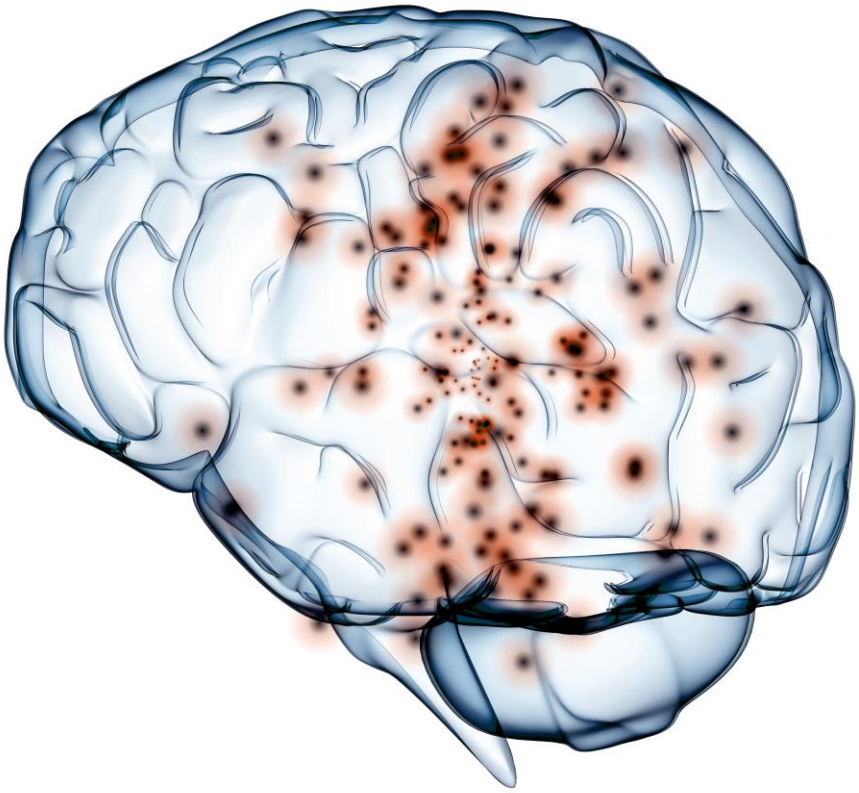
Sharing research data

Concerning human participants

Version 3.0 March 2023



University of Bristol

Research Data Service

Image: Neuronal activity DARPA, Public Domain

University of BRISTOL

CC BY

SCOPE

This guidance is intended to inform researchers about the process of sharing research data which concerns human participants, for use by 3rd parties. For example, to provide evidence of claims made in a journal article. The second part of the guide looks specifically at sharing data from clinical trials. An appendix provides suggestions for wording consent forms in a way which permits data sharing.

Introduction

Research data collected from human participants has often been excluded from data publication. Reasons included the belief that publishing such data was not ethical, it was prohibited by data protection legislation or was too difficult to do safely. This has meant that large amounts of potentially valuable data have been unavailable for further analysis.

This situation has changed thanks to a greater understanding of methods to 'de-sensitise' and to control access to this type of data. Publishers and research funders increasingly require researchers to find a way to provide access to their research data, even if that data initially includes personal information. In order to safely share data, early planning is required. Plans to modify personal data into a form suitable for publication should form part of the wider ethical planning process. Data publication plans should certainly be in place before any new data is collected.

This guide does not cover the collection of personal data, but rather methods to enable safe data sharing and publication. For further information on the responsibilities of the researcher working with personal data see the University Secretary's Office guidance on GDPR and research data.[1]

What is personal data?

Personal data is information that can be used to identify a study participant or subject. Implicit is a risk of discrimination, harm, or otherwise unwanted attention. As a general rule, personal data cannot be shared in its original form.

Under the General Data Protection Regulation (GDPR) and the UK Data Protection Act 2018, personal data is any information relating to an identified or identifiable living individual, which means:

> "a living individual who can be identified, directly or indirectly, in particular reference to a) an identifier such as a name, an identification number, location data or an online identifier, or b) one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of the individual."

Other types of information relating to research participants will also be unsuitable for publication. It is up to the owner of the data to understand and comply with the law and to exercise good judgment when deciding the potential risk of publishing a research dataset.

What do we mean by data 'publication'?

