Sharing of data from clinical research projects – guidance from the Swiss CTU network

Authors (there might be changes in sequence)

Brigitta Gahl, PhD[1], Alan Haynes, PhD[1], Constantin Sluka, PhD[2], Elise Dupuis-Lozeron, PhD[3], Francisca Jörger, PhD[4], Renate Schur, MSc[4], Andri Christen, PhD[5], Sven Trelle, MD[1]

[1] CTU Bern, University of Bern, Switzerland

[2]Clinical Trial Unit, University of Basel, Department of Clinical Research, University Hospital, Switzerland

[3] Clinical Research Centre, Department of Health and Community Medicine, University of Geneva & Geneva University Hospitals, Geneva, Switzerland.

[4] Clinical Trials Center, University Hospital Zurich, Zürich, Switzerland

[5] Swiss Clinical Trial Organisation SCTO, Switzerland

Table of Contents

1. Abstract 4
2. Abbreviations 5
3. Boxes with recommendations 6
4. Introduction 7
5. Legal basis in Switzerland 9
6. Informed consent 11
7. Data management plan 13
8. Anonymization 15
8.1. Goal 15
8.2. Identifying variables 15
8.3. The process of anonymizing data 16
8.3.1 Assessment of the data 16
8.3.2 Detailed specification of required data processing steps 17
8.3.3 Data processing 17
8.3.4 Quality control 17
9. Data structure and format 20
9.1. Data structure 20
9.2. Data format 20
9.3. Character encoding 21
10. Coding of variables 22
10.1. Variable types 22
10.2. Variable labels 23
10.3. Time structures in the data sampling 24
11. Metadata and documentation 25
11.1. Metadata schemes 25
11.2. Additional documentation 26
11.3. Is statistical analysis code needed for data sharing? 27
12. Version control 28
13. Selection of repository 29

13.1. Principles 29
13.2. Time point 30
13.3. Identifying potential repositories 30
13.4. Selection criteria 30
14. Requesting and use of data 32
15. References 34
16. Glossary 37
17. Appendix 44
17.1. Further detailed specification of required data processing steps 44
17.1.1 Example data to be considered for deletion 44
17.1.2 Examples and details on manipulations to decrease precision 44
17.2. Further details on coding of variables 46
17.2.1 Formatting of date and time variables 46
17.2.2 Examples for further documentation of the dataset 47
17.3. Meta data scheme from ISRCTN 51
17.4. Information required for additional documentation 53
17.5. Checklist for selecting a data repository 54

1. Abstract

Objectives: Data sharing has become a requirement of many funding bodies and is becoming a scientific standard in many disciplines. In medical research, however, data sharing can conflict with clinicians' obligation to protect patients' privacy. General recommendations on data sharing exist also for clinical research, but so far lack practical and Swiss-specific aspects. The objective of this document is to provide practical recommendations for all relevant aspects of data sharing in agreement with legislation in Switzerland.

Methods: This document was written by members of the Swiss CTU Network, a network of academic clinical trial units. The process did not follow a formalized Delphi process. After an internal consensus round, this report is now published as pre-print for external review. A second version will incorporate external comments.

We plan to publish this document as a text in progress, as we expect relevant changes in related fields such as the development of further dedicated medical repositories or methodological advances in anonymization techniques.

Results: We developed principles and practical recommendations with respect to informed consent, data management plan, anonymization, data structure and format, coding of variables, metadata and documentation, version control, selection of repository, requesting and use of data. We also provide a summary of legal aspects relevant for the Swiss context.

Conclusions: The intension to share data has an impact not only after a clinical trial or an observational study is completed, but also during the planning period, the conduct and the analysis phase. Clinical researchers need to be aware at the beginning of a study on how to inform patients and at least the amount of work related to preparing data for sharing, metadata, and any further documentation. This report provides details of aspects to be considered, suggests decision criteria, and provides examples and checklists, in order to support data sharing in practice.

2. Abbreviations

ADaM.....Analysis Data Model
AHV .....Alters- und Hinterlassenversicherung (*Old Age Insurance*)
API .....Application Protocol Interface
CDASH .....Clinical Data Acquisition Standards Harmonization
CDISC .....Clinical Data Interchange Standards Consortium
CERN .....Conseil Européen pour la Recherche Nucléaire (*European Organization for Nuclear Research*)
ClinO .....Clinical Trials Ordinance
CRF .....Case Report Form
CSV .....Comma Separated Value
CTU.....Clinical Trials Unit
DAC .....Data Access Committee
DMP .....Data Management Plan
DOI.....Digital Object Identifier
ECRIN.....European Clinical Research Infrastructure Network
ELSI .....Ethical, Legal and Social Implications
EOSC.....European Open Science Cloud
EU.....European Union
FAIR.....Findable, Accessible, Interoperable, Reuseable
FADP.....Federal Act of Data Protection
GCP .....Good Clinical Practice
HIPAA.....U.S. Health Insurance Portability and Accountability Act
HRA .....Human Research Act
HRO .....Human Research Ordinance
ICMJE .....International Committee of Medical Journal Editors
ICPSR.....Inter-university Consortium for Political and Social Research
ICTRP.....International Clinical Trials Registry Platform
ID.....IDentificator
IP .....Intellectual Properties
IPD .....Individual Participant Data
ISRCTN.....International Standard Randomised Controlled Trials Number
KDSG.....Kantonales Datenschutzgesetz (*Cantonal Act of Data Protection*)
MedDRA .....Medical Dictionary for Regulatory Activities
PID .....Patient IDentificator
SAP.....Statistical Analysis Plan
SDTM.....Study Data Tabulation Model
SNOMED.....Systematized Nomenclature of Human and Veterinary Medicine
SPHN.....Swiss Personalized Health Network
TSV .....Tab Separated Value
UK .....United Kingdom
URL.....Uniform Resource Locator
US.....United States
UTF .....Unicode Transformation Format
WHO .....World Health Organisation
XML .....eXtensible Markup Language
ZIP.....Zone Improvement Plan

3. Boxes with recommendations

Box 1: Recommendations concerning consent 12
Box 2: Recommendations concerning the data management plan 13
Box 3: Recommendations concerning anonymization..... 18
Box 4: Recommendations on data structure and format 21
Box 5: Recommendation concerning variables within a shared dataset 24
Box 6: Recommendation for metadata and additional documentation 26
Box 7: Recommendation regarding availability of analysis code..... 27
Box 8: Recommendations selection of repository 31

4. Introduction

Sharing of research data[Glossary][1] has evolved as standard practice in many disciplines. There are two main drivers for the current data sharing movement: one is to enable reproducibility of research results with the goal of increasing transparency and credibility of science (1); the other is to enable reuse of data for new research questions. The International Committee of Medical Journal Editors (ICMJE) considers it an ethical obligation to responsibly share data generated by clinical trials (2). The main reason for this opinion is that trial participants have put themselves at risk by accepting to receive a treatment under study. Slightly different means are needed to achieve these two goals with respect to the amount of data to be shared and with respect to additional documentation that comes with the shared data.

The purpose of this document is to give specific recommendations for each decision that has to be made when sharing data from a clinical trial or an observational study, be it individual participant data (IPD) or aggregate data. A principal investigator (PI) or other study team members who intend to share data should find answers to their questions in this document and whatever else there is to consider. We provide minimal options to fulfil current requirements to still allow for the culture of data sharing to evolve.

Important papers on sharing individual patient data were published. Ohmann et al. developed a consensus document for sharing individual patient data using a consensus-building process among an interdisciplinary task force of research professionals as part of an European project (3). The document provides 10 principles and 50 recommendations to support data sharing and remove obstacles on many different levels such as collaboration culture and incentives, but also on technical and organizational aspects for "making data sharing a reality" (3). Our own statement is rather dedicated to the reality of data sharing clinical researchers are facing. It is our conviction that the establishment of data sharing will affect collaboration among clinical researchers and involvement of the research community.

The FAIR principles (4–6) provide guidelines to improve the Findability, Accessibility, Interoperability, and Reuse of data. They were developed for scientific data in general and focus on machine-operability and the order of the letters represent the dependency of the principles, e.g. data must be Findable to be Accessible, must be Findable and Accessible to be Interoperable and must be Findable, Accessible and Interoperable to be Reuseable. Even though it is very normal for us all to search for digital objects such as scientific papers in a database, this is more complicated when it comes to data objects[Glossary]/artifacts[Glossary]. When retrieving data, a whole package of related descriptions and documentation is needed to understand the data and allow its reuse. As a consequence, structure is needed and thus more rules for data provision. The choice of repository already determines many aspects of findability and accessibility. Usually, a repository has a metadata[Glossary] scheme (see sections 11and 13) that might be specific to the field and hence allows for specific searches. The repository might be linked to other systems to allow for parallel searches in several repositories (see section 11).

[1] Terms that are defined and further explained in the glossary, are marked with [Glossary] whenever they appear for the first time in the text.

Accessibility follows from the data requesting process as defined by the repository. Interoperability of data basically relates to the format, structure, coding, and documentation and is covered in sections 9and 10. Reuse of data depends on many factors and, particularly, of the three other principles. We want to point out that *reproducibility* in the context of shared data might not mean that researchers redoing exactly the same analyses will end up with exactly the same result for each estimate. If precise data such as biomarkers have been jittered or grouped as a means of anonymization[Glossary], derived and estimated values might differ from published values. It is important that this difference is mentioned and quantified in the documentation.

This document was written by Swiss professionals in the field of academic clinical research. Involved persons were identified within the Swiss CTU network and delegated from each clinical trials unit participating in the network. The authors identified relevant topics to be covered and assigned each topic to an individual author. During the writing period, further topics were identified and added. The document was merged and the different parts were consolidated by three members of CTU Bern. Then all authors were asked for feedback to the entire document. The three members of CTU Bern incorporated all feedback reaching consensus among all authors in most aspects (unless explicitly indicated). Afterwards this document underwent language review and is now to be published on https://www.medrxiv.org/ for invited expert review, after another co-authors round. The whole process did not follow a structured Delphi process.

We plan to publish this document as a text in progress, as we expect relevant changes in related fields such as the development of further dedicated medical repositories or methodological advances in anonymization techniques.

