Guidance to support writing a Data Management Plan in clinical research

Goal of this document:

This guidance document aims to assist students and researchers when writing a Data Management Plan for a clinical study. A short version of these guidance notes can also be consulted in DMPonline.be tool.

Contributors: RDM working group faculty of Medicine and Health Sciences and Data Stewards team of the UGent university library

Glossary:

- Clinical trial: is any research study that prospectively assigns human participants or groups of humans to one or more health-related interventions to evaluate the effects on health outcomes. Interventions include but are not restricted to drugs, cells and other biological products, surgical procedures, radiological procedures, devices, behavioural treatments, process-of-care changes, preventive care, etc (WHO definitition).

  NOTE: Make use of the Clinical trials register to register your trial, which includes data from the following providers: ClinicalTrials.gov and EU Trial Register. (take contact met Hiruz)

- Case report forms (CRF): a printed, optical, or electronic document designed to record all data required by the protocol on each trial subject and to be reported to the sponsor.
- Data Management Plan (DMP): is a written document that describes the data you expect to acquire or generate during the course of a research project, how you will manage, describe, analyze, and store those data, and what mechanisms you will use at the end of your project to share and preserve your data. A DMP is ideally created before or at the start of a research project, but updated where necessary as the project progresses (because things may change, or not all details may be known at the outset). **A DMP should be considered a 'living' document**
- Electronic Health Record (EHR): s an electronic version of a patients medical history, that is maintained by the provider over time, and may include all of the key administrative clinical data relevant to that persons care under a particular provider, including demographics, progress notes, problems, medications, etc.
- Good Clinical Practice (GCP) is an international ethical and scientific quality standard for the design, conduct, performance, monitoring, auditing, recording, analyses and reporting of clinical trials. It also serves to protect the rights, integrity and confidentiality of trial subjects. This guideline should be followed when generating clinical trial data that are intended to be submitted to regulatory authorities. The principles established in this guideline may also be applied to other clinical investigations that may have an impact on the safety and well-being of human subjects. GCP requires that the investigator ensures the accuracy, completeness, legibility, and timeliness of the data reported to the sponsor in the case report forms (CRFs) and in all required reports.
- Informed Consent: The process by which a patient learns about and understands the purpose, benefits, and potential risks of a medical or surgical intervention, including clinical trials, and then agrees to receive the treatment or participate in the trial.

What data will you collect or create?

e.g. Data origin, description, data type, data format, data volumes and (link to) existing data

Give a brief resume of the study, focusing on the data being collected or processed (e.g. patient surveys). Has similar or linked research already been performed; for which the data is available? What is the relationship? How does your research complement?

Describe the **type of data** that will be collected during various stages of the study in a structured way (see example below).

> NOTE: If you use articles as background information, as interpretation or to get ideas out of them, you should not consider them as a dataset in the context of a DMP. If, on the other hand, you analyse the articles e.g. re-analysis in terms of a systematic review, text analysis then it should be considered as a dataset.

Describe **all source systems** used such as Electronic Health Record (EHR) data, imaging data, subject files, patient surveys, records kept at the pharmacy, laboratories and at other medico-technical departments involved in the clinical trial.

> EXCEPTION: If you use ePRO (electronic Patient Reported Outcomes), as there is no source document to derive the data entered in the ePRO form from, GCP 6.4.9 says: The identification of any data to be recorded directly on the CRFs (i.e. no prior written or electronic record of data), and to be considered source data must be included in the description of the trial design. Please include this description and the ePRO forms themselves in the application to the Ethics Committee.
>
> NOTE: Data reported on the CRF, that are derived from source documents (e.g. Electronic Health Records) for analysis, should be consistent with the source documents or the discrepancies should be explained.

Outline and justify the chosen **file formats** of your imaging/diagnostic devices & other source systems for which binaries/files are (re)used by listing the file extensions and a short rationale. Describe which CRFs/instruments/surveys will be collected on paper and which will be collected electronically.

