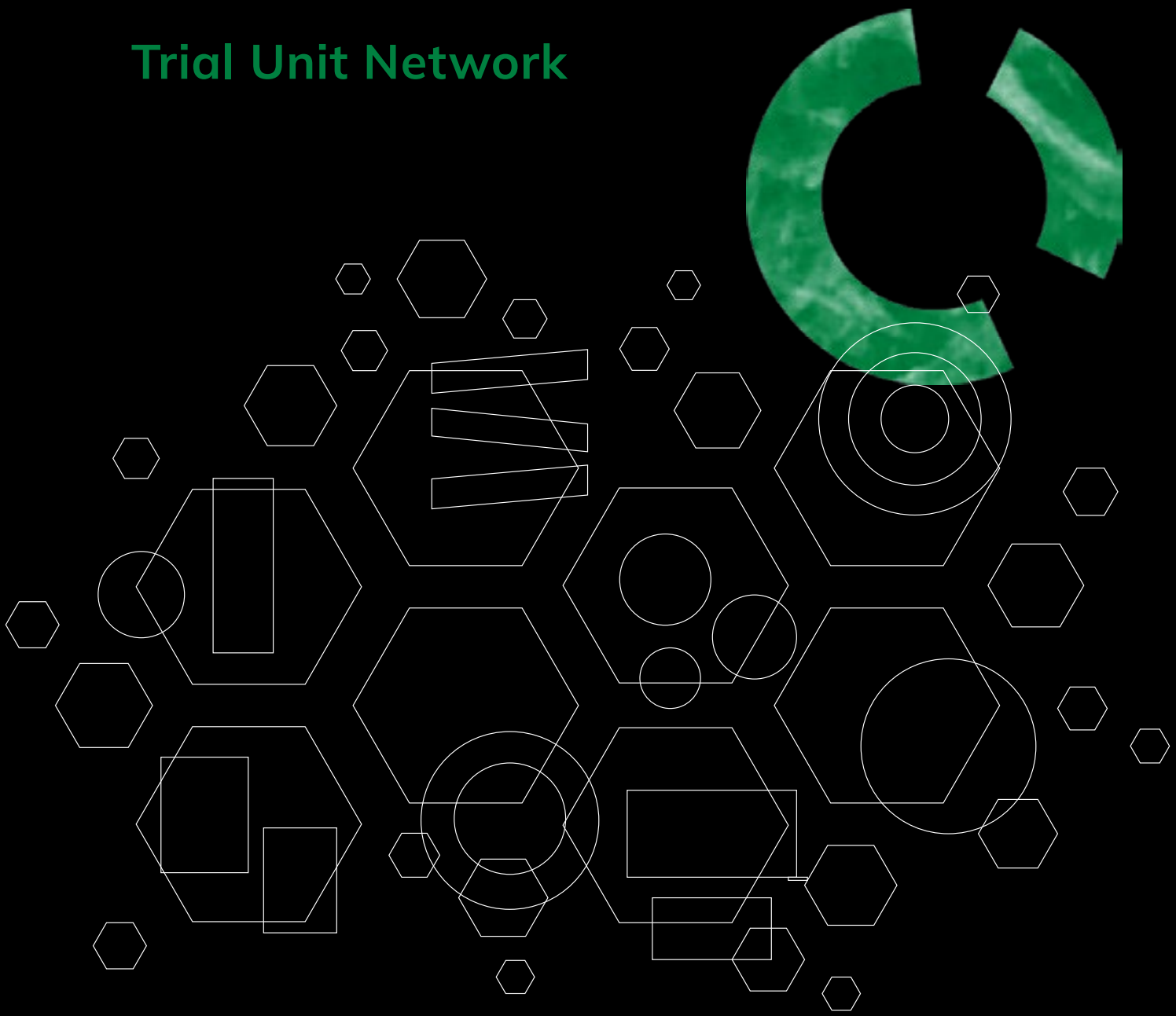
SCTO PLATFORMS

Sharing of Data from Clinical Research Projects

Guidance from the SCTO's Clinical Trial Unit Network



Publisher
Swiss Clinical Trial Organisation (SCTO)
Bern, Switzerland

Authors
Gahl, Brigitta, PhD, Clinical Trials Unit Bern, Statistics & Methodology Platform of the SCTO
Haynes, Alan G, PhD, Clinical Trials Unit Bern, Statistics & Methodology Platform of the SCTO
Sluka, Constantin, PhD, Department of Clinical Research, University Hospital Basel, Data Management Platform of the SCTO
Dupuis-Lozeron, Elise, PhD, Clinical Research Centre, Department of Health and Community Medicine, Geneva University Hospitals
Jörger, Francisca, PhD, Clinical Trials Center, University Hospital Zurich
Schur, Renate, MSc, Clinical Trials Center, University Hospital Zurich
Christen, Andri, PhD, SCTO
Trelle, Sven, MD, Clinical Trials Unit Bern, University of Bern, Statistics & Methodology Platform of the SCTO

Recommended form of citation
Gahl, Brigitta; Haynes, Alan G; Sluka, Constantin; et al. (2021). Sharing of Data from Clinical Research Projects Guidance from the SCTO's CTU Network. Published by the Swiss Clinical Trial Organisation (SCTO). doi: 10.54920/SCTO.2021.02

Contact
To give feedback on this publication or obtain further information, you can contact platforms@scto.ch.
For more information on SCTO Platforms, please visit: www.sctoplatforms.ch.

