version of the shared dataset. The revised shared dataset no longer contains any combinations that seem unusual.

Table 2. Case Study 2: results and recommendations.

Theoretical study population	Adults who have experienced a stressful event
Potential harm	Participants were promised confidentiality. Data contains items about mental health and substance use that could be embarrassing or stigmatizing. **Relative risk of potential harm IF re-identified = high**
Potential to construct a re-identification frame	Nearly all adults have experienced a stressful event. This survey was circulated widely online in over 30 languages. The re-identification frame thus equals the adult population of the world speaking one of these languages and having Internet access.
Relationship of survey sample to re-identification frame	∼10,000 | billions. Given each potential survey unique will have millions of co-equivalents outside the dataset, attribute re-identification in the absence of identity disclosure is not a concern. **Relative risk based on sample proportion = Very, very low**
Direct identifiers	Not collected in original study
Free text with personally identifiable information	Not collected in original study
Select demographic quasi-identifiers of concern	Variables of potential concern: • Age (recorded in years) • Gender – 3 categories – male, female, other • Geography – Specific country of residence • Language of survey Demographic-adjacent variables asking about pandemic-related job activities were considered of lesser concern.
Initial inspection for k-anonymity	Despite the limited number of demographic variables and $N > 10,000$, k-anonymity was not met due to age being recorded in single years and the presence of countries with only a few respondents.
Penetration testing – Univariate and bi-variate checks of select quasi-identifiers for outliers and small groups	4 cases with age over 100. The presence of centenarians is particularly risky – there are web sites that list known centenarians by country. Some respondents from very small countries or regions. Some unusual country–language combinations found. **Relative risk: Some unusual cases and high-risk combinations found**
Recommendations	(1) Top-code age at 80 (general recommendation for datasets that include age). (2) Delete the 'country' variable, while retaining the region variable and the variable ('country30') that contains only countries with more than 30 respondents in order to reduce re-ID risk while retaining reanalysis value. **Assessment: Risk can be substantially reduced if recommendations implemented.**

This, along with the low potential for items within these data to cause harm, lead us to conclude that that there is very low risk in these data being publicly available.

3.2. Case study 2 – international survey of adults regarding impact of stressful events

In this online survey of adults (age 16+), quasi-identifiers in the shared dataset included age, gender, country, and pandemic-related job activities. Several quasi-identifier variables had versions in which information was generalized or suppressed. For example, in addition to specific countries, the shared dataset included variables for 'region' (grouping multiple countries) and 'country30' (including only countries with >30 participants).

See Table 2 for results and recommendations. Relevant findings include cases with unusually high age or from low-population countries or regions, and several unusual country-language combinations. We identified specific recommendations to modify these variables to reduce risk.

Outcome of recommendations: We informed the original research team of these recommendations; the changes were implemented in a new version of the shared dataset. The revised shared dataset no longer contains problematic outliers nor combinations of quasi-identifiers assessed to be unusual. Although the potential harm / stigma from some items within the dataset could be high, the lack of a usable re-identification frame and of geography more specific than the country level makes re-identification extremely unlikely, reducing risk to a minimal level.

4. Discussion

Applying a systematic approach to reducing re-identification risk can help traumatic stress researchers optimize safe and ethical sharing of the data they collect, advancing the ultimate goal of benefiting people impacted by trauma. Implementing these methods may help researchers clarify how to fulfil their (sometimes conflicting) obligations to meet funder and publisher mandates for data sharing while also meeting regulatory mandates to protect participant privacy. A systematic approach that includes quantifiable metrics can also help ethical review bodies make well-informed decisions about how and when traumatic stress research data can be shared.

The two worked case examples demonstrate the application of a systematic approach to assess and mitigate re-identification risk when data are shared. The case examples show the utility of a checklist created by the GCTS FAIR Data Workgroup, intended for use in the context of each researcher's local and national regulatory and ethical requirements.