5. Legal basis in Switzerland

Health-related personal data are considered sensitive data in Switzerland. According to Article 4 paragraph 3 of the Federal Act of Data Protection (FADP) (7), personal data may only be used for the purpose a) indicated to subjects at the time their data are collected, b) that is evident from the circumstances, or c) that is required by law. The use of health-related personal data for research purposes is specifically laid down in a so-called special law, the Human Research Act (HRA) (8).

The Act regulates biomedical research on human subjects at the federal level and is based on internationally recognized principles. Sharing health-related data fulfills criteria for Further Use[2] (9) according to the Act and is regulated by Chapter 4 of the HRA (Art. 32-35). Further Use presupposes that the data are already available, i.e. collected with the necessary justification for another purpose and stored and made available (Art. 24 Human Research Ordinance, HRO, CC 810.301). If data sharing is planned at the time of data collection, e.g. for a clinical trial, the participants must be informed and consent obtained about the intended reuse of the collected data and their right to dissent to that at the time of collection. Article 17 of the HRA states: "If the intention exists to make further use for research of … health-related personal data collected, the consent of the persons concerned must be obtained at the time of such sampling or collection, or they must be informed of their right to dissent.“ However, consent for further use/sharing of data should not be an inclusion criterion for a trial; individuals must be given the possibility to participate in a trial without giving consent for data sharing later. In exceptional cases and under given circumstances (e.g. approval by an ethics committee), the law allows the reuse of health-related data for research that was collected without explicit consent provided it is impossible or very difficult to obtain consent or to provide information on the right to dissent, or this would impose an undue burden on the person concerned. In addition, the privacy and fundamental rights of the individuals must always be ensured (Art. 34 HRA). If the intention is to share only coded data and the data do not contain genetic data, information about potential further use is sufficient unless a participant explicitly disagrees. Explicit consent is not required (Art. 33 HRA).

If personal data are disclosed abroad, adequate data protection must be ensured (Art. 6 Federal Act of Data Protection). Adequate data protection should be part of any data use agreement (see section 12).

Anonymous data, which are not personal and cannot harm persons, are subject to neither FADP nor HRA, may be freely shared. However, as described below it is typically not possible to ensure that individual patient data will remain anonymous for all times to come (see section 8).

Infobox 1: Swiss legal basis in a nutshell

1. Data sharing is considered further use.

[2] The concept of Further Use also applies to biological material but this is not discussed in this statement. The statement is specific for data sharing aspects and does not cover other aspects of Further Use.

2. Consent for data sharing should preferably be obtained at enrolment.
3. Anonymous data does not fall under FADP nor HRA. It is unlikely that individual patient data of a clinical study can be anonymized[Glossary].

248

6. Informed consent

The sharing and use of personal health data for research has implications for patients' rights and interests. The legal requirements for patient information and consent are laid down in the Swiss Federal Act on Data Protection (FADP) and the Human Research Act (HRA) (see section 5).

The Ethical, Legal and Social Implications (ELSI) advisory group, which is part of the Swiss Personalized Health Network (SPHN) (10) initiative, published a framework providing ethical guidance on processing and sharing personal data within SPHN hereafter referred to as the ELSI framework (11). The document takes into account both international guidelines and national law including the HRA with a specific focus on aspects of general consent: "[The] Framework refers to all data types … that can be employed in the context of health research. This includes health-related personal data … that were not originally collected for research purposes, …". The ELSI advisory group considers a general consent (Broad Consent) sufficient for further use of encoded data outside the institution regardless of the original collection purpose and whether data are genetic or nongenetic (*ELSI framework III-1, Guidelines point b).* It is important in this context to have an unambiguous understanding of the term *general consent*. This term is often used in biobanking and related to further use of health-related data and samples collected in routine medical care (12). As described in section 5, sharing data from clinical research projects requires explicit informed consent because the consent given by the patient allows the use of the data to answer the questions/objectives of the project and does not extend to other research purposes. A general consent that was given in the context of routine medical care, for example at time of admission to a hospital, is insufficient for the purpose of sharing clinical trial data. The ELSI Advisory Group provides a broader definition of the term, and states that general consent means "informed consent of a research participant to unspecified further research uses of his or her health-related personal data or human biological material" (in the international academic literature, the closest term to general consent is broad consent). See (9) 3.3.1 for a definition of "further use." In this sense, the framework is applicable to the sharing of clinical trial data. As described in section 5, information and consent about possible data sharing should be done at project enrolment.

Sharing of coded or personal health-related data requires that the transfer of data is traceable at any time to ensure the patients' personal rights to provide information on the type, storage, and reuse (sharing) of her/his data on request or to ensure that data will no longer be available for research if the consent for reuse is revoked (*ELSI framework III-1, III-4)*. This is only feasible if the data are either anonymized (which is in general not achievable, see section 6) or if data are shared on the basis of a contract that we consider the default (see section 14). The sponsor-investigator providing data needs adequate governance structures in place to maintain control over the data such as data sharing agreements[Glossary] specifying the intended use, confidentiality, and the obligation to delete data of persons revoking consent and compliance with data protection. As in all situations, revoked consent has to be immediately addressed (*ELSI framework* says "revocations […] are swiftly acted upon"), but not retroactively. Specifically,

the patient consent status at the moment of database export is relevant. If a patient does not give consent, it should be documented when the patient was asked and what he or she was informed about.

Box 1: Recommendations concerning consent

R1. Sponsor(-investigator)s must ensure that participants are informed about potential data sharing and further use of their data at the time of enrollment in a clinical research project including anonymization of their data.

R2. If sharing of coded data is planned (the usual case):
   a. Sponsor(-investigator)s must ensure that potential participants are informed about the potential sharing of their data. Explicit consent is not needed but the possibility to disagree must be ensured.
   b. Sponsors/sponsor-investigators should ensure that a system is in place that allows access to this information centrally, e.g., by recording disagreements in the study database.
   c. When sharing coded data, the sponsor(-investigator) must have a system in place to ensure participants' rights especially with regard to the use of their data and deletion of the data wherever the data were transferred to.

R3. If sharing of uncoded personal data is planned (not recommended):
   a. Sponsor(-investigator)s must ensure that potential participants are informed about the potential sharing of personal data and the potential anonymization of their data for this purpose. Explicit written informed consent should be sought.
   b. Sponsors/sponsor-investigators should ensure that a system is in place that allows access to consent status of each patient centrally, e.g., by recording the information in the study database.
   c. When sharing personal data, the sponsor(-investigator) must have a system in place to ensure participants' rights especially with regard to the use of their data and deletion of the data wherever the data were transferred to.

R4. For sharing data collected in the setting of clinical routine a general consent of a patient is sufficient unless it explicitly excludes data sharing, the general consent used in the hospital has to be carefully checked.

R5. It is imperative to take into account the consent status of patients. If a patient withdraws consent, data of this patients have to be ignored immediately from the moment of withdrawal on, but analysis already done or data files already provided do not have to be changed.

7. Data management plan

According to (13)(v3.1.0), a Data Management Plan[Glossary] is a document "to identify the overall strategy for data management processes for the trial; a compilation of documents that may include amendments/appendices but are not limited to: Completion Guidelines, Data Quality Plan, CRF Design Document, Database (build) Specification, Entry Guidelines, Database Testing". The Data Management Plan therefore provides an overview of all aspects related to data (management) in a clinical research project. Depending on the details provided in the study protocol, a Data Management Plan might not be needed. However, we recommend that all studies have a Data Management Plan because this supports and facilitates later data sharing activities. Several templates for such a document are freely available over the Internet. We do not recommend a particular one. However, the plan should cover the aspects relevant for data collection, handling, and storage during study conduct (and implementation/conclusion) as well as for data sharing. A possible structure and description of content is shown below. It should be noted that there are now specific journals that specialize in publishing articles on description of datasets and aspects of data management. We make no specific recommendations on this.

Box 2: Recommendations concerning the data management plan

R6. All aspects related to data management including data sharing should be documented before conducting a clinical research project. The document should be considered a living document and regularly updated using a version control system. It might be called Data Management Plan.

R7. Possible structure and content of a data management plan. Not all sections will be relevant to all research projects:

1. Introduction
2. Responsibilities
3. Description of collected/generated data
4. Case Report Form[Glossary] development
5. Clinical Data Management System – study specific implementation
   - 5.1. Implementation of the study database in the Clinical Data Management System
     - 5.1.1. Codebook development
     - 5.1.2. Clinical Data Management System implementation
     - 5.1.3. Medical coding
     - 5.1.4. Data import
   - 5.2. Verification of Clinical Data Management System setup and deployment
   - 5.3. Change management
6. Clinical Data Management System – infrastructure
   - 6.1. Data storage
   - 6.2. Data back-up
   - 6.3. Access to the data

6.4. Granting access to the productive version of the Clinical Data Management System and database
7. Data collection
7.1. Pre-requisites for data entry
7.1.1. Data entry guidelines
7.1.2. Training of users and training documentation
7.2. Entering data
8. Quality control procedures
8.1. Real-time data validation
8.2. External data validation (offline checks)
8.3. Central data monitoring
8.3.1. Definition of Key Performance Indicators (KPIs)
8.3.2. Frequency
8.3.3. Reporting
8.3.4. Clinical Data Management System generated, automatic queries
8.3.5. Manual queries
8.3.6. Follow-up on (persisting) data discrepancies
9. Database closure
9.1. Pre-closure data checks
9.2. Quality assurance audit and database lock
9.3. Database unlock
10. Data transfer and exports
10.1. Data requests and transfer
10.2. Data exports
10.3. Export validation
10.4. Adverse event data reconciliation
11. Clinical Data Management System archiving and provision of final materials to the sponsor
12. Data preservation
13. FAIR data sharing
13.1. Repository[Glossary]
13.1.1. Shared artifacts
13.2. Data request process
13.3. Ethics, legal and security issues
13.3.1. Data protection
13.3.2. Copyright and intellectual property

8. Anonymization

8.1. Goal

Anonymization is the process of handling personal data in such a way that identification of individual persons is impossible or possible only with disproportionate effort. Further data sources and technologies for data linkage might become available at some point, thus the effort needed to identify persons is not known for all times to come (9). As a consequence, data that are anonymized today might not remain anonymized according to this definition. Anonymized data in the strictest sense can be shared openly, but the claim that individual patient data are anonymized is not realistic (14,15). The process we describe below aims at sharing data with researchers in a standardized and institutionalized way based on a standard contract or license in which the data requester agrees not to try to identify patients, not to give the data to other persons, and to maintain data security. In this setting and with these restrictions, present-day anonymization can be considered acceptable. The goal of the anonymization process we describe is to protect participants' privacy to a degree that criminal acts would be necessary for identifying patients at time of sharing the data to identify participants.
Obviously, the anonymization process consists of manipulations that *change* the data. In order not to spoil the benefit of data sharing, it is important to consider the goal of anonymization: protection of patients' privacy, while also considering the usefulness of the data. Of note, anonymizing data needs a lot of work.