Table of Contents

1. Abstract

Objectives: Data sharing has become a requirement of many funding bodies and is becoming a scientific standard in many disciplines. In clinical research, however, data sharing can conflict with the obligation to protect privacy of study participants and especially of patients. General recommendations on data sharing exist also for clinical research, but so far they 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 SCTO's 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 was published as pre-print for external review. This is the second version with feedback from these external reviews incorporated.
We 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 de-identification techniques or changes to the legal situation.

Results: We developed principles and practical recommendations with respect to informed consent, data management plan, de-identification, 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 intention 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, metadata, and any further documentation for being shared. This report provides 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

4. Introduction

4.1 Why share data?

Sharing of research data[Glossary][1] has become standard practice in many disciplines. The two main objectives for data sharing are 1) enabling reproducibility checks of research results and 2) reuse of data for new research questions. In clinical research, data usually relates to individuals, mostly patients. Hence, data sharing may conflict with the duty to maintain patient privacy. There are, however, good reasons to advocate data sharing particularly in health research. The International Committee of Medical Journal Editors (ICMJE) considers it an ethical obligation to responsibly share data generated by clinical trials (1). The main reason for this is that trial participants have put themselves at risk by accepting to receive a treatment under study and to take part in an experiment. These considerations caused funding bodies such as the Swiss National Science Foundation, to request data sharing.

Many people have doubts about the validity and reliability of research results. The proportion of false published findings is estimated as high as 85% (2). In 2014, The Lancet published the article series "Increasing Value and Reducing Waste" that describes underlying problems and provides possible solutions (3–5). Even though the first version of the CONSORT statement was published more than 20 years ago, quality of reporting remains suboptimal (6,7). But even if adherence to reporting guidelines would be perfect, essential information often remains unclear for certain aspects (5). Independent verification of results for both, reproducibility and replicability (see Infobox 1), is therefore impossible for most studies. Examples have shown, however, that this can be important (5).

Collecting high-quality data in health research is costly and time intensive. Further use of data, i.e. using the data beyond the objectives it was originally collected for, is therefore imperative for efficient use of limited resources. This secondary use of data might be motivated by different purposes. The main use may be considered the testing of additional hypotheses or in individual participant data meta-analysis. An-Wen Chan and colleagues provide examples where the testing of secondary hypotheses led to improvement in the evidence-base of commonly used interventions (5). The utility of individual participant data for meta-analysis is indisputable especially with regard to identifying treatment effect modifiers (sometimes called treatment effect heterogeneity) but also for precision of estimates (8,9). Moreover, there are multiple examples that show their practical value in terms of advancing the evidence-base for clinical practice (10). There might be additional uses for individual participant data such as development of prognostic models or for investigating new statistical methods.

[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.

Infobox 1: "reproducibility crisis" and "replication crisis" and its terminology

Discussions on repeating research studies, reproducibility, and replication have gained popularity in science over the last decade. A survey by *Nature* revealed, that more than 50% of the responding researchers think that science is facing a "replication crisis" (11). Several dimensions relate to this phenomenon and the reader is referred to the relevant literature e.g. (12). For this statement, clarification of terminology is warranted as data sharing is mainly (although not only) relevant for reproducibility in this context. Different definitions are in use (12,13) but the terminology by Kirstie Whitaker appears useful here (14).

Reproducible: relates to an independent check of an analysis using the same analysis approach with the same data as in the original study. Reproducibility and reproduction is therefore an aspect of quality assurance/control (15) rather than a specific aspect of science or for its advancement (16).

Robust: relates to using different analyses approaches for the same data to investigate sensitivity of results to underlying (untestable) assumptions etc. Experimental robustness might be defined as varying the experimental set-up (13) although boundaries to *Replication* are fluid.

Replicable: relates to (independently) redoing a research study/scientific experiment. Replicability can relate to the results itself (quantitative replicability i.e. the replication study has similar results) or the conclusions derived from the results (qualitative or inferential replicability i.e. the interpretation and conclusions derived from the results of the replication study match those of the original study). Whether quantitative replicability is possible at all is beyond this statement (13).

Generalizable (and transportable): in clinical trials, generalizability relates to whether results/conclusions can be applied to different populations than the study sample, usually individuals seen in clinical practice. This "extending inferences from a trial to a target population" might be further differentiated into transportability and generalizability (17).

4.2 Purpose of this document

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 or aggregate data. The recommendations take applicable Swiss legislation into account, namely the Human Research Act (18), Human Research Ordinance (19), and the Federal Act on Data Protection (20). However, most recommendations apply also to other legislative contexts. A sponsor, principal investigator (PI) or other study team member who intends to share data should be able to find answers to her/his questions in this document and whatever else there is to consider. We describe options to minimally fulfil current requirements in case resources or motivation are small to still allow for the culture of data sharing to evolve.

The two objectives of data sharing, reproducibility[2] and reuse, may result in different and sometimes divergent requirements. They might therefore require different means with respect to the amount of data and documentation to be shared. It is up to the study team to set priorities where necessary.

[2] *Reproducibility* in the context of shared data does not mean that researchers redoing the same analyses will end up with exactly the same result for each estimate that is intended to be reproduced. If precise data such as biomarkers have been jittered or grouped as a means of de-identification, estimates from a reproducibility study might differ from the original estimates. It is important that this difference is mentioned and quantified in the documentation (see section 10.3.4).