After data has been modified to minimise the potential to identify participants (see 'Removing information vs. modifying information' below), two options exist, depending on the level of risk remaining:

[1] https://www.bristol.ac.uk/secretary/data-protection/gdpr/gdpr-and-research/

Where there is negligible risk of identification and appropriate participant consent is in place, data may be made openly available either in whole or in part online, along with a metadata record, allowing others to discover, use and then cite the data.

Where some risk remains, or explicit consent is not in place, data may be made available via controlled access. A publicly available means of alerting potential re-users to the existence of the data is still required; this is often done via the open publication of a metadata record providing potential re-users with information about the nature of the dataset and how to apply for access.

Controlled access to restricted data

Ideally, research data would be modified into a form which could be made freely available and disseminated as widely as possible. However, this is inappropriate wherever there remains a risk of re-identification. If you wish to ensure data re-users are genuine researchers and that they agree to certain conditions of access, you should deposit your data into a repository which controls access. Access to data is often controlled by requiring potential data re-users to:

- Register and provide contact details
- Provide information about how they will use, store and manage the data if granted access
- Provide information about how they will destroy data or return data to a repository after a specified period
- Meet other conditions, including any specified in the consent forms and information sheets agreed to by research participants

When depositing data in any data repository, it is important to establish whether the process of granting access is managed by you, as the data creator, or by the repository. Data repositories usually have an online portal where users can search for and discover data, though not necessarily gain direct access. There are institutional-specific, discipline-specific and general data repositories. A very commonly used repository service for the social sciences is the UK Data Archive[2].

University of Bristol research data repository

The University has its own research data repository, data.bris[3], which allows researchers to publish datasets under several different access arrangements:

- *Open* – where consent has been given by participants to make anonymised data openly available and risk of re-identification is considered extremely low
- *Restricted* – explicit consent is not in place to openly share data but risk of re-identification is still considered to be low. Data is made available to bona fide researchers only, after they have signed a data sharing agreement
- *Controlled* – risk assessment of re-identification is medium to high or (for historical studies) consent for sharing was not sought from participants. Requests are referred to an appropriate data access committee before data can be shared under a data sharing agreement

[2] http://www.data-archive.ac.uk/

[3] https://data.bris.ac.uk/data/

- *Closed* - data is not available for sharing (except to regulators) because of ethical, IPR or other constraints. Requests for access to 'Closed' data are handled by the Information Rights Officer as FOI requests.

For more detail about how to publish using the Bristol University Repository please contact the Research Data Service[4].

Seeking participant consent to publish data

Whether planning to share data via Bristol's own data repository, the UK Data Archive or another repository, obtaining permission to do so from research participants is essential *even if data is to be anonymised before publication*. This is because some risk of re-identification may remain, even after anonymisation.

Gaining informed consent for data publication from research participants before data is collected is not only best practice, but avoids the expense, delay and loss of use of data involved in attempting to obtain consent later in the research process. Concerns that participants will refuse to participate in research if consent for data sharing is requested are likely to be unfounded as long as the participants are fully informed and have confidence that no identifying data will be shared.

Statements about data publication and sharing in participant information sheets and consent forms should:

1. Avoid precluding data anonymisation and subsequent sharing
2. Be clear on the difference between **administrative data** relating to participants (for example names and contact information), which generally should not be shared, and **research results**, which may be shared
3. State the possibility of future data publication (including storage in a repository) and sharing of anonymised data
4. State any conditions under which access to the data may be granted to others
5. Once a study is complete, a copy of the ethically approved participant information sheet and a blank copy of the consent form should always accompany the data. This is so repository managers and any subsequent users are aware of the conditions agreed to by research participants

The Research Data Service and Research Governance Team have collaborated to produce some template wording suitable for inclusion in consent forms and information sheets, which requests permission to publish anonymised data gathered from participants (see Appendix 1). Once drafted, the consent form and accompanying information sheet must form part of the study documentation to be approved by an appropriate Research Ethics Committee. Guidance on how and where to get ethical approval is available from Research Governance[5].

Where the seeking of consent is not possible

For existing data, when re-contact with research participants is no longer possible or practical, the possibility of sharing data may still exist.

