Research Data Management Policy

1. Background

1.1. Topic

Policy on the management of research data.

1.2. Description

An important ambition of UMC Utrecht is 'Open Science ' for better health care and social involvement. This is about 'open access' (free access to scientific publications), 'open data' and reuse of scientific data, making science more efficient, more widely applicable and reproducible. Data is the foundation of scientific research. Throughout the research cycle, the quality and durability of data must therefore be ensured through careful management. If the quality and durability of data is not in order, the research is not reproducible. Making the data 'open' and reusable is not considered. Research Data Management (RDM) deals with the collection, documentation, processing, storage, sustainable archiving and reusability of the research data. Funders and publishers also take part in above ambition and ask for a data management plan, wherein the researcher demonstrates that the data to be collected, will be available for reuse in the future.

The policy is based on the following documents:

- The Research Code of UMC Utrecht
- Privacy Regulations of UMC Utrecht
- The General Data Protection Regulation (GDPR) and the General Data Protection Regulation Implementation Act
- The Handbook for Adequate Natural Datastewardship (HANDS), prepared by the Dutch Federation of University medical centers (NFU).

1.3. Target audience

All employees involved in scientific research must read and comply with this policy, in particular: department heads, managers research, researchers, quality coordinators, policy advisors, data managers, research coordinators and research nurses .

1.4. Application area

The scope of this policy document is the data management of all scientific research, which is carried out under the responsibility of UMC Utrecht. This also concerns multicenter research in which an external party bears the main responsibility, and the UMC Utrecht collects data as a participating center.

1.5. Roles and Responsibilities

The roles and responsibilities are described in the Research Data Governance Model.

1.6. Aim

Establishing this policy must guide all involved parties in the adequate design and implementation of Research Data Management.

2. Principles

1. FAIR data

For all scientific output we strive for the highest possible degree of 'Fairness', where FAIR stands for:

- **F**indable: datasets are *findable*, provided with a persistent identifier[1] and sufficient metadata[2], with which specific datasets can be found;
- **A**ccessible: datasets are *accessible*, either at data level or at metadata level, and clear user conditions have been drawn up;
- **I**nteroperable: datasets are *interchangeable* and standard terminologies and vocabularies are used wherever possible;
- **R**eusable: datasets are reusable, meet the above criteria and are sufficiently described, so that reuse is scientifically and ethically sound.

2. Privacy by Default/ Privacy by Design

The use of personal data for research is based on:

- A. Privacy by Default : the chosen standard method maximally guarantees the privacy of the data subject[3] (such as use of standard research folder design). The Privacy by Default idea is built into the technology of, for example, data collection and other work processes.
- B. Privacy by Design: consider privacy when designing the study (such as minimizing the amount of personal data and pseudonymising as soon as possible).

3. Dataminimization

Not only in the context of privacy laws, but also in the context of efficiency and scientific integrity, the collection of research data (and in particular of personal data) needs to be demonstrably legitimate, fit for the purpose of the study, and contain no irrelevant data. However, some research requires a broader collection of data (for example to find accidental discoveries in a certain clinical picture in cohort studies). This is described in the Data Management Plan.

4. Data Protection Impact Assessment (DPIA)

A DPIA[4] ensures that the impact of the intended project on the privacy of those involved and the risks for those involved and for the organization is considered before the start of the project. For processing of health information that is directly or indirectly traceable to the person (sensitive personal data), we assume that the outcome of a DPIA will be 'high risk' . Therefore, for each study it is described which measures will be taken to minimize the impact on the privacy of the study participants. This is described in the Data Management Plan.

5. Research Data in control

In line with the general standards of the Netherlands Code of Conduct for Research Integrity 2018 it is ensured that data sources are verifiable, that there is an honest and transparent description of the data collection, that both raw and final data are carefully stored and that the data are as FAIR as possible. It is also ensured that all data and research materials are collected with the appropriate infrastructure, that they are archived in a sustainable manner and are as 'open' as possible and as 'closed' (confidential) as necessary.