5. Origin of the document

5.1 Genesis of this document

This document was written by Swiss professionals in the field of academic clinical research. Involved persons were identified within the SCTO's CTU Network and delegated from each clinical trials unit participating in the network. The authors identified relevant topics to be covered, not following a structured Delphi process, 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. Afterwards this document underwent language review and was published on https://www.preprints.org/ for review (21). We invited national and international experts from different organizations and institutions such as European Clinical Research Infrastructure Network, Swiss Personalized Health Network, universities, clinical trial units, Swiss National Science Foundation, university libraries, university hospitals, ethical committees. We received comments from over 20 individuals which are incorporated in this version.

5.2 Work in progress

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 repositories or methodological advances in de-identification techniques or changes to the legal situation. A formal review cycle of two years is currently foreseen.

5.3 Related recommendations

Recommendations on sharing individual participant data were previously published. Ohmann et al. developed guidance for sharing individual participant data using a consensus-building process among an interdisciplinary task force of research professionals as part of a European project (22). The paper 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" (22). Our own

statement is rather dedicated to the reality faced by clinical researchers when thinking about data sharing.

5.4 FAIR data sharing

The FAIR principles (23–25) provide guidelines to improve the Findability, Accessibility, Interoperability, and Reusability of data. They were developed for scientific data in general and focus on machine-operability. The order of the letters represent the dependency of the principles, e.g. data must be findable to be accessible, and can only be reusable if they are findable, accessible and interoperable. 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]. The choice of repository already determines many aspects of findability and accessibility. Usually, a repository has a metadata[Glossary] scheme, see sections 13 and 15, 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 13. 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 11 and 12. When retrieving data, a whole package of related descriptions and documentation is needed to understand the data and allow its meaningful reuse i.e. make it reusable, see section 13.

6. Some questions related to non-technical aspects of data sharing

6.1 Clinical trials or observational studies?

We think that the recommendations we give apply to data from either clinical trials or observational studies. Clinical trials always involve study-related procedures and are typically funded by research grants. Prospective observational studies are often organized and conducted like clinical trials i.e., they have dedicated funding, a detailed study protocol, study-related procedures etc. Reality tells us that retrospective clinically oriented observational studies are often done without specific funding, based on routinely collected clinical data with relatively slim protocols.
Data preparation for sharing will most likely be similar especially for clinical trials and prospective observational studies. Actually, some large and prominent prospective observational studies are planned with data sharing in mind or have sharing aspects at their core e.g., the Framingham Heart Study. However, for retrospective observational studies, study protocols and further documentation might be less standardized. Providing sufficient documentation might therefore be more difficult and time consuming as compared to clinical trials, see section 13. In addition, the limited resources available for many observational studies might hinder appropriate data sharing.

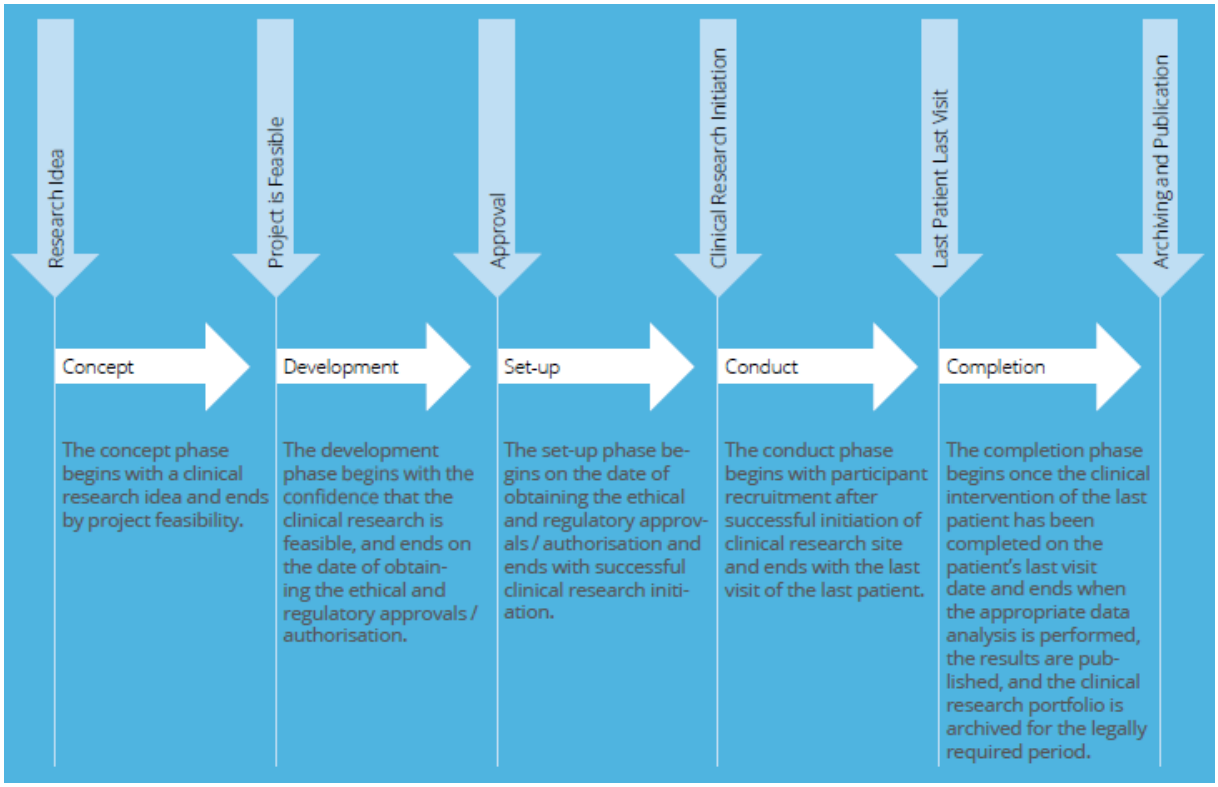
6.2 Who should be responsible?