8.2. Identifying variables

Variables[Glossary] are called directly identifying if they contain personal information by which a participant can be identified with little or no effort, and should in general not be stored within the study database or, if stored, not be possible to export. The Human Research Ordinance mentions explicitly the following data (Art. 25, Paragraph 2): name, address(es), date of birth, unique identification numbers. The U.S. Health Insurance Portability and Accountability Act (HIPAA) provides more details. The following is a non-exhaustive list (16):

- Real names
- All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older.
- Addresses and geolocations/-codes past and present (canton/state might be allowed given that the geographic unit contains more than 20,000 persons; MEDSTAT regions might be more appropriate as they were designed to ensure anonymity (17)).
- Telephone number, email addresses, IP addresses, or any links or aliases/pseudonyms[Glossary] e.g. Facebook, LinkedIn, WhatsApp, Twitter, or links/URLs to personal websites.

- Device/implant identifiers[Glossary] and serial numbers or vehicle identifiers and serial numbers, including license plate numbers.
- Any other (non-health) personal identifier (ID), e.g. hospital ID (or PID), social security numbers (AHV), insurance numbers, passport numbers, account numbers, etc.
- Full-face photographs and any comparable images or biometric identifiers including finger and voice prints.

Identifying data[Glossary] can be variables containing information that are by definition unique to the patient, and therefore the patient can be identified with medium effort, e.g. genetic, genomics, metabolomics, proteomics, micro-array, biomarkers, or similar high-precision data.

Identifying data can be variables containing information which singly or in combination with other data, can be used to identify the patient with some effort (indirect identifiers), e.g.:

- Marker of rare disease or subtype of disease
- Rare medication, treatment, or surgery
- Rare diagnostic tool or machine used
- Rare population
- High-precision variable (while precision depends on the type of data)
- Any unusual variation or combination of variables as mentioned above

8.3. The process of anonymizing data

Anonymization is a multistep process that requires input by several people, among them the sponsor and the statistician. The shared data set should in principle contain only the data that are needed for the intended purpose. For example, to share a dataset underlying a scientific report only the data needed to reproduce the statistics, graphs, tables etc. in the report should be in the dataset[3].

8.3.1 Assessment of the data

It is necessary to assess the whole dataset with all individual variables. This is best done by a statistician or data manager and by the sponsor (because content knowledge might be needed). HIPAA states three criteria relating to a variable or a set of variables that might serve as guidance to assess the risk of re-identification:

1. Replicability: How consistently is a piece of information related to a specific person? For example, while laboratory values vary (low replicability), demographics are more stable (high replicability).

[3] Note that (3) refers to the danger that records in a shared data file might be selected because they are "supporting the conclusions of a specific published paper" (p. 2). This is not an issue of this paper because CTUs conduct analyses according to pre-specified inclusion criteria and are neutral with respect to expected results.

2. Data source availability: Which external data sources could be used to identify a specific person? For example, demographics could be obtained from public registries.
3. Distinguishability: How many persons share a specific combination of characteristics? For example, year of birth and canton is less likely to be unique than complete date of birth and ZIP code.

These criteria are relevant to assess the risk of a linkage attack, the process of re-identification by linking an external data source with person-identifying data to the original data set. In the last decades, several cases of successful linkage attacks have been recorded (18). For example, in 2013 5–7 laboratory values from a known patient were shown to identify the corresponding records in a de-identified[Glossary] biomedical research database (19).

Each variable should be classified whether it is:

- (Potentially) Directly identifying (see section 8.2),
- Indirectly identifying, i.e. identifying in connection with other variable(s). The other variable(s) should be documented, or
- Unproblematic, i.e. neither directly nor indirectly identifying.

8.3.2 Detailed specification of required data processing steps

After categorization, the necessary data preparation steps for the directly and indirectly identifying variables must be defined. This is a non-exhaustive list of potential procedures:

- Deletion: Variables containing directly identifying information unsuitable for manipulation must be deleted. The appendix provides some examples.
- Irreversible pseudonymization[Glossary]: Irreversible pseudonymization is a transformation of a variable into a new variable, where the mapping which renders the process reversible is deleted (database dependent). This usually requires a complex algorithm and is rarely used.
- Manipulations to decrease precision: Too much precision bears the risk of making entries linkable to persons. Possible methods to decrease precision include relative time in the course of the study instead of precise dates and times, rounding of continuous data, grouping and aggregation (categorization), introducing random noise (jittering, perturbation), setting certain values to missing (suppression), data swapping, resampling or subsampling.

The Appendix provides additional details and examples.

8.3.3 Data processing

The steps as defined in 8.3.2 have to be programmed using statistical software and a set of new data files has to be generated.

8.3.4 Quality control

Two persons should perform a quality control and check the de-identified data:

1. Sponsor

   In particular to check:

   - Whether the de-identified data set contains free text variables, in which the text may potentially lead to identification
   - Whether this data set contains other variables which may alone or in combination lead to identification, in particular if infrequent/rare disease or population is involved
   - Whether data need to be lumped into categories

2. Statistician or representative knowledgeable of the data set (e.g. Central Data Monitor, Monitor, Data Manager)

   In particular to check:

   - That any combination of indirectly identifying variables results in a number >1 (e.g. five) records
   - Whether the de-identified data do not contain personal information variables except age without any digit (but not date of birth)
   - Whether the file only contains text variables if specifically requested and that those text variables are appropriately redacted
   - Whether digits have been removed/rounded/jittered
   - Whether dates have been replaced
   - Whether the identification numbers have been replaced with a new random identifier

Whether results based on the new dataset are similar to results using the original dataset must be checked, and if not, where and to what extent they deviate and any deviations should be noted in the same document where the assessment and specifications are described (steps 1 and 2). Every analysis need not be run. Common sense should be applied to select important ones.

The statistician corrects the de-identification[Glossary] coding according to the recommendations resulting from quality control.

If all is in order, the two persons sign a quality control document with a date to document that they did the quality control and what was checked. If multiple (repeated) exports need to be done based on the same code, then this quality control needs to be done only once, except if the sponsor requests a check at each export.

Box 3: Recommendations concerning anonymization

R8. Anonymization should involve at least the sponsor and the statistician.

R9. Directly identifying variables should be removed, IDs should be replaced by random numbers, string variables should be removed, and rare combinations of values identified and lumped together to achieve larger groups of patients.

R10. The anonymization process should be quality controlled and appropriately documented.

9. Data structure and format

Full descriptive information of the data is necessary (see the coding variables section 8) for reproduction of analyses as well as for reuse of the data, which are the two main purposes of data sharing. Details of the de-identification process should be provided for the sake of transparency.

Although the European Clinical Research Infrastructure Network (ECRIN) recommends the Clinical Data Interchange Standards Consortium (CDISC) format for sharing data (3), the use of this standard outside of the pharmaceutical industry is relatively rare, particularly in the academic setting where resources to set up CDISC-compliant databases are limited. While we agree that standardization of items and structure aids secondary data processing and reuse, the current reality is that academic databases are rarely (if ever) designed to CDISC standards. Furthermore, the CDISC defines a variety of formats such as the Study Data Tabulation Model (SDTM) and Analysis Data Model (ADaM) on the database side, and seven different extensible markup language (XML) based formats for data exchange. It is therefore a substantial challenge to understand the full CDISC standard structure, let alone work with it. That being said, utilization of certain features of the format is recommended (such as standardized variable naming and encoding). The ECRIN statement highlights that it is difficult to transfer data to a specific standard unless this is done from the project planning stage. Thus, as far as is possible given constraints of cost, time to implement, and technical capabilities CDISC standards should be employed for new trials at the database design stage.

9.1. Data structure

Clinical research projects typically involve multiple assessments over time (at least two different time points). Data in a study are usually collected on different forms within the case report form. The structure of the database usually reflects this structure, i.e. data are stored in separate tables and keys[Glossary] serve as the link between these tables (relational database). We recommend that the table structure is preserved when preparing a dataset for sharing, that is, each table remains a separate file within the dataset. Careful description of the keys is needed to ensure that users of the data are able to establish the correct link across the different files (see section 11). The original key-value pairs will usually be replaced with new random unique identifiers (see section 8).

9.2. Data format

For older projects, where CDISC standards were not considered, data would ideally be shared in a simple format. Text based comma separated value (CSV) or variants thereof (e.g. tab separated value, TSV) are non-proprietary formats which should be future-proof: changes in future versions of software will not render the data unreadable as they are text based formats. Other formats such as XML, while offering the ability to include data, audit trail, coding and

database structure, are potentially more difficult to work with. Indeed, some widely used statistical software packages have only very basic XML capabilities. Additionally, the FAIR principles suggest that the data should be usable by most users. Using formats such as XML requires a large degree of specialist knowledge simply to read the data into statistical software. Proprietary formats such as SAS, SPSS, Stata data files, and .xlsx files are also less suitable for sharing as they are generally only accessible using that software (although there are packages available for R to handle many formats), and are typically not suited to long-term storage due to changes between versions. As such, text-based formats such as CSV are preferable. There is, however, some variation in recommendations in this respect. While some institutional repositories recommend plain-text-based formats (Georgia State University, World Wide Web Consortium), many others recommend proprietary formats (Inter-university Consortium for Political and Social Research, ICPSR) or a wide range of formats including text-based and proprietary formats (Oregon State University, Stanford University, UK Data Service). Most also suggest delimited text (e.g. comma separate value format) with setup files (codes to read data in and prepare it). However, setup files are containing software code and the programming language dictate which programs can use the data. In general, data are more ready to use only if in the format of the statistical software used for the original statistical, and data are more accessible in any non-proprietary format. We therefore recommend that data are provided in the format as used during analysis and in comma-separated value format. Metadata and documentation should be uploaded in separate files along with the data (see sections 10 and 11).