Any application of these methods must balance multiple aims. In their Anonymisation Decision-Making Framework (ADF), Elliot et al. (2020) note the 'utility principle': 'Anonymisation is a process to produce safe data but it only makes sense if what you are producing is safe useful data.' (p. 15) Researchers and ethical review bodies are faced with balancing these two principles – 'safety' (reducing re-identification risk) and maintaining usefulness of shared data to enable analyses that advance science and practice. For example, grouping participant age into multi-year categories in a sample of trauma-exposed children might preclude important analyses of developmental factors in children's trauma responses or in treatment effectiveness; researchers might choose to group or suppress other variables while retaining age. Systematic assessment of re-identification and disclosure risks in context will help investigators make thoughtful, empirically-grounded decisions about what can be included in a shared dataset as well as how and with whom to share it.

4.1. Reducing risk of re-identification: considerations across the research data lifecycle

At each stage of the research data lifecycle, there are research practices and choices relevant to reducing re-identification risk. See Figure 2 for a summary with action steps. Considerations listed at later stages are relevant to keep in mind during earlier stages. From the beginning, it is particularly useful to consider how and where data might ultimately be shared, to help guide choices regarding data collection and management.

4.1.1. Stage I: Designing study and data structures

During study design, investigators can think ahead to facilitate the process of producing a shareable dataset in keeping with their institutional and national regulations. Some funders now require such planning within the funding application (OpenAIRE, 2022; US National Institutes of Health, 2023). Researchers can use concepts and tools described in this paper to anticipate re-identification risk levels and select data elements balancing future analytical value against the risk or impact of inadvertent disclosure. See Figure 1 resources, e.g. the Future of Privacy Infographic, the IAPP Global Directory of links to regulations.

4.1.2. Stage II: Data cleaning and processing – preparing data for sharing

After the basic steps such as removing direct identifiers, next steps involve potential quasi-identifiers, alone and in combination. To determine which quasi-identifiers could be problematic, researchers can use data analytic methods (described in this paper) and also assess which data points represent potential for harm if disclosure were to occur. Resources in Figure 1 provide additional guidance, e.g. Data Privacy Handbook and McGill Workshop videos, examples and discussion in the Thompson and Morehouse papers, and the sdcMicro R package resources from Templ et al..

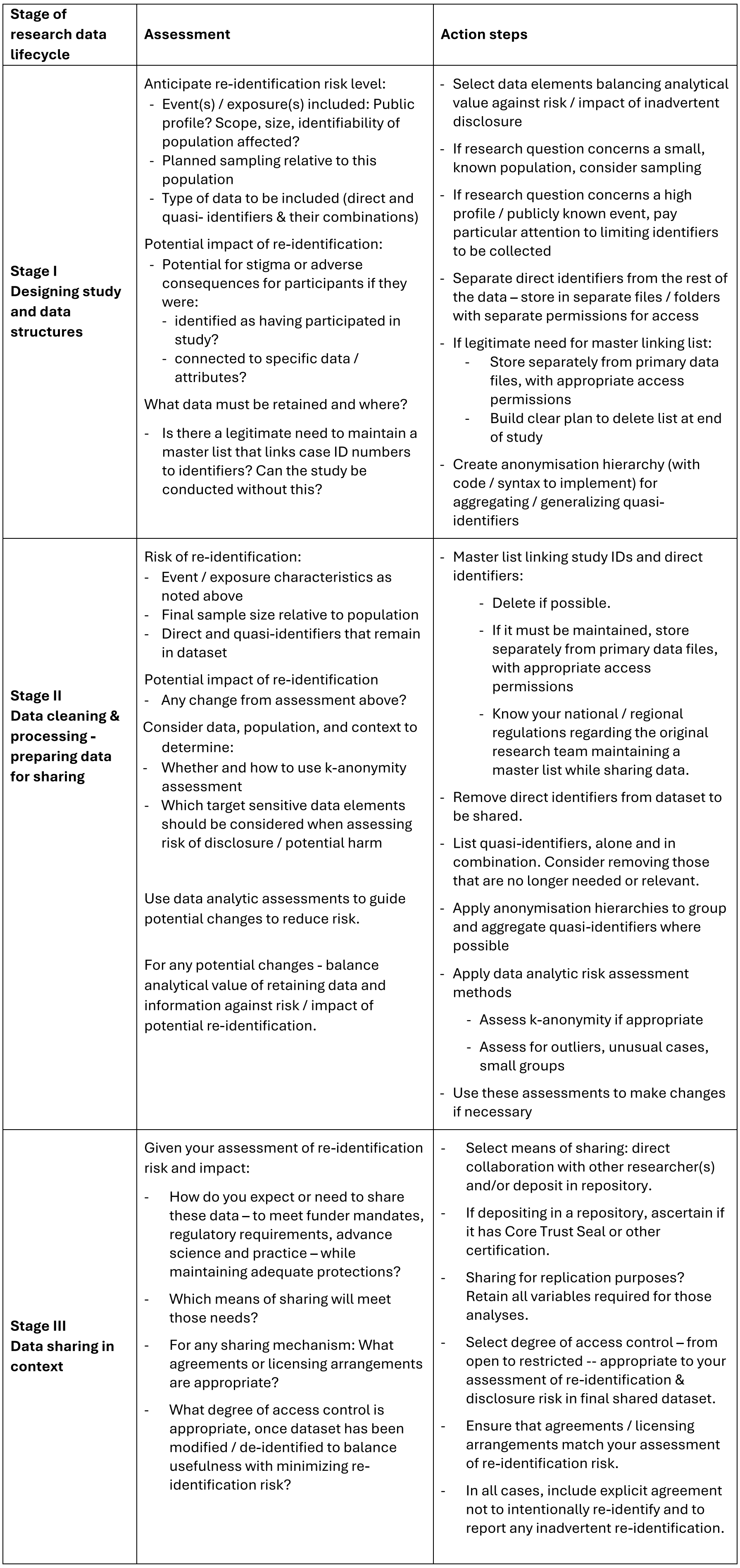
4.1.3. Stage III: Data sharing in context

