Respecting privacy in research data management

Carrying out ethical scientific research is essential to ensure the reliability, quality and transparency of research. Personal data protection is an important element of an ethical scientific research.

Many types of research projects involve people – from clinical trials on patients to demographic data collections, from anthropological to linguistic studies. Lawful personal data processing (see box) implies foreseeing and dealing with privacy issues in accordance with the **'privacy by design' principle**, which requires focusing on these matters right from the planning phase. Another key principle is that of **data minimisation**, pursuant to which only personal data that are strictly necessary to achieve a certain purpose (here: a scientific purpose) should be collected and processed.

'Processing' means any operation or set of operations which is performed on personal data or on sets of personal data. Examples include collection, use, organisation, storage and destruction of personal data.

Personal data

'Personal data' means any **information that identifies** or makes it possible to identify **a natural person**, either directly (e.g. name and surname, online identifier, or personal images) or indirectly (e.g. data relating to their habits, lifestyle, health, financial status, or a code assigned to them within a scientific research project).

'Special' categories of personal data include data revealing racial or ethnic origin, religious or philosophical beliefs, political opinions, trade union

Applicable regulations

"Regole deontologiche per trattamenti a fini statistici o di ricerca scientifica (Ethics standards for data processing for statistical or scientific research purposes)"; "Autorizzazione generale al trattamento dei dati genetici (Aut. Gen. 8/2016) (General Authorisation for the processing of genetic data (Gen. Aut. 8/2016))" and "Prescrizioni relative al trattamento dei dati personali effettuato per scopi di ricerca scientifica (Aut. Gen. 9/2016) (Prescriptions on the processing of personal data for scientific research purposes (Gen. Aut. 9/2016))" issued by the Italian Privacy Authority; Regulation (EU) 2016/679 "General Data Protection Regulation".

membership, genetic data, biometric data, data concerning health, sex life or sexual orientation, legal data and data relating to criminal convictions and offences. These types of data require an extra level of protection because they can lead to discriminations.

Planning privacy management

Start by asking yourself if you really need to collect and process personal data for your specific research activity and by deciding whether to collect new data or reuse existing data collected in the past (also by third parties).

All research participants must receive clear, transparent and appropriate **information on the processing of personal data (privacy notice)** including information about the research project, and the purposes and methods of personal data processing. This document must also indicate the legal basis for the processing (e.g. consent, which can be collected in a number of ways – paper, online, videorecording, etc.) and the rights of data subjects over their own personal data.

To draft this privacy notice, you must **define beforehand** for how long the personal data collected will be retained in an identifying (i.e. non-anonymous) form, with whom such data will be shared and for what purposes.

Remember that the Italian law mandates approval by a dedicated ethical committee for all clinical research involving the recruitment of patients. In the case of the University of Bologna, the committee responsible for approving clinical research projects is the independent ethical committee of Area Vasta Emilia Centro (CE-AVEC).

For non-clinical research projects that involve the collection of personal data, it is in some cases appropriate to ask for the opinion of the Bioethics Committee. It is not yet mandatory, unless expressly requested by, for example, the funding body or the publisher of a publication containing the data.

In the field!

I am a researcher in the humanities and social sciences, and I collect observational data and data from surveys.
How can I manage cross-cutting issues such as privacy?

If you collect personal data, prepare a privacy notice as early as the planning phase, and have it signed by the persons you will interview or from whom you will receive information. Follow current legislation, such as the GDPR, and ask the competent offices for help in drafting your informed consent form.

Personal data security measures

Assess the technical and organisational measures necessary to ensure personal data protection during the active phases of research (collection, analysis, storage).

The first aspect to consider is **choosing an appropriate data storage system.** In order to comply with the General Data Protection Regulation (GDPR), if you need to use a cloud-based solution to facilitate collaboration with third-party partners in personal data processing, this needs to have servers based in a country that ensures an appropriate level of personal data protection, as established by the adequacy decision of the European Commission. Again, with a view to striking a balance between collaboration and protection, you must clarify who needs to access identifying data to conduct the research and **manage access rights to the folders** in which personal data are stored accordingly. Depending on the sensitivity of the data you process, consider encrypting the folders by using dedicated tools.

Another aspect to consider during short-term storage and data analysis is the possibility of anonymising or pseudonymising data. Anonymisation and pseudonymisation modify personal data to make the data subject unidentifiable or less identifiable, respectively. Both techniques involve removing or modifying direct and indirect personal identifiers and may reduce the quality and usefulness of research data, as they imply a loss of information.