9.3. Character encoding

The encoding of files is also an issue, as it determines how special characters (e.g. ä, à, é, è, ö, ü) are interpreted by software. We recommend 8-Bit UCS Transformation Format (UTF-8) (https://en.wikipedia.org/wiki/UTF-8) (20) encoding where possible, as this is a widely recognized encoding system and supports the vast majority of characters. The encoding used should be explicitly stated, ideally in the data management plan.

Box 4: Recommendations on data structure and format

- R11. Retain the database structure in the shared data (five case report forms in the database make five tables in the shared data).
- R12. Use text based formats such as CSV to share data, encoded in UTF-8.
- R13. Also provide data in the original format.

10. Coding of variables

The way data are prepared for sharing affects its general usability as well as its interoperability. For data sharing purposes, as few changes as possible should be made to a dataset after exporting the data from the database as it may not be possible to anticipate all the ways in which data might be used further. Thus, in order to avoid wasted effort, we advise not to recode data for data sharing purposes (within the limitations imposed by anonymization, see section 8). The use of standardized or controlled vocabularies (e.g. SNOMED, MedDRA, CDASH) increases the interoperability of data. Therefore, we recommend the use of standardized vocabulary. However, this should be considered during database development, rather than coding the data afterwards. Some data manipulation and recoding is inevitable, though, when sharing data.

10.1. Variable types

Individual variables come usually in four main types: date/time, text, numeric and categorical (binary and ordinal variables can be thought of as special cases of categorical variables). Each type of variable should be handled in a specific manner.

Date variables should be converted into project days (i.e. days since informed consent or randomization, see section 8). There might be circumstances in which dates/times are necessary such as when seasonal effects are important, as are relationships to historical events. Under such circumstances, we recommend a slightly modified version of the ISO 8601 standard. Date/time variables can be subdivided into three units: date, time, and date-time, each requiring its own handling. The appendix contains further details on formatting standards.

Continuous variables are relatively simple; they should be provided as they are (e.g., 1.5). The number of decimal places should be the same for all observations (if the most precise observation is 1.5, then all observations should have one decimal place: 1.0 instead of 1). Note that it may be desirable to reduce the precision of some variables (see section 8).

Categorical variables comprise binary (yes/no), single-choice (male/female), multiple-choice or ordinal type variables (New York Heart Association Functional Classification scores to classify heart failure). They can typically be provided in two ways: a textual description (such as male/female or yes/no) or a numeric representation (e.g., 1 or 2). From a human readability perspective, it would likely be best to save the textual representation, but data saved in such a manner will typically be considerably larger than that saved with the numeric representation instead, and require more work to make it analyzable. It is thus preferable to save the numeric codes with an additional codebook to provide the meaning of the codes (see Appendix for an example: Table 3, p. 48). The codebook can then be used by statistical software to label the data when it is to be reused, albeit with a little programming. Multiple choice questions should be split into as many binary variables as there are options, e.g., if there are options of diabetes, previous myocardial infarction, and previous stroke there would be three binary variables,

interpreted as yes/no for each. Other methods are available, but require additional work to make them usable for analyses.

We advise that **free text** variables be removed (see also section 8). If the retention of free text is necessary, no special treatment beyond those measures outlined in section 8 is necessary.

Some database systems incorporate into the dataset system-level variables such as row numbers in all tables of a data export. Such variables are often of no use and can typically be removed, but this should be confirmed on a case-by-case basis. **Missing values** should be reported as "NA" and clearly distinguished from non-missing categorical answers like "unknown".

10.2. Variable labels

Variable descriptions are equally important. Without a meaningful name, it is difficult to guess what a particular variable refers to. Short names are preferable for statistical programming and database purposes (some software even imposes limits on the length of names), but this can obscure the meaning of a given variable. Thus, besides the codebook for the meaning of values of (categorical) variables, another file with the labels for each variable is required; for consistency, we call it a labelbook. The labelbook should contain the variable name as it exists in the data (e.g. mi) together with its description (myocardial infarction), any restrictions or dependencies (only if mi == Yes), whether or not the variable is optional, and perhaps some useful notes even if they might also be in other documentation such as the study protocol or data management plan. The level of detail provided in the description depends on context and is likely to evolve over time. We also suggest providing relevant links to the study protocol, for example highlighting endpoints such as "Primary endpoint as described on page XX of the study protocol". A column indicating the data type of each variable is also essential. Different databases use different terms for each type, so a more standardized set of terms is provided in the Appendix (Table 2, p. 47).

Of note, we do not list calculated fields here because calculated values returned from electronic data capture systems usually are being re-calculated using statistical software.

The appendix provides an example of a labelbook with information on the form/table where the variable is collected/stored, variable name, description/label, data type, unit, applicable value label name, and whether the variable is collected as stored or whether values are calculated/derived (Table 3, p. 48).

We would also recommend having a fully annotated version of the (electronic) case report form with example data. Annotations should include variable names, option values, and any logic which defines when a variable should be entered or when a variable/question is shown or hidden in the electronic case report form.

As mentioned previously, system variables can typically be removed as they often include potentially identifying information (at least for the study team). The golden rule, though, is that every variable that exists in the data should be described in the labelbook.

10.3. Time structures in the data sampling

If there is a time structure to the data such as multiple follow ups, it is mandatory to include a visit identifier in the data set which allows the discrimination of the visits for a participant. This is particularly important when an individual form is used multiple times. In principle, this can be done by using a key variable containing the visit identifier (long format data) or by a naming convention such as adding a number at the end of the variable name (stub) indicating the order (wide format data). To reduce empty cells, it is advantageous to separate data by form and we recommend providing data in long format although this must be assessed on a case-by-case basis. The appendix provides an example by looking at fictitious eligibility and blood label values forms (Table 4 and

Table **5**, p. 48).

Forms that do not fit into the normal visit structure (sometimes called unscheduled visits or log forms to record medication or events) can be supplied with a "position" variable to indicate the repetition number of the form (starting at either 0 or 1). The visit structure, definition of unscheduled visits and the starting indices should be reported in the data documentation (see section 11). In Table 6 (p. 50) we see that participant 1 reported taking a medication at two time points, while participant 4 reported taking morphine for a period of time, including changing doses. The remaining participants took no medications.

Another type of necessary information is information about which variables belong to which form, which can be captured in the labelbook, and which forms are collected during which visits. Following our previous logic, we call this a visitbook (Table 7, p. 50). It requires a column for the visit identifier, and a column for which forms occur in each visit. Each row indicates a visit-form combination (i.e. a visit could have multiple forms, and a form could be in multiple visits). An additional column with the name of the visit is also useful. There should also be a graphical representation of the visit structure as shown in of the Appendix (Table 8, p. 50).

Box 5: Recommendation concerning variables within a shared dataset

R14. Prepare data in a long format, with appropriate keys to link tables together.

R15. Document all variables in all tables, and the tables themselves.

11. Metadata and documentation

A data file alone is of limited use, so the concept of data sharing needs comprehensive documentation to go with the data (also see sections 9 and 10). This documentation serves two purposes. It should, on one hand, enable someone not involved with the study to understand and use the data appropriately. On the other hand, it should allow someone with a scientific question in the field of the study to find the data. This section gives a definition of the term metadata and what we think this metadata should contain, at a minimum, to go along with shared clinical research datasets. We also give recommendations with respect to the information and documentation that should be provided. This section therefore deals with findability and re-usability of the data, as claimed in the FAIR principles.

11.1. Metadata schemes

Metadata are data about data, typically structured information such as numbers or classification options, that describe a fixed set of aspects of a data object in a human and, importantly, machine-readable way (21). This definition is in accordance with the concept “metadata scheme” as used in libraries and repositories to denote the fields that describe the stored objects (21,22).

The main purpose of metadata is to find and describe a data object such as a data file, a document, or a whole shared package containing different types of artifacts. Because standardized metadata also allows for interoperability between systems, a data object can be made visible from other points of access (https://www.openarchives.org/pmh/) as far as the involved metadata schemes cover the same aspects. For example, it might be findable via repositories, databases, or registries.

Canham et al. (23) suggest the use of a minimal extension of the DataCite metadata scheme for clinical research data (24) which is a general purpose scheme. Study details can be found basically in one field ("A.3 Study topics"), and the description of the dataset hence remains somehow vague. We think that it is preferable to use a metadata scheme that supports more specific searches. We expect independent reuse of data to evolve into an established scientific research method also in clinical research, so we recommend a metadata scheme that allows researchers to a large extent decide whether or not data are relevant for their research purpose. The World Health Organization (WHO) set out requirements to describe a study (25) while the International Committee of Medical Journal Editors (ICMJE) provided guidelines (26). Section 17.3 in the Appendix shows the set of items required by International Standard Randomised Controlled Trials Number (ISRCTN) deemed essential to describe a study which we consider suitable for data sets in most respects. Provided that clinical trials should be registered in WHO compliant registries, this metadata is already publicly available and might be linked to a dataset in a repository via an application protocol interface (API) in the repository. If this is not available, the data should be entered manually. It is important to ensure consistency across the registry[Glossary] entry and any data repository entries. Although the scheme gives clear guidance

on what information must be provided, it does not mandate how. To improve findability it is recommended to use controlled vocabulary as far as possible. If controlled vocabulary is used, it is important to provide information to the underlying scheme that was used including the version.

11.2. Additional documentation

In addition to metadata, further documentation is needed to make use of the data. As described above in sections 9 and 10, codebooks, labelbooks, and visitbooks provide necessary information. Someone who wants to understand the data also needs to know how it was collected, which sources were used, what hierarchy there was among data sources, and the definitions applied. The context and purpose of the collection is important, as well as what methods were used to ensure data quality. Information that relates to the conduct of the research project is also needed, such as the reason for missingness of certain data and any adaptations that had to be made. If a new tool or drug is investigated, a comprehensive description/brochure of it is also mandatory. Furthermore, the details of data preparation have to be provided such as derivation of variables, and also the process of rounding or jittering data for de-identification (see section 8) has to be described together with its impact on the result, if applied (this information is typically part of the data management plan).