Firstly, you should check to see if the ethically approved information given to the research participants prior to

[4] data-bris@bristol.ac.uk

[5] http://www.bristol.ac.uk/red/research-governance/ethics/

their consent indicated the possibility of future use or sharing of the data. If the information sheets or consent forms used specifically *precluded* data sharing, data publication will not be possible. If the information sheets or consent forms did not mention data re-use/sharing (after anonymisation) then after data have been properly anonymised, it may be possible to share data with selected researchers, under strict access controls. If your research data falls into this category, please contact the Research Data Service[6].

Sharing data that was not collected by you

In most instances, you cannot publish data that you did not collect because you do not own the copyright. As a data re-user, you can evidence your published research claims by citing the original source of the data in all the articles, presentations and any grant applications which are based upon the data.

Exceptions to this are data that have been licensed by the data owner to allow redistribution (e.g. under a Creative Commons or other open license) or in situations where you have paid the copyright holder for permission to redistribute parts of a dataset (i.e. this is permitted under the terms of a paid-for licence).

If you are in either of these situations, remember that, like original researchers, secondary data users have an obligation to ensure that the data are published responsibly and respectfully, and that the legal requirements of data protection legislation are met and the privacy of participants is safeguarded.

Anonymisation

The exact information that needs to be removed or modified during the process of anonymisation will vary depending on the contents of the dataset and the reason that the unmodified data has been deemed to be unsuitable for publication. Firstly, all 'direct identifiers' must be removed from datasets intended for open publication. These include:

- Name
- Initials
- Address, including full or partial postal code
- Spatial location (e.g. latitude and longitude units with enough precision to potentially locate the subject)
- Telephone or fax numbers or contact information
- Email addresses
- Vehicle identifiers
- Medical device identifiers
- Web or internet protocol addresses
- Biometric data
- Facial photograph or comparable image
- Un-anonymised audio or video recordings
- Names of relatives
- Dates relating to an individual (e.g. date-of-birth)

Removal of all direct identifiers may not be an adequate measure to prevent the identification of individuals. Datasets that contain two or more 'indirect identifiers' (listed below) may identify participants when these identifiers are considered together. This is called 'triangulation'. When two or more indirect identifiers are present, we recommend removing or modifying one or more of the indirect identifiers until the risk of identification is negligible. If you are still unsure if the risk of identification is negligible or not, please seek advice

[6] data-bris@bristol.ac.uk

from the University Rights Information Officer[7]. Indirect identifiers include:

- Place/location of treatment, education, service use
- Name of professional or business/service responsible for healthcare, education, service
- Gender
- Rare disease, condition, experience, treatment, or other characteristic
- Risky behaviours (e.g. Illicit drug use)
- Place of birth
- Socioeconomic data, such as occupation or place of work, income, or education level
- Household and family composition
- Body measures (e.g. height, weight)
- Multiple pregnancies
- Ethnicity
- Year of birth or age
- Verbatim responses or transcripts
- Dates of sensitive events
- Small sample sizes i.e. when the number of subjects with a certain characteristic is small

For further guidance on anonymisation see Information Commission, 2012, *Anonymisation: managing data protection risk code of practice*[8].

The risk of data-linking

Data-linking is the merging of two or more separate datasets that contain data about the same people or subjects. Data-linking has many positive applications, helping us to derive greater value from existing datasets.

However, data-linking also poses a risk; alone, a dataset may not contain enough information to identify individuals or place subjects at risk of identification, but when two or more datasets are combined, this may be achievable. This possibility must be considered by researchers planning to publish datasets.

Removing information vs. modifying information

Completely removing information from a dataset ensures the information cannot be used to identify participants or subjects. Sometimes, however, information can be modified enough that it no longer poses risk of identification and can thus remain in the dataset. This involves more effort but is a good option if complete removal of the information significantly de-values the dataset.

Methods of modifying data to limit identification include:

- Combining responses into categories, or a fewer number of categories than in the original dataset. This is a good option if only a small number of people or subjects possess the characteristic. For example, year of birth can be collapsed into 5 or 10-year age bands if only a few people in a dataset share a specific birth year
- Top and bottom coding: collapsing categories into upper and/or lower thresholds. This is a good option if only a small number of people have high or low measurements on a characteristic. For example, if few people report that they have more than five children, these participants can be combined with those who report five children and recoded as '5+ children'

[7] freedom-information@bris.ac.uk
[8] https://ico.org.uk/media/for-organisations/documents/1061/anonymisation-code.pdf

- Rounding dates, times, or measurements reduces the risk of identification when only a small number of people or subjects have a specific value
- Data suppression: involves creating ‘missing data’ if the inclusion of this data poses a risk to identification. Single values may be deleted, or all data for an at-risk research participant or subject

Methods used to modify a dataset must be documented and documentation must be published alongside the modified dataset if it is to retain its value to other researchers.