We believe that data sharing should be independent of individual persons but rather be an institutional obligation i.e. the responsibility of the institution where the main researcher is working at the time of study conduct. The main reason is that fluctuation and mobility of individual researchers is high nowadays whereas institutions, especially universities and university hospitals, are relatively stable and therefore easier to reach. This concept has practical implications: the preparation of data and all documentation as well as the upload to a repository has to be completed shortly after the study is finalized. We are aware that at the time being, it might appear more sensible to leave work to be done *on-demand*, e.g. to publish metadata in a repository and only prepare data and documents in case someone requests it. Use of shared data is not common practice today and it is not clear when it will be, so preparation work carries the risk of being wasted in case no one ever asks for the data. We think that, on the other hand, preparation of data and documentation will in general not be feasible at a later time point. Any detail which is not properly

documented on-time in a systematic way and stored at a proper place is highly susceptible to loss because people involved with the study might no longer be available or might not remember. Institutional data sharing also implies that there is a standard of what data and documents to be shared, so decisions are not left to the discretion of an individual researcher. Increasing experience with data sharing will reduce the time needed for compilation of data and documentation packages.

6.3 How should data sharing be implemented in the course of a study?

Data sharing should be considered already in the development phase of a study. It might otherwise cause unreasonable workload e.g. if data sharing is not planned properly but only considered after the study is finished. It might even be precluded completely e.g. if participants' consent does not allow for it. We use the visualization of the Guidelines for Good Operational Practice of the Swiss Clinical Trial Organisation (26) to link the sections of this document to the study phases.



The **Legal Basis** (Section 7) obviously is relevant in each phase of a research project. **Requesting and use of data** (Section 16) might be worth a consideration early on, e.g. during Concept or Development, to check whether there are already data that could help to answer the study question. During the Development phase, many study documents have to be written and the data base is defined, and the intention to share data has consequences on a couple of specifications. The **Informed Consent** (Section 8) should be stated in a way to enable data sharing. For definition of the

data base, **Data Structure and Format** (Section 11) and **Coding of variables** (Section 12) should be considered. At the latest, the **Data Management Plan** (Section 9) should be written when the case report form is defined, but usually continues to evolve through the Set-up and even the Conduct phase. **Selection of repository** (Section 15) is ideally done when writing the data management plan, but dedicated repositories for a specific field might become more and more available. So, the landscape of available and more appropriate repositories might evolve during study Conduct. Central documents of the study such as the protocol and the statistical analysis plan should be kept up to date if anything changes, so almost no further work is needed to share **documentation** (Section 13). **De-identification** (Section 10) of the data has to be undertaken once the analysis is done during Completion. **Metadata** (Section 13) follow from selection of repository and are defined shortly after Completion, when **documentation** specific for data sharing is finalized. The process of **Requesting and use of data** (Section 16) is also defined by the repository.

6.4 What are the costs of data sharing?

Meaningful data sharing requires work and resources. If it is well-planned and all documentation is kept up-to-date during study conduct, only the data files need to be prepared. But this might also be a time consuming task. We think it is important to mention this here explicitly because our experience shows that efforts are underestimated. We also put the term “meaningful” at the beginning of this section to emphasize that data sharing is useful if done properly and with care which often means with adequate resources. In turn, we believe that if data sharing is done with insufficient resources it might do more harm than good e.g. it can result in increased risk of privacy breaches or wrong secondary analysis because of misunderstandings. The standard we describe here may appear high, even exaggerated. We are aware that our recommendations need a lot of work to be fulfilled, which might appear rather discouraging instead of easing data sharing. However, we think that there is a high risk of sharing useless data if preparation is not done with the necessary care or documentation is outdated. So, we think that considerable effort is indispensable. It should be noted that many funding bodies nowadays financially support efforts for data sharing.

6.5 Co-authorship in case of re-used data?

Criteria for (co-)authorship e.g. from the International Council of Medical Journal Editors comprise contribution to a manuscript and also accountability for its content (27). Therefore, co-authorship does not follow automatically from simply providing data. Additional intellectual contributions to the study that (re-)uses the shared data

are required to justify (co-)authorship. Contribution of data has to be acknowledged in any publication via appropriately citing it, see (28).

7. Legal basis in Switzerland

Health-related personal data are considered sensitive data in Switzerland. According to Article 4 paragraph 3 of the Federal Act on Data Protection (FADP) (20), 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) (29). The Act regulates biomedical research with persons (and data/biosamples) at the federal level and is based on internationally recognized principles. Sharing health-related data fulfills criteria for Further Use[3] (30) 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, 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 or prospective observational study (but also for routinely collected data, see discussion on *general consent* in chapter 8), 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, which applies to clinical trials as well as projects falling under chapter 2 of the HRO, states: "If the intention exists to make further use for research of … [the] 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 such a study; individuals must be given the possibility to participate without giving consent for data sharing later. If no consent for further use was sought at the time of initial data collection, data sharing requires explicit written consent (Art. 28 and 31 HRO). If the intention is to share only coded data, and the data do not contain genetic information, information about potential further use is sufficient unless a participant explicitly disagrees. Explicit consent for data sharing is not required in such cases (Art. 33 HRO). 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). In general, further

[3] 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.

use of data and therefore data sharing with the purpose to answer a research question requires approval by the responsible ethics committee (Art. 33-40 HRO).