The study protocol and statistical analysis plan[Glossary] with amendments contain a large part of the information needed, but researchers have to carefully consider whether this information is enough for each individual project.

Box 6: Recommendation for metadata and additional documentation

R16. We recommend selecting a repository with a metadata scheme that allows for meaningfully detailed search on clinical studies (e.g., search options “patients condition”, “intervention”, “study endpoints”, etc.).

R17. We recommend as a minimum to upload with the data:

a. Readme file describing the data package and containing information to be shared and not contained in the other documents, ideally with a tabular summary of all files (Appendix, 错误!未找到引用源。)
b. Change log to capture different versions of the data set
c. Study protocol
d. Statistical analysis plan
e. Clinical study reports
f. Blank consent form
g. Fully annotated case report form (CRF)
h. Codebook, labelbook, visitbook
i. References to any standardized vocabularies or catalogues used
j. Code for data preparation
k. Description/brochure of a new tool or drug, if applicable

l. Documentation of means undertaken for anonymization
m. Data management plan

11.3. Is statistical analysis code needed for data sharing?

Note that we distinguish between data preparation code and analysis code, and we consider the preparation code to be necessary to go with the data (as it generates an analyzable dataset from the raw data). It is also possible to share the data file after preparation. We see different aspects involved in the question whether sharing analysis code is essential:

- Reproducibility: Undoubtedly, shared code allows for the most precise and quick reproduction of the results because certain analyses might be implemented differently in different software packages, and analyses can be done using different commands within the same software that might even have different implementations. Still, sharing code will often not lead to complete reproducibility because software versions and the underlying operating system might affect usability of the code.
- Detection of errors: Some errors in the analysis can only be detected when scrutinizing the code. Statisticians agree that wrong results are often due to errors in data preparation. From this point of view, sharing of raw data and data preparation code is preferable to sharing data after preparation. Reproducibility of results, even though desirable, does not mean correctness, but is a step in checking it.
- Additional information: Usually, a statistical analysis plan is available for a clinical trial describing in detail all analysis steps. However, statistical code might contain additional details not covered by the statistical analysis plan. Availability of statistical code is therefore essential to fully understand the analyses that were done.

Box 7: Recommendation regarding availability of analysis code

R18. In general, we recommend sharing of code with the dataset and recommend that statisticians keep to programming standards in the scripts, such as:

- Write a master script file that calls all script files of the analysis in the correct sequence.
- Follow a reasonable naming convention.
- Explain each step of the program in (extensive) comments.
- Check logical rigor of the entire code.

12. Version control

Version control allows one to track changes of objects or files through time. Because it may be difficult to tell whether a dataset has been used, simply replacing an object is likely to be undesirable as it would render the DOI referenced by the data user void (or rather, the DOI would be correct, but the dataset it referred to is no longer available or has changed). Version control may not be relevant for all datasets that will be shared. For example, a dataset that accompanies a publication would be unlikely to require version control as it is a static item—it does not change. Similarly, if a questionnaire performed and shared in 2017 was repeated in 2019 but the data were shared separately (2017 data not included), no version control is necessary (although it may be helpful to refer to the other dataset in the metadata). Conversely, extracts from registries might need version control if new data are periodically added to the dataset. Similarly, if the originally shared dataset from a clinical trial is shared but only some variables are cleaned and a second dataset is shared with all variables cleaned, this would ideally be a revision. New data (variables or observation) or changes to data are reasons to make a new version. Replacing only parts in the data object is easier than creating a whole new data object.

Where version control is considered necessary, a new DOI should be assigned to the object. Ideally the new objects DOI would indicate that it is a child of the original object. For example, dataset X is assigned a DOI of 1234. A year later, new data are added to X and the dataset is shared. A DOI of 1234.1 would indicate that it is a child of the original dataset (the main part of the DOI has remained the same, but has an extra part appended). If this is not possible and the new dataset is assigned a completely different DOI (e.g. 5678), then the original DOI should be entered into the metadata of the new dataset, and vice versa, to establish a link between the objects.

Box 9: Recommendations regarding version control

R19. Objects whose content has changed - new data appended to the original dataset (variables or observations) should be versioned.

R20. A related DOI should be assigned to the new dataset, rather than creating a whole new object. At the minimum, the DOI of the different versions should be stored in the metadata of all objects.

13. Selection of repository

Infobox 2: Data repository versus (clinical trial) registry

Registries: A clinical trial registry is a collection of records about clinical trials according to an agreed upon set of metadata (27). In registries accepted by the World Health Organization (WHO) and included in their International Clinical Trials Registry Platform (ICTRP), see 9.1, these records contain a minimum amount of information as defined in the WHO Data Set (25). As of 2019, this data set does not define or require attached artifacts or files. Confusingly, the WHO calls the database behind its Search Portal "Central Repository" (27), when it is in fact a registry.

Data repositories: In contrast, a data repository is a (digital) collection of digital datasets. Although not mandatory, the term nowadays implies a function to make these datasets findable, accessible, and reusable (5) and allows for longer term storage. Technically, a repository consists at least of a backend, a database to store metadata and information, and file server to store the datasets and other digital artifacts, and a web-based frontend that allows users to access the backend.

13.1. Principles

According to the FAIR data principles, research data should be findable, accessible, interoperable, and reusable (4,5), see section 2. Principle F3 mandates that "(meta)data are registered or indexed in a searchable resource" (4). Although the principles do not explicitly mention data repositories, principle F3 implies that research data should be stored in an appropriate repository that follows all principles (5). The European Clinical Research Infrastructure Network (ECRIN) data sharing statement is more explicit and states, that "data and trial documents made available for sharing should be transferred to a suitable data repository" (3) and we support this view. According to the FAIR data principles, research data should be findable, accessible, interoperable, and reusable (4,5), see introduction. Principle F3 mandates that "(meta)data are registered or indexed in a searchable resource" (4). Although the principles do not explicitly mention data repositories, principle F3 implies that research data should be stored in an appropriate repository that follows all principles (5).

When selecting a repository, clinical researchers therefore should ensure that the repository respects all FAIR data principles as a minimum. Although there are alternative initiatives like CoreTrustSeal (28), the FAIR principles seem to be the most widely accepted. However, other initiatives might evolve over time and become generally agreed standards. Given the lack of generally agreed standards and certification processes, researchers will need to assess the suitability of a repository for their purposes.

13.2. Time point

Ideally, the appropriate repository is identified before writing the Data Management Plan (see section 7) and then described therein. We assume that a sponsor/investigator uses the same repository for all her/his projects so this should be feasible.

13.3. Identifying potential repositories

So far, no repository exists that is specific for clinical research projects. Therefore, clinical researchers need to identify an appropriate repository by themselves. Many institutions involved in clinical research, like universities, currently maintain their own institutional repository. This might be a good starting point in the evaluation process. Alternatively, universities usually have a central contact point that supports researchers with issues related to data sharing and open science in general (29).

For projects that were funded by extramural grants, there might be specific requirements for a repository or even a specific repository mandated. For example, the Bill & Melinda Gates Foundation maintains a list of approved repositories for publications published in Gates Open Research (30). It is also expected that the planned European Open Science Cloud (EOSC) will affect how data from projects funded by the European Union will be shared (31). Repository registries maintain a searchable database of repositories. The largest one is probably r3data, a collaborative project of large European academic institutions. R3data can help locating topic specific repositories, which may be a better choice than an institutional repository because data are more likely to be found in a search for that particular topic. Furthermore, Swiss academic research institutions are currently developing a digital repository for long-term preservation and publishing of research data, Olos (32), to support the publication needs of funders and help researchers to manage research data.

Another choice might be Zenodo, which is based at CERN (European Organisation for Nuclear Research). There are also for-profit/commercial repositories such as FigShare and Dryad, although we do not explicitly recommend their use.

13.4. Selection criteria

After having identified a set of potential repositories, a researcher will need some explicit criteria to select a repository. We suggest an approach to structure this process which is based on a report by the Digital Curation Centre in Edinburgh (33), shaped as a checklist (Table 10, p. 54). Some items are very specific, others cannot be defined exactly and require adaptations on a project basis and not all aspects might be assessable.

Another useful resource are the levels of digital preservation by the National Digital Stewardship Alliance (34).

Box 8: Recommendations selection of repository

R21. Select a suitable repository, and include this information in the data management plan. Institutional repositories might be a good choice.

R22. Make data as open as possible, but as closed as necessary (FAIR)

765

14. Requesting and use of data

Principle 6 of the ECRIN statement (3) states: “In the context of managed access, any citizen or group that has both a reasonable scientific question and the expertise to answer that question should be able to request access to individual participant data and trial documents.” This begs the question of who decides whether a question is reasonable and an individual/group has the relevant competencies. Decisions made by the original project team could be seen as biased. Accordingly, the ECRIN statement (3) suggests that ideally each repository would have independent boards to assess the "scientific merit, potential impact and appropriateness of the proposed secondary analyses". With slightly different priorities, such a board might also be referred to as Data Access Committee[Glossary] (DAC). A DAC might evaluate and approve data requests within a reasonable response time. This would of course require separate boards or DACs for different subject areas. From our point of view, it is a good idea to have a board of specialists/DACs supporting new research on existing data, but it might be difficult to find the resources for this work. From a legal point of view, there are few minimal requirements that have to be fulfilled in order to receive data:

1. The data requester has to confirm that the purpose of the data request is scientific, that the research project will be conducted in accordance with the local legislation (Human Research Act, authorization from ethics committee) and rules of conduct (Good Clinical Practice). Any different purpose would have to be explicitly mentioned in the informed consent (see (35) and see section 6).
2. The data requester has to confirm that she/he:
    - 2.1. Will not try to identify individual persons in the data
    - 2.2. Will not give the data to other persons
    - 2.3. Will maintain data security
    - 2.4. Will report any accidental finding to the data provider

We think that publishing the metadata and sharing the data after checking these two requirements will be the usual process in clinical research. The requesting process is obviously determined by the repository, so we only sketch some possible implementation options. With minimal use of resources, requirement 1 might be covered by a checkbox on the request form that a requester has to tick. If she/he does not, a pop-up window might occur saying that the request is going to be rejected. Requirement 2 needs the requester to be a person able to confirm in a legally binding way. There are established ways to check whether an action is done by a human over the Internet, but in the context of data sharing we assume by default that the requester has an academic affiliation, which will be used to verify the requester’s identity. A requester without academic affiliation might turn to the data provider directly. The requester might confirm the items of requirement 2 by signing a contract or by ticking a checkbox of a license agreement (35). The agreement might contain an example text of how the original study

and its investigators should be acknowledged in any kind of publication to ensure that data generators receive appropriate recognition (36). All requests are stored by the repository to be traced by interested persons such as the principal investigator. If there is a DAC/board of specialists it makes sense that a data request comprises a proposal together with an authorization from the ethics committee (unless the request comes from a country without ethics committees). The proposal briefly describes the aims and objectives of the planned study or reanalysis of the requested data, the planned analysis, the data that are needed and the time frame of the study. The DAC/board of specialists evaluates and approves the request, checks the requesters identity and informs the principal investigator.