SHARING DATA FROM CLINICAL TRIALS

Many academic publishers, professional organisations and research funders, both public and commercial, now actively support the sharing of data from clinical trials. Sharing information about new health interventions ensures the results of trials are maximised, negative impacts do not needlessly recur and that valuable resources are not wasted by inadvertently repeating trials. The Declaration of Helsinki[9] is an international agreement stating that trial information should be publicly accessible.

However, researchers also have a duty to protect the interests of trial participants and sharing trial data can seem to conflict with this.

There are ways to address this challenge and successfully balance both disclosure and protection of participant information; the growing consensus is that anonymised trial data should be shared openly while potentially identifiable information is made available via controlled access.

A number of different levels of data disclosure are examined below. A study might decide to share:

1. Knowledge that the trial has been conducted (registration)
2. A brief summary of the trial's results
3. Full details about the trial's results and methods
4. Full results and methods along with individual patient-level data

Trial registration

As a minimum level of disclosure, it is a requirement in many countries that clinical trials be publicly registered.

In the UK the National Health Service Health Research Authority (HRA) require clinical trials which investigate a medicinal product, a medical device or other type of novel intervention, or randomised clinical trial to compare interventions in clinical practice, to be publically registered as a condition of favourable ethical approval[10]. Be Part of Research[11] (formerly the UK's Clinical Trials Gateway) then publishes information about registered trials.

Funded clinical trials at the University of Bristol are also registered through the Research Registration Checklist[12]. Please note however that this checklist is an internal university risk assessment and does not have any connection to the public registration of clinical trials.

Summary of trial results

Far fewer studies, perhaps as little as 20%[13] share the results of their research. And it is far less likely that an unsuccessful trial will share its results as compared with a successful trial.

The UK was originally a part of the EU Clinical Trials Regulation, establishing an EU-wide database of trial registrations and results together, which came into force in 2020. However, with the passage of Brexit and

[9] https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/
[10] https://www.hra.nhs.uk/documents/1018/research-transparency-sept-2015.pdf
[11] https://bepartofresearch.nihr.ac.uk/
[12] http://www.bristol.ac.uk/red/research-governance/ethics/sponsorship/
[13] Prayle AP (2012). Compliance with mandatory reporting of clinical trial results on ClinicalTrials.gov: cross sectional study, https://doi.org/10.1136/bmj.d7373

withdrawal of UK from the European Union, the appropriate authority for UK-based clinical trials is the Health Research Authority (HRA). From 1st January 2022, the HRA automatically registers clinical trials with ISRCTN[14] to ensure research transparency[15]:

This will start with clinical trials of investigational medicinal products (CTIMPs) that are submitted through combined review in the new part of IRAS. It is still a standard condition of a Research Ethics Committee (REC) favourable opinion for clinical trials to be registered on a publicly accessible database, this requirement has not changed.

HRA also indicates that any clinical trials submitted via IRAS before 31st December 2021 should register themselves with an established international register such as ISRCTN or ClinicalTrials.gov.[16]

AllTrials[17], a campaign group supported by organisations such as the NHS, the Wellcome Trust and the Medical Research Council is in favour of greater trial disclosure. AllTrials recommends that peer-reviewed journal articles, clinical study reports and similar reports are uploaded to public trial registries.

Upload of data to current registers may not yet be technically possible. However, where a document is available elsewhere online, a link or reference (for example the DOI of an Open Access journal article or a published dataset) might instead be provided as text within a trial's registration record.

Full trial details

Where full reports exist, such as Clinical Study Reports or their equivalent, both the EU Regulation (Section 39) and research transparency campaign groups suggest that they too be shared. Such detailed reports contain a wealth of information which may be of use to researchers.

Data should be anonymised before publication, with any information which might identify a participant removed but retained (see Patient level data, below). Examples are personal narratives or descriptions of individual adverse events. Consent given by participants must expressly state that data, once anonymised, can be shared beyond the original trial.