---

[1] A **persistent identifier** (PI) is a permanent link and unique label to a digital object, independent of the storage location. The unique label ensures that the digital object is retrievable from the internet at any time, even when the name of the digital object or the storage location changes.

[2] Metadata are data about the research data, such as a description of the data source, a codebook of the dataset or a description of the data collection method.

[3] data subject : the identified or identifiable natural person of who personal data are collected

[4] DPIA=Data Protection Impact Assessment and PIA=Privacy Impact Assessment are used interchangeable. In this document we agree with the internationally used DPIA. In NL the term GEB (Gegevens Effect Beoordeling) is often used.

6. Data Sharing Statement

In the context of Open Science, the sharing of scientific data with other researchers is encouraged. A data sharing statement is a statement about which research data will be shared after completion of the research, how the data will be made available and under which conditions this will be done. The data sharing statement provides a good description of why data can or cannot be shared and is therefore a compulsory part of the Data Management Plan for all scientific research. This requires compliance with laws and regulations in the field of privacy. This is based on the principle "As open as possible, as closed as necessary".

7. Researchers access to data expertise

Like other Dutch UMCs, UMC Utrecht follows the NFU Guidelines for data stewardship. We view data as legitimate and citable products of research, and are transparent about how the data can be viewed or used[6]. The support of researchers by data management experts is organized differently per division. In the coming year, data stewardship will be formalized, as described by NFU Data4lifesciences, to provide a more policy-oriented approach to data management within the Divisions.

3. Implementation framework

3.1. Preparation of the research

3.1.1. Data Management Plan (DMP)

In the context of integrity and transparency of research and FAIR data for Open Science, UMC Utrecht researchers are required to fill in a DMP for scientific research[7] using the UMC Utrecht specific DMP template, unless a funder has a mandatory template[8]. The UMC Utrecht DMP template supports researchers with examples of answers, and with information about the specific subjects. Data managers can be consulted if necessary. With this , the UMC Utrecht aims for transparency in data management processes and compliance with legislation and regulations. The content of the DMP is in line with national and international initiatives[9].

3.1.2. Storage and organization of files

For research purposes, UMC Utrecht offers the Research Folder Structure. Separate storage of personal data and research data is possible here. The storage location is recorded in the Data Management Plan. The storage location is under management at UMC Utrecht unless otherwise stated in the Data Management Plan. The principal investigator is responsible for the proper management of the research folders. The organization and access to folders is carried out by an authorized employee on behalf of the principal investigator.

_____

[6] Research Code UMC Utrecht

[7] For all types of scientific research, inclusive all non WMO compliant research

[8] National and international development is that scientific institutions take responsiblility of the DMP content, instead of funders with a number of required items, as is described in

[9] ERC Information for ERC grantees

3.2. Data Collection

3.2.1. Collecting data

For the collection of data for research, automated data collection is used as much as possible, using the available infrastructure of the UMC Utrecht. For choosing the most suitable data collection system, a decision aid is available and the Data Manager can be consulted. When personal data must be processed in software hosted by an external party, a processor agreement with the software provider is needed. Audit trails, as used for years in Good Clinical Practice (ICH-GCP guideline), are mandatory for all scientific research.

3.2.2. Registration of research participation

In view of patient safety, it is important that every participation in scientific research is registered in the electronic patient database (EPD). This also applies to healthy volunteers. In case, for valid reasons, it is not desirable to register healthy subjects in the EPD, one can deviate from registration in the EPD (see SOP Performing WMO-study). This must be documented in the DMP.

3.2.3. Re-use of data

Re-use of healthcare data (retrospective studies) is possible after pseudonymization of the data (see also 3.3.2). Data that are collected in healthcare, are made available for research in the Research Data Platform (RDP). For specific expertise and guided searches in sources one can use the Utrecht Patient Oriented Database (UPOD). Making data available for reuse must always comply with applicable laws and regulations. For the use of external data sources for research, researchers turn to the supplier of the data and jointly draw up a Data Access Agreement. Often this is based on documents from the supplier.

3.2.4. Raw data storage

Collected data, whether this has been done via re-use or via new data collection, needs to be stored and frozen (read only) as 'raw data' immediately after collection. Raw data are not to be edited or changed. This is necessary, so that, at any time, one can trace back to the original data.