15. References

1. Naudet F, Sakarovitch C, Janiaud P, Cristea I, Fanelli D, Moher D, et al. Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: Survey of studies published in the BMJ and PLOS Medicine. BMJ. 2018;360.
2. Taichmann D. Data Sharing Statements for Clinical Trials: A Requirement of the ICMJE. Ann Intern Med [Internet]. 2017; Available from: http://www.icmje.org/news-and-editorials/data_sharing_june_2017.pdf
3. Ohmann C, Banzi R, Canham S, Battaglia S, Matei M, Ariyo C, et al. Sharing and reuse of individual participant data from clinical trials: Principles and recommendations. BMJ Open. 2017;
4. FORCE11. The FAIR Data Principles - FOR COMMENT [Internet]. 2016 [cited 2018 Dec 6]. Available from: https://www.force11.org/group/fairgroup/fairprinciples
5. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. Comment: The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;
6. GO FARE. FAIR Principles - GO FAIR [Internet]. 2018. Available from: https://www.go-fair.org/fair-principles/
7. The Federal Assembly of the Swiss Confederation. CC 235.1 Federal Act of 19 June 1992 on Data Protection (FADP). 1992; Available from: https://www.admin.ch/opc/en/classified-compilation/19920153/index.html
8. Confederation TFA of the S. Federal Act on Research involving Human Beings. Http://WwwAdminCh/Opc/En/Classified-Compilation/20061313/IndexHtml [Internet]. 2011; Available from: http://www.admin.ch/opc/en/classified-compilation/20061313/index.html
9. EDI ED des I. Koordinationsstelle Forschung am Menschen (kofam) [Internet]. Koordinationsstelle Forschung am Menschen (kofam) c/o Federal office of public health FOPH CH-3003 Bern. 2017. Available from: http://kofam.ch/en/applications-and-procedure/projects-that-do-not-require-authorisation/
10. Network SPH. Swiss Personalized Health Network [Internet]. Swiss Personalized Health Network. 2019. p. https://www.sphn.ch/en.html. Available from: https://www.sphn.ch/%0A
11. Swiss Personalised Health Network. Ethical Framework for Responsible Data Processing in Personalized Health Research [Internet]. 2018. Available from: https://www.sphn.ch/dam/jcr:6fb78ffa-95c8-4372-bfb1-5c9b1e2cb53d/Ethical_Framework_20180507_SPHN.pdf
12. Husedzinovic A, Ose D, Schickhardt C, Fröhling S, Winkler EC. Stakeholders' perspectives on biobank-based genomic research: Systematic review of the literature. Eur J Hum Genet. 2015;23(12):1607–14.
13. Trial Master File Reference Model [Internet]. https://tmfrefmodel.com/. 2014. Available from: http://tmfrefmodel.com/
14. Chevrier R, Foufi V, Gaudet-Blavignac C, Robert A, Lovis C. Use and understanding

of anonymization and de-identification in the biomedical literature: Scoping review. Vol. 21, Journal of Medical Internet Research. Journal of Medical Internet Research; 2019.

15. O'Neill L, Dexter F, Zhang N. The risks to patient privacy from publishing data from clinical anesthesia studies. Anesth Analg. 2016 Jun 1;122(6):2017–27.

16. Services C for M& M. The health insurance portability and accountability act of 1996 (HIPAA). Online http//www C hhs gov/hipaa [Internet]. 1996; Available from: https://www.federalregister.gov/documents/2009/10/30/E9-26203/hipaa-administrative-simplification-enforcement.

17. Statistik B für. MEDSTAT. https://www.bfs.admin.ch/bfs/de/home/statistiken/gesundheit/nomenklaturen/medsreg.html.

18. Garfinkel. De-identification of personal information. NISTIR 8035. 2015;

19. Atreya. Reducing patient re-identification risk for laboratory results within research datasets. J Am Med Inf Assoc. 2013;20(95).

20. Pike R. UTF-8 [Internet]. 2019. Available from: https://en.wikipedia.org/wiki/UTF-8

21. Oulun Yliopisto. Research Data Guide [Internet]. Managing research data: Data documentation and metadata. 2018. Available from: http://libguides.oulu.fi/Researchdata/Data_documentation

22. University ta trobe. Durable file formats. 2019; Available from: https://latrobe.libguides.com/dataorganisation/fileformats

23. Canham S, Ohmann C. A metadata schema for data objects in clinical research. Trials [Internet]. 2016 Dec 24;17(1):557. Available from: http://trialsjournal.biomedcentral.com/articles/10.1186/s13063-016-1686-5

24. DataCite. DataCite Metadata Schema for the Publication and Citation of Research Data. 2011;29.

25. WHO | WHO Data Set. WHO [Internet]. 2018 [cited 2018 Dec 6]; Available from: http://www.who.int/ictrp/network/trds/en/

26. ICMJE. Up-Dated ICMJE Recommendations [Internet]. 2019. Available from: http://www.icmje.org/icmje-recommendations.pdf

27. WHO | Glossary [Internet]. Who. Geneva, Switzerland: World Health Organization; 2018 [cited 2018 Dec 6]. Available from: https://www.who.int/ictrp/glossary/en/

28. CoreTrustSeal – Core Trustworthy Data Repositories [Internet]. [cited 2018 Dec 6]. Available from: https://www.coretrustseal.org/

29. Sciences swiss academies of arts and. Swiss Academies Factsheets. 2019. p. http://www.akademien-schweiz.ch/index/Publikatione.

30. Bill & Melinda Gates Foundation. Gates Open Research - Data Guidelines [Internet]. Seattle, WA: Bill & Melinda Gates Foundation; [cited 2018 Dec 6]. Available from: https://gatesopenresearch.org/for-authors/data-guidelines

31. European Commission. European Open Science Cloud (EOSC) | Open Science - Research and Innovation - European Commission [Internet]. [cited 2018 Dec 6]. Available from: https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

32. Swissuniversities. Olos [Internet]. 2018. Available from: https://www.dlcm.ch/olos

33. Whyte A. Where to keep research data: DCC checklist for evaluating data repositories (v.1.1) [Internet]. Edinburgh: Digital Curation Centre; 2015 [cited 2018 Dec 10]. Available from: www.dcc.ac.uk/resources/how-guides

34. Preservation D, Questions FA, Based L, Format O. Levels of Digital Preservation Support [Internet]. 2014 [cited 2018 Dec 10]. p. 1–5. Available from: https://ndsa.org/activities/levels-of-digital-preservation/

35. IWATSUBO T. アルツハイマー病根本治療薬の臨床応用に向けて : Alzheimer's Disease Neuroimaging Initiative (ADNI)の取り組み(最前線,アルツハイマー病). Vol. 43, Farumashia. 2007. p. 894–8.

36. Alter G, Gonzalez R. Responsible practices for data sharing. Am Psychol. 2018 Feb 1;73(2):146–56.

37. Kirkwood BR, Sterne JAC. Essential Medical Statistics. Medical statistics. 2003. 513 p.

16. Glossary

Term	Definition	Reference
Anonymization	Process by which any way of linking data in a data set with a natural person is irreversibly removed/destroyed or only possible with disproportionate effort. *De-identification* or *Pseudonymization* with destruction of the *Key* are needed as a minimum for this process. It must be noted that the required measures for anonymization must be defined on a case-by-case basis because a combination of not directly identifying information might enable identification of a natural person. The Human Research Act acknowledges that absolute irreversible anonymization is impossible. Disproportionate effort is given if linking: 1. Is only possible with considerable criminal energy, or 2. Requires extensive technical infrastructure and know-how.	Schweizerischer Bundesrat p. 8096. Eidgenössisches Department des Inneren p. 69-70.
Anonymized (health-related) data	(Health-related) Data which cannot (without disproportionate effort) be traced to a specific person. See also *Anonymization*	Human Research Act Art. 3i. and General Data Protection Regulation (EU)
Artifact	The term artifact is used because relevant study information might be recorded in a variety of different ways, including records, documents and data. An artifact is therefore any information that is captured during a clinical trial that meets the purpose or definition described in the protocol. In some cases, the artifact is a single document, data set or piece of information but in other cases it could be represented by multiple document types or data types.	https://tmfrefmodel.com/wp-content/uploads/2018/03/tmf-rm-deliverable-user-guide-v1-2018-03-16.pdf

Case Report Form	(1) A printed, optical or electronic document designed to record all of the protocol-required information to be reported to the sponsor for each subject/patient in a clinical trial. (2) A record of clinical study observations and other information that must be completed for each subject in a clinical trial, per study protocol mandate. CRF can refer to either a CRF page (which contains one or more data items linked together for collection and display) or a casebook (which includes all CRF pages on which a set of clinical study observations and other information can be or have been collected, or the information collected by completion of such CRF pages for a subject/patient in a clinical study).	The Free Dictionary
Coded data (set)	De-identified data that can be linked to a specific person via a *Key* (code). This means that the data are anonymized for any person who accesses the data and who has no direct access to the *Key*. However, the conditions under which the *Key* is stored and can be accessed are critical for qualifying data as coded: 1. Storage of the *Key* must be separate from the storage of the data. No person directly involved in a research project or who works as a subordinate to someone who wants to use coded data may have access to the *Key*. This includes but is not limited to investigators, study nurses/coordinators, statisticians, and data managers. Precautions must be taken to ensure that only authorized persons have access to the *Key* (see 2) and each access must be documented (date and who accessed it for what reason).	Human Research Act Art. 3h. HRO Art. 26-27.