If personal data is shared abroad, adequate data protection for the transfer process and storage at the receiver side must be ensured (Art. 6 Federal Act of Data Protection). Adequate data protection should be part of any data use agreement (see section 16).

Anonymous data, which are not personal and cannot harm persons by definition, are subject to neither FADP nor HRA, and may be freely shared. However, as described below it is typically not possible to ensure that individual patient data are or will remain anonymous (see section 10).

Infobox 2: Swiss legal basis in a nutshell

1. Data sharing is considered further use.
2. Consent for data sharing should preferably be obtained at enrolment.
3. Anonymous data does not fall under FADP nor HRA. However, it is unlikely that individual patient data of a clinical study can be anonymized[Glossary].

8. 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 7).

The Ethical, Legal and Social Implications (ELSI) advisory group, which is part of the Swiss Personalized Health Network (SPHN) (31) initiative, published a framework providing ethical guidance on processing and sharing personal data within SPHN hereafter referred to as the ELSI framework (32). 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 otherwise (*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 the context of biobanking and related to further use of health-related data and samples collected in routine medical care (33). As described in section 7, sharing data from clinical research projects requires explicit informed consent because the consent given by the patient allows 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 usually 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"). In this sense, the framework is applicable to the sharing of clinical trial data. As described in section 7, 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 (see also section 10 for data processing before data can be shared). This ensures patients' personal rights to provide information on the type, storage, and reuse (sharing) of her/his data on request and ensures 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 10) or if data are shared on the basis of a contract. We

consider the latter option most appropriate and should be the default setting (see section 16). 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 de-identification of their data.

R2. If sharing of coded data is planned:
   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. Sponsor(-investigator)s should ensure that a system is in place that allows access to this information centrally, e.g., by recording disagreements in the study database.

R3. If sharing of uncoded personal data is planned:
   a. Sponsor(-investigator)s must ensure that potential participants are informed about the potential sharing of personal data and the potential de-identification of their data for this purpose. Explicit written informed consent should be sought.
   b. Sponsor(-investigator)s 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.

R4. For sharing data collected in 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 patient has 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.

9. Data management plan

According to (34)(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 and 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 (R7). 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

10. De-Identification

10.1. Goal

Sharing patient data with someone not involved in the patients' treatment or diagnostics conflicts with the obligation to protect the patients' privacy. In this section, we describe means to decrease the risk that individual patients are identified on the basis of shared research data. We do not think that all the techniques described below must be applied in every project. Instead, risks and benefits of sharing a specific data set need to be considered carefully especially taking into account what patients consented to (section 8) and then the appropriate techniques should be used.[4]

According to the Swiss Federal Law, *anonymization* is the act of processing personal data in such a way that identification of individual persons is impossible or possible only with disproportionate effort (18). This definition conforms with other regulatory frameworks such as the European General Data Protection Regulation (35). However, 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 (30). As a consequence, data that are anonymized today might not remain anonymized according to this definition. This implies that individuals might be identifiable in the future even if data is considered anonymized at the time it is shared/prepared for sharing. The legislation takes this into account: it is the time point at which the dataset is prepared and the technical capabilities at that time that determine whether data is *anonymous* or not. However, if data from clinical studies is thoughtlessly treated and shared as *anonymized*, the negative consequences in case of a future breach are likely immense for clinical research in general and data sharers in particular. This alone implies for us, that clinical research data should rarely, if ever, be treated and shared as anonymized data even after appropriate processing. Researchers should carefully weigh the benefits and risks and, if in doubt, should consider their data identifying. Furthermore, the claim that individual patient data are anonymized in a strict sense, even after appropriate processing, appears unrealistic and we refer to the literature for a discussion on the topic (36,37). In short, the amount and precision of information that is available in the data collected in clinical research, even after extensive processing, plus the technical capabilities nowadays and general availability of additional data are at the core of the problem and make identification of individuals e.g. via linkage not a disproportionate effort.

[4] It is also relevant how the target repository defines the requesting process (section 16), but repositories might further develop interoperability, so data may at some point be accessible through different portals. Requirements from selection of repository hence are less clear and it is reasonable to assume a standard contract or license.

Within this document, we use *de-identification* to term the means to protect participants' privacy in a way that criminal acts would be necessary for identifying patients with shared data. We see de-identification as part of a standardized and institutionalized data sharing process, in which the data requester, based on a standard contract or license, agrees not to try to identify patients, not to give the data to other persons, and to maintain data security (see sections 16). In this setting and with these restrictions, de-identification is acceptable, given the potential benefits of shared data. Note that the term *coded data* in the sense of section 8 refers to a specific form of de-identified data. Coded data imply the existence of a key that allows for re-identification but only under very specific conditions i.e. because patients must have the opportunity to withdraw consent at any time and because it must be possible to inform individuals whenever relevant findings are identified during the research process (obligation to inform) see also Articles 26 and 27 in (35). Obviously, the de-identification process consists of manipulations that *change* the data. So, the means undertaken should protect patients' privacy while maintaining usefulness of the data. Of note, de-identification of data might be time consuming and require specific expertise in data management and the research field itself.