Example

- Study phase I: systematic review
    - Literature database: .pdf articles in EndNote database
    - Expected size: 500Mb
- Study phase II: survey
    - Online questionnaire through REDCap system, stored in RedCap as xyz and exported as .csv file
    - Survey of healthy volunteers, target n=50, to be recruited at regional hospitals
    - Expected size: 200Mb
- Study phase III: blood sample analysis
    - Collected from xxx patient cohort
    - eCRF in REDCap populated with patient parameters, as defined in the study protocol
        - Expected size: to be determined
        - File format?
    - Biomarker analysis
        - Elisa readout, digital file, .csv format, <1Mb
        - FACS, .fcs and .jpeg, 500Mb

Patient data for research purposes or data collected in a research project where the hospital is the sponsor, should be stored on secure storage of Ghent University Hospital. Ask ICT UZ Ghent for the best storing option related to the **data volumes** you (expect to) collect. If Ghent University storage can be used, you can follow the guidelines for storage for UGent researchers. Having an idea of how much data you will generate (e.g. in MB/GB/TB) and how the data volume will grow, can help you identify potential challenges in storing and transferring your data.

FAIRsharing.org is an online resource on **data standards** e.g. minimal required information for biological and biomedical investigations, standards to support acquisition, exchange, submission and archive of clinical research data (DOI: https://doi.org/10.25504/FAIRsharing.r87bgr).

If you intend to work with **existing or third-party data**, consider what (if any) costs are involved, and to what extent you are allowed to reuse and share these data. Regulatory or contractual restrictions (e.g. confidentiality agreements) might apply.

How will the data be collected or created? How will you guarantee data quality during data collection or creation?

How will you design the data collection tool (electronic or paper Case Report Form [CRF]) for accurate and appropriate capture of data? How does the data collection tool meets the regulatory requirements and the programme standards, user friendliness of completion and data entry, while meeting the needs of the protocol? How will you keep track of different versions? Refer to the SOP for CRF design when available.

Describe how data collection will be verified (double data entry, etc) and how collected data will be verified for accuracy, consistency and completeness: e.g. manual checking, automatic checking, discrepancy handling. Refer to SOPs or built-in procedures of your eCRF manager when available.

Source Data Verification, the process of cross-referencing data recorded in a CRF to the original source information, can be performed by a study monitor. Contact HIRUZ for monitoring (hiruz.ctu@uzgent.be).

NOTE: The investigator/institution should maintain adequate and accurate source documents and trial records that include all pertinent observations on each of the site's trial subjects. Source data should be attributable, legible, contemporaneous, original, accurate, and complete. Changes to source data should be traceable, should not obscure the original entry, and should be explained if necessary (e.g. via an audit trail). The investigator should ensure the accuracy, completeness, legibility, and timeliness of the data reported to the sponsor in the Case Report Forms (CRFs) and in all required reports. Data reported on the CRF, that are derived from source documents, should be consistent with the source documents or the discrepancies should be explained. An audit trail should be maintained (i.e. changes or corrections should be dated, initialled, and explained and should not obscure the original entry). This applies to both written and electronic changes or corrections. (Good Clinical Practice (GCP) guidelines).

Note that secure applications can assist in this process; such as REDCap (Research Electronic Data Capture). Contact HIRUZ (Health, innovation and research institute) for more information and support (hiruz.ctu@uzgent.be).

Specific guidance:

- To protect patients' privacy at all times, persons without a health profession and treatment relationship with the patient cannot **access the Electronic Health Record (EHR)** even if the patient gives his/her/its consent.
  (http://www.ejustice.just.fgov.be/cgi_loi/change_lg.pl?language=nl&la=N&cn=2002082245&table_name=wet)
- If you use a statistical analysis tool, generic spreadsheet or generic database application to collect subject data which can be structured by instrument and events (for longitudinal projects), consider using a more appropriate tool (e.g. REDCap (Research Electronic Data Capture)) to comply to the Good

Clinical Practice (GCP) requirement for accurate reporting, interpretation and verification; i.e. by retaining records of all changes and corrections (audit trail).

- Before analysis, the Principal Investigator must confirm that all collected data in the system are correct. For studies using **ePRO (questionnaire-based studies**), mechanisms should be built in to ensure that checks are in place to minimise the chances of error when completing the questionnaire.
- SOPs such as **eCRF Completion Guidelines** and a **Data Validation Plan** describe the workflows and procedures followed by the study staff to ensure data quality.