	2. Decoding i.e. identifying a person is only allowed under the following conditions: a. Breaking the code is necessary to avert an immediate risk to the health of the person concerned. b. A legal basis exists for breaking the code. Breaking the code is necessary to guarantee the rights of the person concerned, and in particular the right to revoke consent.	
Controlled access	Refers to the way a data set is shared. In a controlled access model, the data are only shared with an entity if they meet certain conditions and on request.	Keerie C et al. 2018.
Data	Pieces of information. Within this document, we use a narrow definition of data, denoting the content of structured data files.	
Data Access Committee	A Data Access Committee (DAC) is a body of one or more individuals who are responsible for data release to external requestors based on consent and/or National Research Ethics terms. A DAC is typically formed from the same organization that collected the samples and generated any associated analyses. Multiple datasets may be affiliated to a single DAC.	European Genome-phenome Archive
Data Management Plan	Document that outlines how data are to be handled both during and after a research project including data preservation.	Wikipedia
Data Object	An entity available in electronic format (document, text, program, zip file). In the setting of clinical research data sharing: data and associated documents related to a clinical trial and typically stored in a repository.	Canham and Ohmann. Trials (2016) 17:557

Data Sharing/Transfer Agreement	Contract or license that describes the conditions	
Data Validation Plan	Document that describes the process of data validation, e.g. which variables have to be checked and what consistency rules have to be met. It might include checks on chronological sequence, completeness, identification of duplicates, checks of range and distribution shape of variables.	
De-identified	See *De-identification*	
De-identification	Process by which all directly identifying data is either removed, altered or censored from a data set. It must be noted that the term *de-identification* as such has no legal basis in Switzerland but rather is a concept originating in the USA based on rules set forth in the Health Insurance Portability and Accountability Act (HIPAA). For the purpose of this document, de-identification relates only to directly identifying data. De-identification is usually insufficient for data sharing.	US Office for Civil Rights 2012
External party	Receiver of de-identified data whose access to the data was not explicitly consented to by the patients (could be a researcher or data repository). Alternative phrase: third party.	
Identifier	A number or string that identifies/labels a unique object. *Identifiers* in a clinical study project usually follow an encoding system; in other words, there are rules behind the generation of the *identifier*. Such rules might be a pseudonymization algorithm (see *pseudonym*) or a sequential numbering system. Identifiers are therefore often referred to as *ID code*, *ID number*, *record ID*, or *unique identifier* (UID) in the clinical research context.	Wikipedia

Identifying data (directly or indirectly)	Any information that solely (directly) or jointly with other data enables identification of a natural person among a data set.	
Key	A piece of information that allows decrypting encrypted data. In the clinical research context this is usually a participant/patient log/list that allows linking a (unique) *identifier* (record) with the *identifying data* usually the full name, birth date, and hospital/practice identification number. The *key* is usually stored on site under restricted access (e.g. in the study-binder).	Wikipedia
Limited data (set)	A data set that has been de-identified and which contains only the absolute minimum number of *variables* required to conduct an analysis by an *External party*. It includes *variables* needed to derive variables which are needed to conduct the analyses by the *External party* unless these *variables* increase the risk for identification.	
Metadata	Data about data; a vector of structured information, typically numbers or classification options that describes a fixed set of aspects of a data object in a human and machine-readable way.	
Open access	Refers to the way a data set is shared. In an open access model, the data is shared publicly and can be accessed without restriction or request.	Keerie C et al. 2018.
Personal Data (health-related)	Any information relating to an identified/specific or identifiable natural person (*data subject*); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity.	Cantonal Data Protection Act (Kantonales Datenschutzgesetz, KDSG) Art.2 Par. 1 (Federal Act on Data Protection Art. 3a).

		EU Directive 95/46/EC 4.
Pseudonym	A pseudonym or alias is a unique name (or more generally, a string consisting of alphabetic and potentially numeric characters) used to conceal *identifying data*. The pseudonym is generated using a set of rules (pseudonymization algorithm). A pseudonym can be generated with or without the possibility of restoring the underlying *identifying data* (reversible or irreversible pseudonymization). If the same algorithm is used across systems, pseudonymization allows for data to be linked to the same person across multiple data records or information systems without revealing the identity of the person. It must be noted that the term does not appear in any of the following laws: HRA, ClinO, HRO. Derivation of a new variable from other variable(s) using simple rules like calculating age from date of birth and enrolment date is not considered pseudonymization as this does not generate a unique attribute.	
Pseudonymization (reversible or irreversible)	See *Pseudonym*	
Registry	A clinical trial registry is an entity that houses clinical trial registers i.e. a record containing information about a clinical trial (27). In registries accepted by the World Health Organization (WHO) and included in their International Clinical Trials Registry Platform (ICTRP), these records contain a minimum amount of information as defined in the WHO Data Set (25). As of 2019, this data set does not define or require attached artifacts or files. Confusingly, the WHO calls the database behind its Search Portal "Central	

	Repository" (27), when it is in fact a registry.	
Repository	Collection of digital datasets. Technically, it consists at least of a backend, a database to store metadata and information; a file server to store the datasets and other digital artifacts; and a web-based frontend that allows users to access the backend. Although not mandatory, the term implies that there is a function to make these datasets findable, accessible, and reusable (5) and allows for longer term storage.	
Statistical Analysis Plan	A statistical analysis plan is a document that contains a more technical and detailed elaboration of the principal features of the analysis described in the protocol, and includes detailed procedures for executing the statistical analysis of the primary and secondary variables and other data.	
Third party	See *External party*	
Variable	A measured or recorded attribute that characterizes an object, e.g. a participant. A variable is the operationalized way in which the attribute is represented for data processing i.e. a variable contains attributes. There are different types of variables (data types). The most common ones are: nominal/categorical with the special case of binary (only two categories), ordinal, numeric/continuous, date & time, string.	(37)

17. Appendix

17.1. Further detailed specification of required data processing steps

17.1.1 Example data to be considered for deletion

- Names, address, etc. have to be deleted, see section 8.2.
- All freetext variables should be deleted unless the content is checked and redacted where necessary to ensure privacy.
- Any internal record identifier of the clinical database.
- Any identification numbers that are not needed for analysis purposes such as biosample/kit numbers etc.
- Any variables that contain data that is particular or has low prevalence e.g. multiples (twins, …), special comorbidities.

17.1.2 Examples and details on manipulations to decrease precision

- Dates (time): The enrolment date (time) should be set to zero. All other date variables (including date of birth) should be replaced by variables containing time relative to the enrolment date using the appropriate scale e.g. years for age, or days for study visits (relative study day). Consider to deliver age bands (e.g. 5 year bands) instead if the disease or population is infrequent or rare. To protect persons in rare age groups, those above 89 should be grouped together in a “90 or older” category. Use accordingly for young ages as appropriate.
  Alternatively, a random offset can be added to all dates in the data for a specific person. It is recommended to use different offsets for each person, as long as relative differences between persons are not relevant. For some dates, e.g. birthdays, or when seasonal effects are of interest, other methods such as the generalization into certain categories like month or years, may be required.
- Geographic information: Consider whether aggregation to MEDSTAT(17) or other higher level unit is appropriate.
- Unusual data: If a variable contains data that allows identification of individuals because it is special or has low prevalence consider grouping or aggregation into categories.
- Height and body weight: Consider whether Body Mass Index (BMI) is sufficient and derive BMI and delete height and body weight.
- Renal function: Consider whether Serum Creatinine can be replaced by estimated Glomerular Filtration Rate (GFR).
- (High precision) continuous/numerical data: Round data to the next higher digit or introduce random jitter on the last digit (Perturbation).

- Identification numbers that are needed for analysis, participant ID, study site ID (cluster ID, country ID, etc.):
- All identification numbers must be replaced by a unique random number. It is important to ensure that records with the same identification number, e.g. participant or study site identifier, are assigned the same new random number. The general process is:
    1. Check all data files for the variable (identification number) of interest.
    2. Collect the maximum amount of data i.e. make sure that you get all identification numbers of interest across all data files and save in a separate data file.
    3. Randomly shuffle the IDs (1. generate a new variable with random numbers (no seed[4]), 2. sort data accordingly, and 3. replace the new variable with integers in ascending order (new ID). Make sure that the new variable contains only unique numbers).
    4. Merge the new ID into all relevant data files.
    5. Delete the original ID from all relevant data files.
    6. Repeat for other identification numbers.

    If the number of records is unique for a particular identification number e.g. study site ID, consider to aggregate.

General approaches:

- Aggregation (generalization) might be a strategy to achieve de-identification and should be considered if other manipulations remain unsatisfactory. For example, numerical data can be transformed into categorical variables and categorical variables may be combined into new (less informative) categories. As outliers have a larger risk of re-identification, one could aggregate outliers only and leave non-outlier values unchanged.
- Replacing the observed value of specific record with "missing", thereby increasing the frequency of certain rare combination (suppression).
- Data swapping: For a fraction of records, values of quasi-identifiers might be exchanged, with the possibility of adding constraints on which pairs of records can be swapped. For example, given two "similar" records, one may swap the values of one quasi-identifier, e.g. age.
- Resampling: One identifies the probability distribution of the quasi-identifying data and replaces its values with a random sample from its distribution. Care must be taken if correlations with other variables need to be preserved.
- Subsampling: Only a subsample of the data might be shared, thereby reducing the risk of re-identification.

[4] Alternatively, a random seed might be used, but removed from any documentation after the final dataset was created and underwent the anonymization process.

17.2. Further details on coding of variables

17.2.1 Formatting of date and time variables

- Date variables should be provided in the ISO 8601 standard of year-month-day (e.g. 12th October 2018 would be 2018-10-12).
- Time with seconds should be coded as hours:minutes:seconds (e.g. 07:59:45 or 15:32:01). Where seconds are unavailable, leaving away seconds is considered acceptable (e.g. 15:32), so long as all observations are coded consistently (same applies to minutes).
  - Where data come from multiple time zones, the offset from Coordinated Universal Time (UTC) should be added (e.g. 15:32+01:00 for Central European Time). Conversion to UTC is encouraged.
- Date-time variables should follow the rules for both date and time, and have the date part followed by the time part, separated by a space (e.g. 2018-10-12 07:59:45 or 2018-10-12 07:59; the strict ISO 8601 standard separates dates and times by T, but the space is readily recognized as a date-time variable by statistical software).
  - As with times, the offset from UTC is vital for datasets including multiple time zones.