10.2. 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 that might be helpful in this context. The following is a non-exhaustive list (38):

- 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 (39)).
- 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, 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

10.3. The process of de-identifying data

De-identification is a multistep process that requires input by several people, among them the sponsor and data manager or 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[5].

10.3.1 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:

[5] Note that (22) 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). We think that control about tendentious selection of records from a research database is in general outside the sphere of influence of data sharing.

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).
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 (40). 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 (41).

Each variable should be classified whether it is:

- (Potentially) Directly identifying (see section 10.2),
- 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.

10.3.2 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[6], rounding of continuous data, grouping and aggregation (categorization), introducing random noise (jittering, perturbation), setting

[6] Here, the issue with linkage with external information becomes immediately obvious: nowadays trials must be registered in publicly available registers. These entries usually contain the enrolment from which it is then straightforward to restore the exact dates (depending on the precision of the relative time).

certain values to missing (suppression), data swapping, resampling or subsampling.

The Appendix provides additional details and examples.

10.3.3 Data processing

The steps as defined in 10.3.2 have to be programmed using (statistical) software and a set of new data files has to be generated.

10.3.4 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/programmer 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 de-identification

R8. De-identification should involve at least the sponsor and the statistician/data manager.

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 de-identification process should be quality controlled and appropriately documented.

11. Data structure and format

Full descriptive information of the data is necessary (see the coding variables section 12) 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 (22), 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, 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 possible given constraints of cost, time to implement, and technical capabilities, CDISC standards should be employed for new trials at the database design stage.

11.1. 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 can establish the correct link across the different files (section 13). The original key-value pairs will usually be replaced with new random unique identifiers (see section 10).

Alternatively, it is also possible to share a flat file which contains all data of a study in one file. Depending on the complexity/structure of the data, such a data set might be more difficult to understand for the data requester than the original structure, though.

11.2. 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 if provided in the format of the statistical software used for the original analysis but 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 (sections 12 and 13).

11.3. 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) (42) 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 (e.g., 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.

12. 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 de-identification, see section 10). 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.

12.1. Variable types

Individual variables usually come in four main types: date/time, text/string, 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 10). 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 10).

Categorical variables comprise binary (yes/no), single choice (male/female), multiple-choice or ordinal type variables (e.g., New York Heart Association Functional Classification scores to classify heart failure or adverse event grading schemes). 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 section 19 for an example: Table 3). 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 section 10). If the retention of free text is necessary, no special treatment beyond those measures outlined in section 10 is necessary.

Some database systems incorporate system-level variables into the dataset 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".

12.2. 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 section 19, Table 2.

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

The appendix (section 19) 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).

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.

12.3. 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 time (e.g., 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. Section 19 provides an example by looking at fictitious eligibility and blood label values forms (Table 4 and Table **5**).

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 (section 13). In Table 6 (section 19) 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 (section 19, Table 7). 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 section 19 (Table 8).

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.

13. 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 (see also sections 11 and 12). This documentation enables someone not involved with the study to understand and use the data appropriately. In addition, metadata allows to find the data (section 5.4). This section gives a definition of the term metadata and what we think the documentation should contain, at a minimum, to go along with shared clinical research datasets.

13.1. 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 (43). 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 (43,44).

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 (45) as far as the involved metadata schemes cover the same aspects.

Canham et al. (46) suggest the use of a minimal extension of the DataCite metadata scheme for clinical research data (47) 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 (48) while the International Committee of Medical Journal Editors (ICMJE) provided guidelines (27). Section 19.3 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 are 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.

13.2. Additional documentation

In addition to metadata, further documentation is needed to make use of the data. As described in sections 11 and 12, 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 should be provided, such as derivation of variables, and also the process of rounding or jittering data for de-identification (see section 10) has to be described together with its impact on the result, if applied.

The study protocol and statistical analysis plan[Glossary] with amendments contain a large part of the information needed, but researchers must 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 (section 19.4)
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 vocabulary or catalogue used

j. Code for data preparation
k. Description/brochure of a new tool or drug, if applicable
l. Documentation of means undertaken for de-identification
m. Data management plan

13.3. 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), unless a single flat ready-to-analyze data file is shared. 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.

14. 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 8: 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.

15. Selection of repository

Infobox 3: Data repository versus (clinical trial) registry

Registries: A clinical trial registry is a collection of records about clinical trials according to an agreed set of metadata (49). 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 (48). 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" (49), 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 (24) and allows for longer term storage. Technically, a repository consists at least of a backend, a database to store metadata and information, and a file server to store the datasets and other digital artifacts, and a web-based frontend that allows users to access the backend.

15.1. Principles

According to the FAIR data principles, research data should be findable, accessible, interoperable, and reusable (23,24), see section 2. Principle F3 mandates that "(meta)data are registered or indexed in a searchable resource" (23). 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 (24). 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" (22) and we support this view.

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 (50), 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.

15.2. Time point

Ideally, the appropriate repository is identified before writing the Data Management Plan (see section 9) and then described therein. We assume that a sponsor/investigator uses the same repository for all her/his projects so this should be feasible.

15.3. 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 (51).

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 (52). 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 (53). 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 (54), 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.

15.4. 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 (55), shaped as a checklist (section 19.5, Table 10). 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 (56).

Box 9: 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)

16. Requesting and use of data

Principle 6 of the ECRIN statement (22) 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 (22) 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 must 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) of the acting person and rules of conduct (Good Clinical Practice). Any different purpose would have to be explicitly mentioned in the informed consent (section 8).
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 (57). 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 (58). 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.