How will you document the data?

How will a description be provided of the implemented system for performing data collection, entry and handling? Describe how the data will be received and tracked. Describe if protocols used for data handling are available.

Document your data at the study level and at the data level. Think of reports, user guides, lab notebooks, informed consent forms, … as well as inventories of your datasets, processing scripts, data labels, codes, …

Standardize your documentation as much as possible using clinical ontologies and vocabularies. See also http://bioportal.bioontology.org/ . Use medical coding dictionaries to help categorize medical terms so that they can be analyzed/reviewed. e.g. ICD-10 codes for classifying diseases, MedDRA codes for medical products, CTCAE - Common Terminology Criteria for Adverse Events, etc . Keep record of the version of the medical coding dictionary used.

How will you manage any ethics and confidentiality issues?

If the project concerns an experiment on humans, or research on residual biological samples, or retrospective research using previously collected data it has to be submitted to and approved by the EC before the project can be moved to production and data collection can start. In general, all participants will need to sign an informed consent before data collection can start.

Contact HIRUZ (Health, innovation and research institute) for more information and support. HIRUZ has ICF templates and guides the Ethics Committee (EC) submission procedure. Contact HIRUZ for monitoring (hiruz.ctu@uzgent.be).

You must collect the data according to the the Belgian law of August 22nd 2002 on patient rights and Belgian law of 30 July 2018 on the protection of natural persons with regard to the processing of personal data; and to abide by all guidelines provided by the EC. You must follow the General Data Protection Regulation (2016/679) for data collection.

How will you manage intellectual property rights issues?

When the ownership or user rights to data are transferred from or to Ghent University Hospital, this must be perpetuated in a data transfer agreement (contact: hiruz.contracten@uzgent.be).
If this is done within the framework of an ongoing study, a data processing agreement must be attached as an annex to the study contract (contact: hiruz.ctu@uzgent.be).

How will you store and backup data during research?

Consider storage options of Ghent University Hospital and Ghent University. Patient data for research purposes should be stored on storage of Ghent University Hospital. Refer to the appropriate policies and procedures.

How will you ensure that stored data are secure?

Describe the applicable IT security policies. Describe appropriate access to the collected research data: consider definition of access levels for users and authorization and withdrawal of users.

Note that to protect patients' privacy at all times, persons without a health profession and treatment relationship with the patient cannot access the Electronic Health Record (EHR) even if the patient gives his/her/its consent.

GCP guidelines request that you maintain a security system that **prevents unauthorized access** to the data used in clinical research. For safely storage and handling patient data, use the UZ Ghent facilities and guidelines.

The General Data Protection Regulation (art32, recital 83) gives the controller a catalogue of criteria to be considered when choosing methods to **secure personal data**. Those measures should ensure an appropriate level of security, including confidentiality, taking into account the **state of the art** and the costs of implementation in relation to the risks and the nature of the sensitive personal data to be protected. **Encryption** as a concept is explicitly mentioned and is the best way to protect data during transfer and one way to secure stored personal data.

Which data should be retained for preservation and/or sharing?

List which part of the data (refer back to your data (type) description) can or should be preserved after the end of the project and why (not).

The following legal requirements can apply:

- ➔ According to The Belgian Act of 7 May 2004 on experiments on human persons, or the Belgian Experiments Act, data must be preserved for 20 years.
- ➔ Guideline on the content, management and archiving of the clinical trial master file (paper and/or electronic): the sponsor and the investigator shall archive the content of the clinical TMF for at least 25 years after the end of the clinical trial."
- ➔ The Clinical Trial Regulation: archiving of the Trial Master File: 25 years
- ➔ The term for preservation of the essential documents of a clinical trial is at least 25 years after the end of the clinical trial (preservation by the sponsor and the investigator). (VERORDENING (EU) Nr. 536/2014)

NOTE: The data to be collected is also specified in the protocol of your clinical trial.

What is the long-term preservation plan for the selected datasets?