While few trauma datasets can be made openly available, most can be made accessible (the 'A' in FAIR). Depositing data in an established repository affords long-term preservation and creates an ongoing resource. Examples in our field include the Child Trauma Data Collection (2024), and the Grief in Daily life Archive (Grief-ID Archive) (Pociūnaitė-Ott & Lenferink, 2024) How data are shared is relevant to mitigating identity or attribute disclosure risk. The Checklist (Global Collaboration on Traumatic Stress FAIR Data Workgroup, 2024) helps guide researchers in balancing assessed risk, analytical value of data, and regulatory requirements to select an appropriate level of open vs restricted access. Most established repositories (Core Trust Seal, 2020) offer varying levels of access controls – from fully public to very restricted (Marcotte et al., 2023). Each level provides opportunities to mitigate risks via technical, legal, and/or behavioural mechanisms. No matter how data are shared, researchers should expect data recipients to explicitly agree not to try to re-identify individuals and to report any inadvertent re-identification. For an example of such language for publicly available data, see Figure 3.

4.2. Special considerations for trauma research

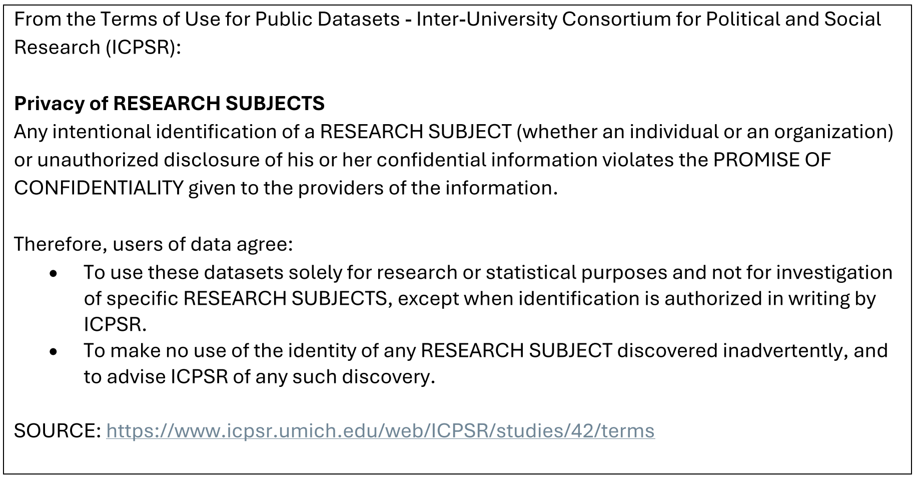
The particular nature of traumatic stress research, its content, data, study designs, etc., can influence the likelihood and potential impact of re-identification. The first potential harm from re-identification is the violation of trust if confidentiality is broken. The nature of trauma exposure(s) and public attention means that re-identification risk is influenced not only by population and sample size, but also by the degree of 'public profile' (including social media) versus more private awareness of specific events. The concept of the re-identification frame is particularly relevant to trauma datasets for this reason. The potential impact of identifying participants or disclosing sensitive attributes is influenced by the nature of social stigma for some trauma exposures and trauma-related mental health responses. And there are cultural variations in stigma and in privacy norms regarding mental health or trauma exposure.

Greater public profile, higher stigma, and smaller populations can each increase risk or impact of re-identification. With regard to a particular dataset, it may be useful to characterize trauma exposures and outcome measures along dimensions of public /



Stage of research data lifecycle	Assessment	Action steps
Stage I **Designing study and data structures**	Anticipate re-identification risk level: - Event(s) / exposure(s) included: Public profile? Scope, size, identifiability of population affected? - Planned sampling relative to this population - Type of data to be included (direct and quasi- identifiers & their combinations) Potential impact of re-identification: - Potential for stigma or adverse consequences for participants if they were: - identified as having participated in study? - connected to specific data / attributes? What data must be retained and where? - Is there a legitimate need to maintain a master list that links case ID numbers to identifiers? Can the study be conducted without this?	