Open Data

Sharing sensitive data

How you share your data depends on any applicable legal and ethical factors – this is especially true in fields where the data collected relates to human research participants, such as medicine or the social sciences. You may have questions around how to protect research participants while openly sharing your data, but datasets that contain personal data can often be shared by ensuring you have informed consent for data sharing, have applied appropriate anonymization techniques, and/or controlled access to the data.

Consent

Research data can be made available for future reuse by ensuring that consent is sought from participants. Participants should be informed how the research data will be stored, preserved , shared, and reused in the long-term, and how confidentiality will be maintained.

Consent procedures must be tailored for the specific research context, methods and sample, the nature of the data (personal, sensitive, level of detail), the format of the data (surveys, written, recordings) and the planned data uses and handling. This will influence the type of consent and consent process used. You can find detailed guidance in the sections on gaining written or oral consent, and consent forms of the UK Data Service **here**.

Please note

Researchers are obliged to seek consent, but participants are free to decide if and how you can share their data. Researchers should impartially advise participants about the risks and benefits of research participation and data sharing. Participants then decide what they will consent to, and researchers must honor the wishes of the participants. If the participants don't give their consent, you can't make your data available for reuse.

Anonymization

Anonymization alters direct identifiers (name, postal code, phone number, etc.) and indirect identifiers (occupation, gender, location etc) such that individuals cannot be identified in a dataset. Both quantitative and qualitative data can benefit from anonymization using the following techniques:

Remove

Remove direct identifiers from your dataset. Better still, plan ahead and avoid collecting identifiable data that is not needed. Where direct identifiers such as a personal name cannot be easily removed while maintaining usefulness of the dataset, use pseudonymization instead (for qualitative data).

Generalize

Where possible, replace disclosive information with more generalized information – whilst still maintaining meaning. For instance, reference to a Catholic Church could be generalized to a place of worship.

Aggregate and reduce

Variables such as age and location can be aggregated or reduced to decrease the precision of the variable. For instance, recording birth year rather than birthdate is a form of reduction.

Log

Keep a record of each step of anonymization and be sure to keep this file separate from the anonymized data file.

Reassess

Continually assess, as you conduct the anonymization process, whether the risk of disclosure has been removed.

Please note

- Anonymization may impact the usefulness of data. Be sure to apply an appropriate level of anonymization. If data is becoming completely unusable you might want to choose to use a controlled access repository, instead of continuing to remove things from the dataset.
- Consider any linked datasets that are (or will be) available. In some cases, anonymized datasets can be at risk of disclosure when combined with other open datasets. Should this be the case for your dataset, controlled access may be required.
- Search and replace is useful for anonymizing datasets but be cautious, search and replace will not apply to misspelled words and may replace words not intended.

Controlled access

There may be cases where data cannot be fully anonymized and openly sharing data is not feasible. It may still be possible to make your data accessible, in line with the FAIR Principles, to authenticated users via a controlled-access repository. This depends on what the ethical board approving your study said about data sharing, and the level of permission granted. Bear in mind that mixed levels of access control can be put into place; you can combine controlled access to potentially disclosive data while openly sharing non-disclosive data. Click here for a list of protected access repositories.

Can't share the data? Share the metadata

In cases where data cannot be shared for legal and/or ethical reasons, you can openly publish a description of your data (known as a 'metadata record'). This helps others to discover your data and provides essential information about how the data can be accessed and cited. For example, you could post a "data codebook" or "data dictionary" in a repository that describes the variables used in your dataset. In this document, you can cite the article in which it appears in order to connect the data descriptor to the paper.

Please note
The metadata record for sensitive data should never contain identifying information like names.

Third party data

In cases where data has been obtained from a third party, restrictions may apply to the availability of the data. You should write a Data Availability Statement describing how other researchers can access and reproduce your data.

Please note
Be aware of any access and reuse restrictions where your dataset contains data derived from a third party.

Toolbox

- UK Data Service
- The Anonymisation Decision-making Framework
- Amnesia
- QAmydata
- sdcMicro
- ANDS Publishing and sharing sensitive data
- EDCTP Knowledge Hub
- How to make a data dictionary