Describe the archiving of the project data to ensure security and confidentiality of the data, to allow comprehensive reconstruction of the completed work and to ensure regulatory requirements regarding retention. This could mention: the electronic archiving system, reference data (normal ranges, coding), timing and length of archiving, submission and retrieval procedure.
Archiving will be authorised by the Sponsor following submission of the end of study report. The investigator and sponsor specific essential documents will be retained for at least 20 years after end of trial. At that moment, it will be judged whether it is necessary to retain them for a longer period, according to applicable regulatory or other requirement(s).

Data may be provided a repository, provided that the data set is pseudonymised as well as possible so re-identification should be practically impossible. Personal Health Information as collected in clinical research is sensitive data. General Data Protection Regulation (GDPR) (art 25) recommends **pseudonymization** of personal data as part of "Data protection by design and by default". Personal data which have undergone pseudonymization are considered information on an identifiable natural person (and thus are protected by the GDPR) if the personal data which have undergone pseudonymization could be attributed to a data subject by the use of additional information e.g. the **subject identification log**. Without this additional information, re-identification should be practically impossible, taking into account current known techniques and technological development.
Often a distinction is made between direct and indirect identifiers and indirect identifiers can lead to identification when additional data are leveraged. Unless permitted by the Ethics Committee, **no direct identifiers** (such as name, email, address, telephone number, among others) can be stored together with the research dataset.

Which repository can I use for preservation of my data?
Search re3data.org and select the filter option Life Sciences (or medicine) when searching for an **international repository** e.g. for sequence data, microarray data, bioassay data, substance or sequence-based reagents, human clinical data, genetic tests or validated measurement instruments. ClinicalTrials.gov is an example. It is a registry and results database of publicly and privately supported clinical studies of human participants conducted around the world.
Verify the quality of the repository and check if depositing the data in this repository is in accordance with the information in the Informed Consent Form.

There are many **regional centralized information repositories** for record level data of specific disease areas and medical procedures. These registries serve a variety of objectives; validation of procedures, case ascertainment, quality control, processing and analysis, coding and storage, reporting, making data accessible, evaluate health status or the quality of care provided.
Examples: Belgian Cancer / Transplant / Neuromuscular diseases / Diabetes / Medical Implant Devices / Cystic Fibrosis / Orthopedic Prosthesis.

Are any restrictions on data sharing required?

List reasons for not (immediately) sharing. Think of personal data, sensitive, confidential data or data with valorisation potential. Personal data for example could restrict sharing of part of the research data if proper consent was not obtained or when the informed consent forms contain limitations on dissemination. Also data which is obtained as part of a collaborative study (e.g. with industry) may be subject to restrictions for sharing.

> NOTE: The dissemination policy should be also stated in the protocol of your clinical trial.
>
> NOTE: Genetic information of living humans is considered not only personal data, but sensitive personal data under GDPR.

Describe whether data will be shared upon publication of the results, upon finalization of the project funding, x year after first or last patient inclusion, … Explain if there is a need for a delay.

It is possible that a journal requests that the trial protocol, the anonymised participant level dataset, and the statistical code for generating the results to be submitted as separate appendices to the full study report files and be made publicly available at the journal's website after acceptance of publication. These data can be submitted provided that the data set is adequately prepared and when there are no further restrictions, as mentioned above. More information can be found here (UGent intranet: please make sure you are logged in).

Specific guidance:

- Sharing of personal data between healthcare professionals: via e-health applications (e.g. CoZO)

Who will be responsible for data management?

The Principal Investigator (PI) is responsible for the conduct of the clinical trial at a trial site and authorises the delegation of all data management activities. It is recommended to also list the delegated roles in the data management plan, when already available.

> NOTE: The role of Principal Investigator (PI), Project Coordinators, Trial Associates, Data Quality Manager, Monitors, Auditors, Statistician of the research project is defined in the **delegation log**.
> These roles are described in a delegation log and should be reflected in the EDC - Electronic Data Capture or CDMS - Clinical Data Management System by only giving the necessary user rights to the different roles. HIRUZ can provide a template Delegation log and offers an EDC system.
>
> NOTE: For collaborative research projects, the PI of the coordinating center is called the Coordinating Investigator (CI), and each site has a PI.

The sponsor is responsible for implementing and maintaining quality assurance and quality control systems with written SOPs to ensure that trials are conducted and data are generated, documented (recorded), and reported in compliance with the protocol, GCP, and the applicable regulatory requirement(s).