- Select data elements balancing analytical value against risk / impact of inadvertent disclosure - If research question concerns a small, known population, consider sampling - If research question concerns a high profile / publicly known event, pay particular attention to limiting identifiers to be collected - Separate direct identifiers from the rest of the data – store in separate files / folders with separate permissions for access - If legitimate need for master linking list: - Store separately from primary data files, with appropriate access permissions - Build clear plan to delete list at end of study - Create anonymisation hierarchy (with code / syntax to implement) for aggregating / generalizing quasi-identifiers
Stage II **Data cleaning & processing - preparing data for sharing**	Risk of re-identification: - Event / exposure characteristics as noted above - Final sample size relative to population - Direct and quasi-identifiers that remain in dataset Potential impact of re-identification - Any change from assessment above? Consider data, population, and context to determine: - Whether and how to use k-anonymity assessment - Which target sensitive data elements should be considered when assessing risk of disclosure / potential harm Use data analytic assessments to guide potential changes to reduce risk. For any potential changes - balance analytical value of retaining data and information against risk / impact of potential re-identification.	- Master list linking study IDs and direct identifiers: - Delete if possible. - If it must be maintained, store separately from primary data files, with appropriate access permissions - Know your national / regional regulations regarding the original research team maintaining a master list while sharing data. - Remove direct identifiers from dataset to be shared. - List quasi-identifiers, alone and in combination. Consider removing those that are no longer needed or relevant. - Apply anonymisation hierarchies to group and aggregate quasi-identifiers where possible - Apply data analytic risk assessment methods - Assess k-anonymity if appropriate - Assess for outliers, unusual cases, small groups - Use these assessments to make changes if necessary
Stage III **Data sharing in context**	Given your assessment of re-identification risk and impact: - How do you expect or need to share these data – to meet funder mandates, regulatory requirements, advance science and practice – while maintaining adequate protections? - Which means of sharing will meet those needs? - For any sharing mechanism: What agreements or licensing arrangements are appropriate? - What degree of access control is appropriate, once dataset has been modified / de-identified to balance usefulness with minimizing re-identification risk?	- Select means of sharing: direct collaboration with other researcher(s) and/or deposit in repository. - If depositing in a repository, ascertain if it has Core Trust Seal or other certification. - Sharing for replication purposes? Retain all variables required for those analyses. - Select degree of access control – from open to restricted -- appropriate to your assessment of re-identification & disclosure risk in final shared dataset. - Ensure that agreements / licensing arrangements match your assessment of re-identification risk. - In all cases, include explicit agreement not to intentionally re-identify and to report any inadvertent re-identification.

