Introduction to anonymisation techniques for social sciences research data – Q&A November 2024

19 November 2024

Public

Version: 01.00

“There are some health data classified as "sensitive" (e.g. some NHS data). Would you be able to clarify the definition or how this is different from special category data or directly identifiable data?”

The term sensitive data is not explicitly defined in the UK GDPR but is commonly used to describe information that, due to its nature, requires more advanced security and safeguards. In the context of NHS data, sensitive data typically includes medical records, patient diagnoses, or test results. The sensitivity of such data stems from its potential impact on individuals if disclosed, such as harm to privacy, stigma, or discrimination. While not a legal term, the concept of sensitive data reflects the practical need for robust protection of information that could significantly affect an individual.

Under UK GDPR, special category data is a legally defined category of personal data that requires higher protection because of its nature. It includes data revealing health information, racial or ethnic origin, religious beliefs, sexual orientation, genetic and biometric data (when used for identification), and more. Health data held by the NHS, for example, falls within this category.

Directly identifiable data refers to information that can directly identify an individual without needing additional data. Examples include names, NHS numbers, passport numbers, or other unique identifiers. While directly identifiable data might not always be inherently sensitive (e.g., a name alone is not typically sensitive), it becomes highly significant when combined with other types of data, such as medical information.

In the English context, the distinctions between these categories are particularly relevant to health and social care organizations like the NHS. While sensitive data is an informal term that highlights the practical need for stronger safeguards, special category data is a formal legal concept that imposes stricter processing requirements. On the other hand, directly identifiable data focuses on an individual’s identification and may overlap with both sensitive and special category data when linked to personal details like medical records.

“Would you treat "my brother was eaten by a lion six months ago" in an interview answer as an indirect identifier, or a non-identifying variable? Do researchers need to make a judgment depending on context?”

This is such an interesting question and the answer is it depends. An indirect identifier is a piece of information that, while not directly identifying on its own could reasonably be combined with other available information to identify an individual. The specificity of the event

(a person being eaten by a lion) combined with the time frame (six months ago) could make it possible to identify the individual, especially if the incident was rare, publicly reported, or occurred in a specific geographic region. For example, in a region where such events are rare or widely covered in the media, this information could lead to re-identification when paired with other data.

Alternatively, if the event were more generic (e.g., in a region where interactions with lions are unfortunately more common and less publicly reported) or the research context did not allow access to additional identifiable details (like geography or family connections), this statement might be treated as a non-identifying variable. In this case, it would be an anecdote or piece of qualitative data that doesn't uniquely point to any one person.

"Is it correct that if we ask people to waive their anonymity we need to tell them what we would use the survey data for?"

That is correct, however you should always, whenever possible, inform participants what their data will be used for, even in the case of an anonymised survey. This aligns with ethical research practices and legal obligations under the UK GDPR, ensuring transparency and building trust with participants.

Before even discussing the possibility of waiving anonymity with participants, researchers must ask themselves key questions to justify this choice. This ensures that waiving anonymity serves a meaningful purpose and is not done unnecessarily. As researchers we should always ask ourselves does waiving anonymity provide insights or outcomes that could not be achieved through anonymised data? For example, are personal stories or direct quotes with attribution crucial for the research findings to carry weight, does personal data provide critical context that anonymized responses cannot etc..

The differentiation between collecting directly identifiable data e.g. names, for use within the research project and wanting to share this information at the end of the project for future reuse is also very important.

"Is it ok to ask respondents to waive their anonymity in every survey? (We don't know for sure we will use the data but it would be good to capture as we are wanting to increase the value of our data to be able to use it for specific pension schemes). Or should surveys always be anonymous?"

The above response touches on general waiving anonymity, but when considering collecting directly identifiable data to enable linkage, it is essential to explicitly discuss the purpose,

scope, and implications with participants. This can be addressed in the Participant Information Sheet and consent form. Participants should be fully informed about what data will be linked, who will perform the linkage, and how it will be used. You should always allow participants to opt in or out. However, it is worth acknowledging that there are situations where unconsented studies and linkages occur, such as when using data for purposes that meet legal exemptions or rely on a lawful basis such as public interest or legitimate interests under the UK GDPR. In these cases, researchers must demonstrate compliance with data protection regulations, ensure transparency through appropriate public notices, and implement robust safeguards to minimise risks to participants' privacy. These situations are exceptions rather than the norm and require careful ethical and legal consideration.

Final note is that collecting personal information "just in case" is not compliant with data minimisation principles under UK GDPR. By clearly distinguishing linkage from data sharing and addressing it transparently in the consent process, researchers can ensure ethical compliance and participant trust.

"It was mentioned that Task in Public Interest can be used not only for processing personal data, but also for sharing research data. This implies that researchers might be able to bypass obtaining informed consent. Is that right? I also don't think it is clear how researchers "use" a legal basis. How is the legal basis applied in practice?"

There are some nuances here to consider. Under UK GDPR there are six lawful bases for processing personal data, one being consent. Researchers in the UK based at universities will often use Public Task as the legal basis for any research conducted, however this doesn't negate ethical obligations, including transparency and minimising harm. Public task can also justify sharing data for broader scientific benefit, provided safeguards like anonymisation and access control are applied. However ethical research practices require consent, even if legally it's not mandatory. This respects participants' autonomy and ensures transparency. Additionally, participants must still be informed of the legal basis, even if explicit consent is not required. This is part of the GDPR's transparency requirements.