Anonymisation is the processing of data so that the data subjects can no longer be identified. Anonymous data are such for everyone, including the researchers that collected them in the first place. To anonymise data, you need to select all possible direct and indirect identifiers and modify them using the most appropriate strategies. Pay special attention to attribute combinations that can lead to the identification of certain individuals, and to small population samples. To avoid inference disclosure, you can use measures such as generalisation, aggregation, top and bottom coding to hide identifiable outliers, data perturbation, etc. Anonymised data are no longer regarded as personal data under the GDPR.

Pseudonymisation is the processing of identifying data so that they can no longer be attributed to a specific person in the absence of additional information, such as an encryption key. Pseudonymised data are still regarded as personal data under the GDPR. Data pseudonymisation involves removing or encrypting any directly identifiable piece of information. For encryption, you should use random codes for each person and store the encryption key, possibly encrypted, separately from the encrypted data file.

In the field!

I need to pseudonymise two datasets, a quantitative one and a qualitative one. *What can I do?*

To pseudonymise quantitative data:

- Remove or replace all information that enables direct identification (e.g. name, surname, address, telephone number, email address, IP address, etc.) using a random code for each person.
- Encrypt the key that allows to re-identify each record and store it separately from the dataset.
- Generalise or remove indirect identifiers (e.g. age, occupation, etc.) from the dataset.

To pseudonymise qualitative data:

- For text, e.g. interview transcripts, use pseudonyms and general descriptions and mark replacements with [square brackets]. Example: [Person 1] works for [a financial organisation] in Belgium.
- For audio and/or video, use dedicated tools to blur faces and modify voices.

I am a medical researcher, and I need to anonymise my quantitative data.
What strategies can I use and what results should I expect?

- Generalising or removing indirect identifiers reduces the level of detail in the data. E.g. change "Age 27" into "Age group 21-30"; change "Schizoid personality disorder" into "Mental and behavioural disorder".
- Top and bottom coding hides outliers in the data. E.g. "Age group above 70", "Salary below 1,658 euros/month", etc.
- Data perturbation modifies the value of numerical data by adding 'noise' and replacing real values with simulated or average values.

Long-term preservation of personal data

The retention period for identifiable personal data must be decided at the start and disclosed to research participants in the privacy notice. It is unlawful to retain data longer than it is necessary to achieve the purpose for which they were collected. If you need to deposit identifiable personal data in open access to ensure the transparency and reproducibility of research, be aware that this is only possible if there is an appropriate legal basis to do so (e.g. the consent of the data subject).

If this is not the case, you must consider depositing them in a secure repository that allows for restricted access to the data and requires authorisation. Sometimes there may even be a committee responsible for evaluating requests for data access and reuse.

As already mentioned, anonymised data are no longer regarded as personal data under the GDPR. For this reason, careful anonymisation is the best strategy to make data available in a repository at the end of research.

Useful links

Intranet page on Personal data processing for scientific research (in Italian)
https://intranet.unibo.it/Ateneo/Web1/Pagine/PrivacyRicerca.aspx

Laws and regulations:

- Regole deontologiche per trattamenti a fini statistici o di ricerca scientifica (Ethics standards for data processing for statistical or scientific research) (in Italian)
  https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9069637
- Autorizzazione generale al trattamento dei dati genetici (General Authorisation for the processing of genetic data) (in Italian)
  https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/5803688
- Prescrizioni relative al trattamento dei dati personali effettuato per scopi di ricerca scientifica (Aut. Gen. 9/2016) (Prescriptions on the processing of personal data for scientific research purposes (Gen. Aut. 9/2016)) (in Italian)
  https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9124510#5
- Regulation (EU) 2016/679 (General Data Protection Regulation, GDPR) https://eur-lex.europa.eu/eli/reg/2016/679/oj
- Adequacy decision https://ec.europa.eu/newsroom/article29/items/614108 |
  https://www.garanteprivacy.it/temi/trasferimento-di-dati-all-estero

Useful tools:

- Encrypting data https://www.veracrypt.fr/en/Home.html |
  https://docs.microsoft.com/it-it/windows/security/information-protection/bitlocker/bitlocker-overview
- Blurring faces in a video https://coehelp.uoregon.edu/using-openshot-to-blur-a-face-in-a-video/
- Modifying recorded voices https://www.qualitative-research.net/index.php/fqs/article/view/512/1106
- Anonymising data https://amnesia.openaire.eu/ | https://arx.deidentifier.org/ | https://github.com/sdcTools/sdcMicro