Figure 2. Reducing re-identification risk: Considerations at each stage of research data lifecycle.



From the Terms of Use for Public Datasets - Inter-University Consortium for Political and Social Research (ICPSR):

Privacy of RESEARCH SUBJECTS
Any intentional identification of a RESEARCH SUBJECT (whether an individual or an organization) or unauthorized disclosure of his or her confidential information violates the PROMISE OF CONFIDENTIALITY given to the providers of the information.

Therefore, users of data agree:

- To use these datasets solely for research or statistical purposes and not for investigation of specific RESEARCH SUBJECTS, except when identification is authorized in writing by ICPSR.
- To make no use of the identity of any RESEARCH SUBJECT discovered inadvertently, and to advise ICPSR of any such discovery.

SOURCE: https://www.icpsr.umich.edu/web/ICPSR/studies/42/terms

Figure 3. Exemplar language delineating data users' responsibilities with regard to re-identification.

private knowledge, degree of stigma, and population affected. When marginalized populations, such as refugees or sexual minorities, are the focus of trauma research, data may also be used by governments (or other actors) with intentions that conflict with the interests of those groups. All of these contextual factors should play a role in researchers' assessment of how to protect against potential risks as they prepare data to be shared.

4.3. Future directions and conclusion

There are several directions for future work. We should characterize current strengths and gaps in trauma researchers' practice of assessing and ameliorating re-identification risk, so that training can be optimized. The current case studies assessed two publicly available and relatively low-risk datasets. It would be useful to apply these methods to a range of traumatic stress research datasets, especially those that present more difficult de-identification challenges (smaller samples, publicly known events) and then report the results to inform best practices. We might develop tools enabling researchers to share their use of these methods with ethical review bodies to ensure that risk-benefit analyses for data sharing are accurately understood.

In conclusion, the trauma field (like most health / social science research) has not generally employed systematic, empirical approaches to assessing or mitigating re-identification risk. With contemporary tools, we have the opportunity to do better. We hope this introduction sparks greater interest in learning, implementing, and improving these methods for traumatic stress research.

Note

1. We conducted a basic search for publications in the traumatic stress field addressing systematic approaches to data de-identification, i.e. methods for assessing or ameliorating re-identification risk. In September 2024, we searched PTSDPubs (a comprehensive database of traumatic stress-related publications [US Department of Veterans Affairs National Center for PTSD, 2025]), as well as three leading journals focused on trauma and its impact: *European Journal of Psychotraumatology*; *Psychological Trauma: Theory, Research, Practice, and Policy*; *Journal of Traumatic Stress*. Using the search terms ('k-anonymity' OR 'k anonymity' OR k-anon* OR 'de-identification' OR 'deidentification' OR de-identif* OR reidentif* OR 're-identification' OR re-identif* OR reidentif*) we found just one relevant publication, which addressed anonymising and sharing qualitative traumatic stress data (Campbell et al., 2023).

Acknowledgements

The authors gratefully acknowledge the assistance of Chris Hoeboer and Miranda Olff, the investigators responsible for the Global Psychotrauma Screen (GPS) Global dataset.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was funded in part by the Trauma Data Institute.

Data availability statement

Data used in Case Studies were obtained from two publicly available datasets, as referenced in the paper.