To also bear in mind here is for the processing of special categories of personal data an additional legal basis needs to be identified, and researchers must always document both the lawful basis for processing personal data and the additional condition for special category data. While explicit consent is one of the additional conditions for processing special category data UK researchers based at universities will often use the condition noting that processing is necessary for archiving, scientific research, historical research or statistical purposes.

What does this mean in practice? Always contact the DPO at your organisation to check what lawful bases are used by the organisation you are based at, for example authorities that are not public bodies might rely on legitimate interests. The lawful basis used also has a direct effect on individual rights and further fantastic and in-depth information is available from the ICO at https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/lawful-basis/a-guide-to-lawful-basis/.

"How can we work together with the research participants? In order to reduce their concerns about the use of their personal information in our research?"

Working collaboratively with research participants is essential for addressing any concerns they may have about the use of their personal information. While the Participant Information Sheet plays a critical role in this process, it's all about active engagement. The PIS provides participants with clear and transparent information about the purpose of the research, the types of personal data being collected, and how it will be used, stored, and protected. It also outlines participants' rights, such as their ability to withdraw from the study or request the deletion of their data, and it reassures them that all data will be handled in compliance with ethical and legal standards. However, building trust and reducing concerns requires going beyond formal documentation.

Active engagement with participants throughout the research process can make a significant difference in fostering trust and alleviating concerns. Open communication is key; we should always create opportunities for participants to ask questions and receive clear, tailored explanations about how their data will be used. Involving participants in decisions about data use, such as whether it should remain identifiable or be anonymised, can help them feel more in control and more comfortable about their involvement.

It is also important to keep participants updated throughout the study, providing progress updates and communicating any changes in how their data will be handled if that is to happen. This helps to maintain transparency and builds trust over time.

"If research is place-specific and in-depth/longitudinal, how can you ensure data submitted for secondary use is both confidential (for participants) and useable (for future research)? Thinking particularly about qualitative data here and how to find the balance."

Place-based research presents unique challenges because the location itself is key to the data's meaning and value. In such cases, removing or anonymising the place entirely will undermine the usability of the data for secondary research. To address this while maintaining participant confidentiality and minimising risk, the three-pronged strategy of consent, anonymisation, and access control is essential.

For consent clearly explain to participants that the location will remain identifiable because it is integral to the research.

The anonymisation should focus on protecting people, for example pseudonymise participants' names, roles, or relationships, generalising specific details e.g. provide month and year instead of day month and year, aggregate socio-demographics for example age instead of using their raw age use age bands.

As anonymisation is limited by the need to retain the place, access control must then be carefully considered. For example permission only access, where ethical approval from the institution the secondary research is based at is needed and approval from the original data creator must be in place before the repository safely transfers the data.

These strategies allow to maintain the integrity of place-based qualitative research while respecting ethical responsibilities to participants. By acknowledging the importance of place and building safeguards around it, usable data can be shared ethically and legally.

"When psedonomysing names, any suggestion about how to choose psedonoysed name? Name can suggest participant's background - is it recommend to use a similar name which may suggest similar background or a completely different name, or depending on the study?"

When assigning pseudonyms it is good practice to ask participants about their preferences whenever possible. Consulting participants allows them to have a say in how they are represented, which respects their autonomy and fosters trust in the research process. It also reduces the risk of misrepresentation or reinforcing stereotypes, as participants can choose names that they feel accurately reflect their identity or background.

However, in situations where consulting participants is not feasible always balance the need to protect confidentiality with maintaining cultural or demographic representation where relevant. Pseudonyms should always be appropriate, avoid stereotyping, and the process behind choosing the pseudonyms chosen should be documented.

Involving participants where possible enhances ethical practice, but when this is not an option, careful consideration and transparency remain essential.

> **"How to anonymise data when conducting arts based research methods? I am conducting a mixed methods project. Starting with quant data followed by qual data. I aim to use arts-based methods with possibly an exhibition/community engaged workshop/event."**

The consent protocols are ever so critical here. Participants must understand how their contributions will be used, whether they will be anonymised, pseudonymised, or attributed, and any risks of identification. Ideally they should be given the choice.

In terms of anonymisation, this does depend on what was discussed and agreed with participants, but strategies might include pseudonymising names, generalising contextual details and using composite representations for example by creating a collage to protect individual identities.

From a practical point of view, with arts-based researchers we have seen that participants are keen for their outputs to be shared for example drawings or pictures and they usually do not require attribution, they prefer to maintain their anonymity but happy for the outputs to be shared.

Regarding the public facing activities again the consent protocols are most important and the Participant Information Sheet should clearly inform them of all activities and all risks involved.

> **"This is probably a question for the consent training but is it recommended to obtain separate consents for data disclosure by different identifiability? (e.g. consent for sharing their data in anonymised data, consent for sharing their data in de-identified data)"**

Yes it is always advisable to use very granular consent forms and this is actually the only way to allow participants to make clear, informed decisions about how their data will be used and shared when we provide them with all the applicable options.

We are supported by the Universities of Essex, Manchester, Edinburgh, University College London and Jisc. We are funded by UKRI through the Economic and Social Research Council