17. References

1. 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
2. Ioannidis JPA. Why most published research findings are false. In: Getting to Good: Research Integrity in the Biomedical Sciences. 2018. p. 2–8.
3. Correction Chalmers I, Bracken MB, Djulbegovic B, Garattini S, Grant J, Gülmezoglu AM, et al. How to increase value and reduce waste when research priorities are set. Vol. 383, The Lancet. 2014. p. 156–65.
4. Bhopal RS. Increasing value and reducing waste in biomedical research. Vol. 388, The Lancet. 2016. p. 562.
5. Chan AW, Song F, Vickers A, Jefferson T, Dickersin K, Gøtzsche PC, et al. Increasing value and reducing waste: Addressing inaccessible research. Vol. 383, The Lancet. 2014. p. 257–66.
6. Clyne B, Boland F, Murphy N, Murphy E, Moriarty F, Barry A, et al. Quality, scope and reporting standards of randomised controlled trials in Irish Health Research: an observational study. Trials [Internet]. 2020 Dec 8;21(1):494. Available from: https://trialsjournal.biomedcentral.com/articles/10.1186/s13063-020-04396-x
7. Jin Y, Sanger N, Shams I, Luo C, Shahid H, Li G, et al. Does the medical literature remain inadequately described despite having reporting guidelines for 21 years? – A systematic review of reviews: an update. J Multidiscip Healthc [Internet]. 2018 Sep;Volume 11:495–510. Available from: https://www.dovepress.com/does-the-medical-literature-remain-inadequately-described-despite-havi-peer-reviewed-article-JMDH
8. Debray TPA, Moons KGM, Valkenhoef G, Efthimiou O, Hummel N, Groenwold RHH, et al. Get real in individual participant data (IPD) meta-analysis: a review of the methodology. Res Synth Methods [Internet]. 2015 Dec 19;6(4):293–309. Available from: https://onlinelibrary.wiley.com/doi/10.1002/jrsm.1160
9. Kanters S, Karim ME, Thorlund K, Anis A, Bansback N. When does the use of individual patient data in network meta-analysis make a difference? A simulation study. BMC Med Res Methodol [Internet]. 2021 Dec 13;21(1):21. Available from: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-020-01198-2
10. Kawahara T, Fukuda M, Oba K, Sakamoto J, Buyse M. Meta-analysis of randomized clinical trials in the era of individual patient data sharing. Int J Clin Oncol [Internet]. 2018 Jun 12;23(3):403–9. Available from: http://link.springer.com/10.1007/s10147-018-1237-z
11. Baker M. 1,500 scientists lift the lid on reproducibility. Nature [Internet]. 2016 May 26;533(7604):452–4. Available from:

https://www.nature.com/articles/533452a

12. Fidler F. Reproducibility of Scientific Results. In: The Stanford Encyclopedia of Philosophy [Internet]. 2018. Available from: https://plato.stanford.edu/archives/win2018/entries/scientific-reproducibility

13. Goodman SN, Fanelli D, Ioannidis JPA. What does research reproducibility mean? Sci Transl Med [Internet]. 2016;8(341):341ps12. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27252173

14. Whitaker K. Reproducibility in brains science. 2017;

15. Gelfond JAL, Heitman E, Pollock BH, Klugman CM. Principles for the ethical analysis of clinical and translational research. Stat Med [Internet]. 2011 Oct 15;30(23):2785–92. Available from: https://onlinelibrary.wiley.com/doi/10.1002/sim.4282

16. Drummond C. Reproducible research: a minority opinion. J Exp Theor Artif Intell [Internet]. 2018 Jan 2;30(1):1–11. Available from: https://www.tandfonline.com/doi/full/10.1080/0952813X.2017.1413140

17. Dahabreh IJ, Robertson SE, Steingrimsson JA, Stuart EA, Hernán MA. Extending inferences from a randomized trial to a new target population. Stat Med [Internet]. 2020 Jun 30;39(14):1999–2014. Available from: https://onlinelibrary.wiley.com/doi/10.1002/sim.8426

18. The Federal Assembly of the Swiss Confederation. CC 810.30 Federal Act of 30 September 2011 on Research involving Human Beings (Human Research Act, HRA) [Internet]. 2020. p. https://www.fedlex.admin.ch/eli/cc/2013/617/en#cha. Available from: https://www.admin.ch/opc/en/classified-compilation/20061313/index.html

19. The Swiss Federal Council. Ordinance on Human Research with the Exception of Clinical Trials (HRO). 2021;(September 2013):1–30. Available from: https://www.admin.ch/opc/en/classified-compilation/20121177/index.html

20. The Federal Authorities of the Swiss Confideration. FADP: Federal Act on Data Protection of 19 June 1992 (Status as of 1. January 2014) [Internet]. 2014. p. 1–24. Available from: http://www.admin.ch/ch/e/rs/c235_1.html

21. Gahl B, Haynes a, Sluka C, Dupuis-Lozeron E, Jörger F, Schur R, Christen A TS. Sharing of data from clinical research projects – Guidance from the SCTO's CTU Network [Internet]. 2020. Available from: https://www.preprints.org/manuscript/202006.0344/v1

22. 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;7(12).

23. FORCE11. The fair data principles. (訳) NBDC研究チーム. FAIR原則 (「The fair data principles」和訳) [Internet]. 2016. Available from: https://biosciencedbc.jp/about-us/report/fair-data-principle/

24. 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;3.