ORCID

Nancy Kassam-Adams http://orcid.org/0000-0001-7412-1428
Kristi Thompson http://orcid.org/0000-0002-4152-0075
Marit Sijbrandij http://orcid.org/0000-0001-5430-9810
Grete Dyb http://orcid.org/0000-0002-7138-3665

References

Bonomi, L., Huang, Y., & Ohno-Machado, L. (2020). Privacy challenges and research opportunities for genomic data sharing. *Nature Genetics*, *52*(7), 646–654. https://doi.org/10.1038/s41588-020-0651-0

Campbell, R., Goodman-Williams, R., Engleton, J., Javorka, M., & Gregory, K. (2023). Open science and data sharing in trauma research: Developing a trauma-informed protocol for archiving sensitive qualitative data. *Psychological Trauma: Theory, Research, Practice, and Policy*, *15*(5), 819. DOI:/10.1037tra0001358

Child Trauma Data Collection. (2024). https://www.icpsr.umich.edu/web/DSDR/series/2420.

Core Trust Seal. (2020). Core trust seal: Why certification? Retrieved February 4, 2020, from https://www.coretrustseal.org/why-certification/.

Elliot, M., Mackey, E., & O'Hara, K. (2020). *The anonymisation decision-making framework 2nd edition: European practitioners' guide*. UKAN Publications. https://ukanon.net/wp-content/uploads/2024/01/adf-2nd-edition-european-practitioners-guide-final-version-cover-2024-version-2.pdf

European Parliament. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance) 2016. OJ L 119 04.05.2016, 1. http://data.europa.eu/eli/reg/2016/679/oj.

Galatzer-Levy, I. R., Huang, S. H., & Bonanno, G. A. (2018). Trajectories of resilience and dysfunction following potential trauma: A review and statistical evaluation. *Clinical Psychology Review*, *63*, 41–55. https://doi.org/10.1016/j.cpr.2018.05.008

Global Collaboration on Traumatic Stress FAIR Data Workgroup. (2024). Checklist for reducing re-identification risk in traumatic stress research data, v1.0. 2024. https://www.global-psychotrauma.net/fair-tools.

Haering, S., Kooistra, M. J., Bourey, C., Chimed-Ochir, U., Doubková, N., Hoeboer, C. M., Lathan, E. C., Christie, H., & de Haan, A. (2024). Exploring transdiagnostic stress and trauma-related symptoms across the world: A latent class analysis. *European Journal of Psychotraumatology*, *15*(1), 2318190. DOI:10.1080/20008066.2024.2318190

Hodson, S., Jones, S., Collins, S., Genova, F., Harrower, N., Laaksonen, L., Mietchen, D., Petrauskaité, R., & Wittenburg, P. (2018). Turning FAIR data into reality. Interim report of the European Commission Expert Group on FAIR data. https://doi.org/10.5281/zenodo.1285272.

Hoeboer, C., & Olff, M. (2024). Global Psychotrauma Screen (GPS) - Global dataset. https://osf.io/gvaut/.

International Association of Privacy Professionals (IAPP). (2023). A practical guide to anonymization standards across the EU and UK. https://iapp.org/news/a/a-practical-guide-to-anonymization-standards-across-the-eu-and-uk.

Joo, M.-H., & Kwon, H.-Y. (2023). Comparison of personal information de-identification policies and laws within the EU, the US, Japan, and South Korea. *Government Information Quarterly*, *40*(2), 101805. https://doi.org/10.1016/j.giq.2023.101805

Kassam-Adams, N., & Olff, M. (2020). Embracing data preservation, sharing, and re-use in traumatic stress research. *European Journal of Psychotraumatology*, *11*(1), 1739885. DOI:10.1080/20008198.2020.1739885

Keyan, D., Garland, N., Choi-Christou, J., Tran, J., O'Donnell, M., & Bryant, R.A. (2024). A systematic review and meta-analysis of predictors of response to trauma-focused psychotherapy for posttraumatic stress disorder. *Psychological Bulletin*, *150*(7), 767–797. https://doi.org/10.1037/bul0000438

Marcotte, J., Rush, S., & Ogden-Schuette, K. (2023). Tiered access to research data for secondary analysis. *Journal of Privacy and Confidentiality*, *13*(2), 1–12. https://doi.org/10.29012/jpc.825