As many trial registries do not yet have the facility to attach documents to trial records, researchers may wish to use the University of Bristol Research Data Repository for this purpose. Low-risk, anonymised trial documents can be deposited into the repository and made available as 'Open' data without the need to restrict access. If you'd like to use Bristol's Data Repository please contact staff for more details[18].

Patient level data

In line with data protection legislation, data which contains personal information cannot be shared in its original form. Data which might potentially be used to identify a patient or participant should only be shared via controlled access. Some groups feel that this is well

[14] https://www.isrctn.com
[15] https://www.gov.uk/guidance/clinical-trials-for-medicines-apply-for-authorisation-in-the-uk#registration-of-your-clinical-trial
[16] https://www.hra.nhs.uk/planning-and-improving-research/research-planning/research-registration-research-project-identifiers/
[17] http://www.alltrials.net
[18] http://www.bristol.ac.uk/staff/researchers/data/contacts/

worthwhile, given the potential value of the information. Trial registries do not yet have the required mechanisms to do this and so an alternative route must be found. Historically the research group responsible for conducting a trial may have elected to make patient level data available 'on request' after checking the intentions of potential re-users.

However, this isn't ideal as the burden and costs of controlling access, perhaps for an indefinite period, then fall to clinical or research staff. But there is a more significant problem associated with the informal sharing of trial data. The Wellcome Trust suggest that data from individual trials should ideally be 'linked' and centrally searchable if its true value is to be realised. For this reason, it is recommended that data be deposited into a data repository.

The University of Bristol Research Data Repository can be used to provide controlled access to clinical data. Datasets held by the repository are linked, via the international DataCite[19] scheme, to similar datasets held all over the world and DataCite provide a central search facility[20].

There are also regulations in place for the retention of clinical trial *records*. Please see the University's Records Retention Schedule (IGP-04)[21] and 'Guidance on the retention of research records and data' for more information.[22]

Other barriers to sharing trial data

There are unlikely to be commercial barriers to sharing trial data. Indeed, many commercial research funders, such as GlaxoSmithKline and Roche, provide anonymised clinical trial data to researchers (approved by an independent review panel) through the Clinical Study Data Request website[23]. If your trial is commercially funded the terms of data availability should be stated in your commercial research contract, so this should always be checked first to ensure sharing of any data does not breach commercial restrictions. If data availability terms are not specified or you're not sure whether the responsibility to share trial data lies with you or your funder please contact the DREI Contract team for guidance[24].

[19] https://www.datacite.org/
[20] http://search.datacite.org/ui
[21] http://www.bristol.ac.uk/media-library/sites/secretary/documents/information-governance/records-retention-schedule.pdf
[22] Guidance on the retention of research records and data, PDF
[23] https://clinicalstudydatarequest.com/
[24] http://www.bristol.ac.uk/red/contracts/

APPENDIX 1: CONSENT FORM WORDING TO FACILITATE DATA SHARING

1: Future data usage

Where open access to anonymised data is planned:
"I understand that after the study, anonymised data may be made available as "open data". This means the data will be publicly available and may be used for purposes not related to this study. However, it will not be possible to identify me from these data."

Or, from the Health Research Authority:
"I understand that the information collected about me will be used to support other research in the future, and may be shared anonymously with other researchers."

For data that will only be shared under controlled conditions:
"I understand that only health professionals and members of the research team will have access to my personal data and that the overall anonymous data from the study may be seen and used by other researchers, for ethically approved research projects, on the understanding that confidentiality will be maintained."

2: Withdrawal of research data from study (Option one)

If identifiers are *irreversibly* removed during the course of a study, data contributed by a particular individual cannot then be identified and subsequently removed. In this case the following phrase may be appropriate:
"Please confirm that you understand that you may withdraw your data, without giving a reason, until the point at which your data is anonymised. After this point it will no longer be possible to identify your data".

3: Withdrawal of research data from study (Option two)

If identifiers are instead *reversibly* removed during the course of a study, it remains possible to be identified and subsequently remove data regarding a particular individual, up until the point that the data is fully anonymised for publication. In this case the following phrase may be appropriate:
"Please confirm that you understand that you may withdraw your data, without giving a reason, until the point at which your data is made publicly available in an anonymised form".

Note:
Under no circumstances should participants be given the option to withdraw their data after data has been published openly. It is impossible to recall publicly available data.