25. Go-Fair. FAIR Principles [Internet]. Go-Fair. 2021. p. 1–3. Available from: https://www.go-fair.org/fair-principles/

26. SCTO. Guidelines for Good Operational Practice - Swiss Clinical Trial Organisation [Internet]. 2017. Available from: https://www.scto.ch/en/publications/guidelines.html

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

28. FORCE11: Data Citation Synthesis Group, 研究データ利活用協議会: リサーチデータサイテーション小委員会翻訳. Joint Declaration of Data Citation Principles. [データ引用原則の共同宣言 : 最終版] [Internet]. 2014. p. 1–4. Available from: https://japanlinkcenter.org/rduf/doc/rduf_rdc_jddcp_ja.pdf

29. Confederation S, Assembly TF. Federal Act on Research involving Human Beings. Http://WwwAdminCh/Opc/En/Classified-Compilation/20061313/IndexHtml. 2014;1–22.

30. 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/

31. Network SPH. Swiss Personalized Health Network. Swiss Pers Heal Netw [Internet]. 2019;16:https://www.sphn.ch/en.html. Available from: https://www.sphn.ch/%0A

32. 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

33. Husedzinovic A, Ose D, Schickhardt C, Fröhling S, Winkler EC. Stakeholders' perspectives on biobank-based genomic research: Systematic review of the literature. Vol. 23, European Journal of Human Genetics. 2015. p. 1607–14.

34. Trial Master FIle. In: Definitions. 2020.

35. European Parliament and Council of European Union (2016). Regulation (EU) 2016/679 [Internet]. Official Journal of the European Union 2016 p. 156. Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN

36. 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.

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

38. Kops SR. The Health Insurance Portability and Accountability Act of 1996 (P.L. 104-191). Benefits Q. 1997;13(2):8–13.

39. Statistik B für. MEDSTAT.

https://www.bfs.admin.ch/bfs/de/home/statistiken/gesundheit/nomenklaturen/medsreg.html.

40. Garfinkel. De-identification of personal information. NISTIR 8035. 2015;
41. Atreya R V., Smith JC, McCoy AB, Malin B, Miller RA. Reducing patient re-identification risk for laboratory results within research datasets. J Am Med Informatics Assoc. 2013;20(1):95–101.
42. Pike R. UTF-8. 2019.
43. Oulun Yliopisto. Research Data Guide [Internet]. Managing research data: Data documentation and metadata. 2018. Available from: http://libguides.oulu.fi/Researchdata/Data_documentation
44. University ta trobe. Durable file formats. 2019; Available from: https://latrobe.libguides.com/dataorganisation/fileformats
45. Open Archives Initiative Protocol for Metadata Harvesting [Internet]. 2013. Available from: http://www.openarchives.org/pmh/
46. Canham S, Ohmann C. A metadata schema for data objects in clinical research. Trials. 2016 Dec;17(1):557.
47. Starr J, Ashton J, Brase J, Bracke P, Gastl A, Ziedorn F. DataCite Metadata Schema for the Publication and Citation of Research Data (Version 2.1). Hann TIB doi [Internet]. 2016;29. Available from: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:DataCite+Metadata+Scheme+for+the+Publication+and+Citation+of+Research+Data#1
48. World Health Organization. WHO Data Set [Internet]. WHO. World Health Organization; 2019. p. 1. Available from: https://www.who.int/ictrp/network/trds/en/#.XqcBOQAL0uo.mendeley%0Ahttp://www.who.int/ictrp/network/trds/en/
49. WHO | Glossary [Internet]. WHO. Geneva, Switzerland: World Health Organization; 2018 [cited 2018 Dec 6]. Available from: http://www.who.int/ictrp/glossary/en/
50. CoreTrustSeal – Core Trustworthy Data Repositories [Internet]. Available from: https://www.coretrustseal.org/
51. Sciences swiss academies of arts and. Swiss Academies Factsheets. 2019. p. http://www.akademien-schweiz.ch/index/Publikatione.
52. Bill & Melinda Gates Foundation. Gates Open Research. Seattle, WA: Bill & Melinda Gates Foundation; 2017.
53. HLEG EOSC. European Open Science Cloud | Open Science - Research & Innovation - European Commission [Internet]. 2016. Available from: https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud
54. Swissuniversities. Olos. 2018.
55. Whyte A. Where to keep research data. DCC checklist for evaluating data repositories [Internet]. Edinburgh: Digital Curation Centre; 2015. Available from: http://www.dcc.ac.uk/resources/how-guides-checklists/where-keep-research-data

56. Preservation D, Questions FA, Based L, Format O. Levels of Digital Preservation Support. 2014;1–5.

57. 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). Farumashia. 2007;43(9):894–8.

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

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

18. 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	https://tmfrefmodel.com/wp-content/uploads/2018/03/tmf-rm-deliverable-user-guide-v1-2018-03-16.pdf

	a single document, data set or piece of information but in other cases it could be represented by multiple document types or data types.	
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 look 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	Human Research Act Art. 3h. HRO Art. 26-27.

	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). 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.	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 investigator site file/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 (49). 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 (48). 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" (49), 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 (24) 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.	(59)

19. Appendix

19.1. Further detailed specification of required data processing steps

19.1.1 Example data to be considered for deletion

- Names, address, etc. have to be deleted, see section 10.2.
- 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.

19.1.2 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(39) 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[7]), 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.

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

- 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.

19.2. Further details on coding of variables

19.2.1 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.

19.2.2 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	Labelname	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_1dp	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

19.3. 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		

19.4. 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

19.5 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	