Mondschein, C. F., & Monda, C. (2019). The EU's General Data Protection Regulation (GDPR) in a research context. In P. Kubben, M. Dumontier, & A. Dekker (Eds.), *Fundamentals of clinical data science* (pp. 55–71). Springer International Publishing. DOI:10.1007/978-3-319-99713-1_5

Morehouse, K. N., Kurdi, B., & Nosek, B. A. (2024). Responsible data sharing: Identifying and remedying possible re-identification of human participants. *The American Psychologist*. Advance online publication. https://doi.org/10.1037/amp0001346

Olff, M., Primasari, I., Qing, Y., Coimbra, B. M., Hovnanyan, A., Grace, E., Williamson, R. E., Hoeboer, C. M., & The GPS-CCC Consortium. (2021). Mental health responses to COVID-19 around the world. *European Journal of Psychotraumatology*, *12*(1), 1929754. DOI:10.1080/20008198.2021.1929754

OpenAIRE. (2022). How to comply with Horizon Europe mandate for Research Data Management. Retrieved January 1, 2025, from https://www.openaire.eu/how-to-comply-with-horizon-europe-mandate-for-rdm.

Perry, A., & Zenk-Möltgen, W. (2024). When to use the k-rule? - criteria for managing uniqueness and de-anonymization risk in social science survey data. *Transactions on Data Privacy*, *17*(3), 123–146.

Pociūnaitė-Ott, J., & Lenferink, L. (2024). FAIR intense longitudinal data archive on prolonged grief in daily life. DOI:10.13140/RG.2.2.12452.80004

Prakash, K., Kassam-Adams, N., Lenferink, L. I. M., & Greene, T. (2023a). Data sharing and re-use in the traumatic stress field: An international survey of trauma researchers. *European Journal of Psychotraumatology*, *14*(2), 2254118. DOI:10.1080/20008066.2023.2254118

Prakash, K., Kassam-Adams, N., & Greene, T. (2023b). Dataset: Data sharing and re-use: An international survey of traumatic stress researchers' opinions and experiences. https://doi.org/10.17605/OSF.IO/P2VY5.

Research Data Management Support, Huijser, D., Moopen, N., Flores, J., Beltrán, M., de Bruijn, K., de Bruin, J., Capel, D., Dijkstra, F., Folkers, J., Franzke, A., de Graaf, J., van den Hout, S., Huigen, F., Janssen, R. D. T., Jovic, K., Kessels, L., Kleerebezem, S., de Koning-van Nieuwamerongen, D., … & Weijdema, F. (2024). *Data privacy handbook*. https://uu.nl/privacyhandbook.

Ritchie, F. (2017). The 'Five Safes': A framework for planning, designing and evaluating data access solutions. Data for Policy. https://doi.org/10.5281/zenodo.897820

Sadeh, Y., Denejkina, A., Karyotaki, E., Lenferink, L. I. M., & Kassam-Adams, N. (2023). Opportunities for improving data sharing and FAIR data practices to advance global mental health. *Cambridge Prisms: Global Mental Health*, *10*, e14. https://doi.org/10.1017/gmh.2023.7

Sensitive Data Expert Group. (2020). Sensitive Data Toolkit for Researchers Part 2: Human Participant Research Data Risk Matrix. https://doi.org/10.5281/zenodo.4088954.

Sweeney, L. (2002). Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(05), 571–588. https://doi.org/10.1142/S021848850200165X

Thompson, K., & Sullivan, C. (2020). Mathematics, risk, and messy survey data. *IASSIST Quarterly*, *44*(4), 1–13. https://doi.org/10.29173/iq979

US Department of Health and Human Services. (2012). Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. https://www.hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html#standard.

US Department of Veterans Affairs National Center for PTSD. (2025). PTSDPubs Database. Jan 1 2025. https://www.proquest.com/ptsdpubs/index.

US National Institutes of Health. (2023). Data sharing approaches. January 1 2025; https://sharing.nih.gov/data-management-and-sharing-policy/sharing-scientific-data/data-sharing-approaches#after.

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. DOI:10.1177/2515245917749652

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J., da Silva Santos, L. B., & Bourne, P. E. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18