17.2.2 Examples for further documentation of the dataset

Table 1: Codebook example

Labelname	Code	Value label
yn	0	No
yn	1	Yes
sex	1	Male
sex	2	Female
route	1	Oral
route	2	IV
route	3	Anal
unit	1	mg/dL
unit	2	mg
unit	3	ug/dL
unit	4	ug
unit	5	g
freq	0	less frequent
freq	1	daily
freq	2	twice daily
freq	3	every 8 hours
freq	4	every 6 hours
freq	5	more frequent

Table 2: Recommended data type names. These types would be referenced in the labelbook

Data type	Description
Str	Free text (short for string). See above for notes
Int	Integer
Num	Numbers without specific accuracy
Num_Xdp	Number with X decimal places (e.g. num_1dp for values with 1 decimal place)
Date	Date variables (formatted to ISO 8601 standards)*
Time	Time variables (formatted to ISO 8601 standards)*
Datetime	Date and time variable (formatted to ISO 8601 standards)*
Cat	Categorical variable (e.g. male/female/undifferentiated/unknown)
Bin	Binary variables (e.g. yes/no)

* would ideally be converted to study time (e.g. days since randomization/informed consent/some other reference point); see section 6.

Table 3: Labelbook example

Form	Variable	Label	Type	Unit	Label name	Note
	visit	Visit ID	Int			
	pid	Participant ID	Int			
	position	Position in repeating form sequence	Int			
elig	sex	Sex	Cat		sex	
elig	age	Age	Int	Years		
elig	ic	Informed Consent given	Cat		yn	
elig	ic1	Age 18 years or older	Cat		yn	
elig	ic2	Recurrent kidney stone disease	Cat		yn	
elig	ex1	More than 5 instances of kidney stone disease	Cat		yn	
elig	ic_date	Date of Informed Consent	Date			
lab	lab_bl_yn	Blood sample taken	Cat		yn	
lab	lab_bl_rbc	Red blood cell count	num_1 dp	mcL		
lab	lab_bl_ldl	Blood LDL cholesterol	Int	mg/dl		
drug	uvisit	Unscheduled visit ID	Int			
drug	position	Drug name	Str			
drug	route	Administration route	Cat		route	
drug	Dose	Dose	Num			see unit for relevant units
drug	Unit	Unit	Cat		unit	
drug	Freq	Frequency	Cat		freq	
drug	freq_det	Frequency details	Str			if freq = 0 or 5
drug	Start	Start	Date			
drug	ongoing	Ongoing?	Cat		yn	
drug	End	End	Date			

Table 4: Structure of dataset with one row per participant (part of eligibity form)

Visit*	pid	sex	age	ic1	ic2	ex1	ic	ic_date
1	1	1	58	1	1	0	1	2016-01-09
1	2	2	54	1	1	0	1	2016-01-15
1	3	1	54	1	1	0	1	2016-07-11
1	4	1	41	1	1	0	1	2016-09-01
1	5	1	32	1	1	0	1	2017-09-11
1	6	2	36	1	1	0	1	2017-09-28
1	7	2	30	1	1	0	1	2017-10-24
1	8	2	51	1	1	0	1	2018-10-27

* The visit variable in this case is optional as the eligibility form is only used once.

Table 5: Structure of dataset with multiple rows per participant (part of blood laboratory values form)

visit	pid	lab_bl_yn	lab_date	lab_bl_rbc	lab_bl_chol
1	1	1	2016-01-09	5.1	123
1	2	1	2016-01-15	5.6	144
1	3	1	2016-07-11	4.7	103
1	4	0			
1	5	0			
1	6	0			
1	7	1	2017-10-24	5.2	110
1	8	1	2018-10-27	4.2	90
2	1	0			
2	2	0			
2	3	1	2016-08-05	4.8	66
2	4	1	2016-10-02	4.5	142
2	5	1	2017-10-12	4.7	103
2	7	0			
2	8	1	2018-11-25	6.1	125
3	1	1	2016-03-10	5.5	140
3	2	1	2016-03-20	5.4	130
3	3	0			
3	4	1	2016-11-06	6	129
3	5	0			
3	7	1	2017-12-20	5.2	111
3	8	1	2018-12-28	4.5	121

Table 6: Example for a log form data table

pid	position	drug	route	dose	unit	freq	freq_det	start	ongoing	end
1	0	amoxicillin	1	500	2	2		2016-02-25	0	2016-03-05
1	1	amoxicillin	1	500	2	2		2018-10-20	0	2018-11-01
2	0									
3	0									
4	0	morphine	1	15	2	5	every 4 hours	2017-01-20	0	2017-01-25
4	1	morphine	1	30	2	5	every 4 hours	2017-01-26	0	2017-01-30
4	2	morphine	1	5	2	1		2017-01-31	0	2017-02-05
5	0									
6	0									
7	0									
8	0									

We see that participant 1 reported taking a medication at two time points, while participant 4 reported taking morphine for a period of time, including changing doses. The remaining participants took no medications.

Table 7: Visitbook example (first three visits only)

visit	visitlabel	form	formname
1	Baseline visit	elig	Eligibility
1	Baseline visit	lab	Laboratory values
2	1 month	visit	Visit info
2	1 month	lab	Laboratory values
3	2 month	visit	Visit info
3	2 month	lab	Laboratory values

Table 8: Visit structure

Form	Baseline	1 month	2 month
Eligibility	X		
Laboratory values	X	X	x
Visit info		X	x

17.3. Meta data scheme from ISRCTN

Options are added in curled brackets if provided, an empty filed on the right hand side indicates free text, "M" denotes mandatory fields.

General data

Public title	M	
Overall trial status		
Recruitment status		
Plain English Summary	M	Who can participate? What does the study involve? Where is the study run from? When is the study starting and how long is it expected to run for? Who is funding the study? Who is the main contact? Trial website

Contact information

Type	M	{Public, Scientific}
Primary contact	M	
ORCID ID		
Contact details	M	
Additional contact		
Type		{Public, Scientific}
ORCID ID		
Contact details		

Additional identifiers

EudraCT number		
ClinicalTrials.gov number		
Protocol/serial number	M	

Study information

Scientific title	M	
Acronym		
Study hypothesis	M	
Ethics approval	M	
Study design	M	Free text
Primary study design	M	{Not Specified, Interventional, Observational, Other}
Trial setting		{Not Specified, Hospitals, GP practices, Other, Home, Internet, Community, Schools}
Trial type	M	{Not specified, Diagnostic, Other, Prevention, Quality of life, Screening, Treatment}
Patient information sheet		

Condition	M	Free text
Intervention	M	Free text
Intervention type	M	{Not specified, Drug, Supplement, Device, Biological/Vaccine, Procedure/Surgery, Behavioral, Genetic, Other, Mixed}
Phase		
Drug names		
Primary outcome measure	M	
Secondary outcome measures	M	
Overall trial start date	M	
Overall trial end date	M	
Reason abandoned (if study stopped)		

Eligibility

Participant inclusion criteria		
Participant type	M	{Not Specified, Healthy volunteer, Patient, Health professional, Carer, All, Mixed, Other}
Age group	M	{Not Specified, Adult, Senior, Neonate, Child, All, Mixed, Other}
Gender	M	{female, male, both}
Target number of participants	M	
Participant exclusion criteria	M	
Recruitment start date	M	
Recruitment end date	M	

Locations

Countries of recruitment		
Trial participating centre	M	

Sponsor information

Organisation	M	
Sponsor details	M	
Sponsor type	M	{Not defined, Charity, Government, Hospital/treatment centre, Industry, Other, Research council, Research organisation, University/education}
Website		
Privacy	M	{Show all contact details, Hide telephone and email details}

Funders

Funder type	M	
Funder name	M	
Alternative name(s)		
Funding Body Type		
Funding Body Subtype		
Location		

Results and Publications

Publication and dissemination plan		
Intention to publish date		
Participant level data	M	{Available on request, Not expected to be available, Stored in repository, Other, Not provided at time of registration, To be made available at a later date}
Basic results (scientific)		
Publication list		
Publication citations		

17.4. Information required for additional documentation

Table 9: Information on supplied documentation

Title	Size	Type	Format
Study_Protocol_final	420 KB	Text	Pdf
Data_preparation	67 KB	Stata script	Do
Statistical_analysis_plan	180 KB	Text	Pdf
Consent_form	56 KB	Text	Word
analysis_final	120 KB	Stata script	Do

17.5 Checklist for selecting a data repository

Table 10: Selection criteria

Item	Yes	No	Unsure	Potential indicators	Explanation
Is the repository trustworthy?	□	□	□	• Certifications or public institution behind the repository?	
Will my data, information, and documentation be hosted?	□	□	□	• Any restrictions on file type? • Any restrictions on file size?	
Will any legal requirements be met?	□	□	□	• Licensing • Storage of sensitive data	
Does the repository support the sharing process?	□	□	□	• On request …	See chapter 11
FAIR data principles	□	□	□	Does the repository make the data findable, accessible, interoperable, and as reusable as possible for as long as required?	In order to sustain the value of the data, the repository has to comply with the FAIR principles.
Basic functionality	□	□	□	• Single landing page per dataset • Unique identification number • Digital Object Identifier	
Does the repository allow for enough and the right meta-information? Is the metadata scheme specific for medical research?	□	□	□	• Specific metadata fields on disease, intervention, outcome etc.	See chapter 8
Long term preservation, sustainability	□	□	□	• (might not be possible to assess)	Is there any plan in how long term preservation is ensured? For how long is storage guaranteed (for example, the repository of the Open Science Framework has a preservation fund that ensures hosting for 50+ years (based on present costs)).
Does the repository track usage and provide sufficient statistics?	□	□	□	• Page views for each object/dataset • Number of downloads per object/dataset	