3.3. Data preparation

3.3.1. Validation

All adaptations for validation and preparation of a suitable dataset for analysis are documented in such a way that the research results are verifiable. For all scientific research it is mandatory to keep an audit trail. When the raw data consists of care data which are erroneously recorded in the electronic patient files (EPD), the usual care processes are followed to adjust that information. This means that only the care provider can change this data, and not the researcher.

3.3.2. Anonymize/Pseudonymize

Directly identifying personal data must be stored separately from the research data, and research data must be anonymised or pseudonymised. Care data are almost always indirectly traceable to the person and completely anonymizing will therefore often not be possible. A pseudonymised file can indirectly be traced back to a natural person, as long as a key file exists. An important reason to still pseudonymise data is that this removes the direct traceability to a person. Technical facilities (such as the RDP and the Research Folder Structure) and procedural provisions ensure that the researcher does not have access to the directly identifying data. In case of indirect traceability to a person in research files, appropriate measures are taken to remove the indirect traceability to an individual when publishing or sharing data[10].

3.3.3. Storage of analysis file

After validation and preparation, and before final analysis and reporting, the analysis file is to be frozen. The analysis file is pseudonymised, or anonymized. For separate storage of personal data and research data, the Research Folder Structure can be used. In a frozen data collection it is no longer allowed to change or add data. Statistical analyses are only performed on the frozen datasets.

3.3.4. Requirements for data storage

For all types of research, the storage location is physically and digitally protected. The policy of UMC Utrecht regarding information security is described in the Information Security Directive. UMC Utrecht provides a number of storage locations that meet these requirements, such as bulk storage and departmental disks. The DMP must describe where the data is stored, so that it is clear for every employee in the research team, where the data can be found.

3.3.5. Data exchange during a study

Exchanging data with internal colleagues and external partners during a study, is done in a safe manner according to the regulations of the UMC Utrecht, which can be found under this link. Prior to the exchange

and sharing of data with external parties, a Data Transfer Agreement has been drawn up, which establishes the conditions under which data are shared.

3.4.   Metadata and Documentation

Metadata provide the information needed to enable a broad scientific community to find, share and understand the content and context of a dataset. When compiling metadata, standardization is sought as much as possible. With the mandatory DMP, UMC Utrecht supports researchers in the implementation of good metadata of research data, which is a prerequisite for FAIR data.

3.5.   Data Analysis

The scripts for data analysis are saved and commented (thus creating an audit trail), so that, when the script is read by an experienced colleague researcher/statistician/data analyst, it is completely clear why the steps in the analysis have been taken and which choices have been made.

3.6.   Data Archiving

3.6.1.   Transfer of custody after completion

When a principal investigator leaves after completion of the research, the management of the research folder and all data must be transferred to the new responsible person or to the department head. Data collected for scientific research, in which UMC Utrecht is end responsible or performs part of the research, are stored in storage locations managed by UMC Utrecht, unless other contractual agreements have been made about this.

_____

[10] Suitable measures, e.g. aggregation of data (use of birth year in stead of birth date) or minimization of data in light of 'the more variables you combine, the higher the risk of indirect identification'.

3.6.2.   Documentation and Publication

To ensure that all relevant data remain accessible, detailed agreements must be made prior to the start of the study, about all relevant data and the format in which they will be made available or transferred after the study has ended (see also ICH-GCP 4.9. 0, 4.9.4 and 4.9.5.). In order to archive the research data, a data package is created, consisting of data, metadata and documentation. Which specific metadata and documents are archived, depends on the type of research. The data package contains all information for the research to be replicated. The Data Management Plan and the Division Datamanager offer support in the selection of files in a data package.

3.6.3.   Storage period

Concerning the retention period, there are different legal standards for different types of research: this is at least 15 years for research with humans (Basic Selection Document), unless otherwise contracted. Anonymous datasets receive, with applicable conditions, an infinite storage period by depositing the sets in a dedicated archive (see Data Archiving), so that the data is available for future scientific research.

3.6.4.   Data Archiving

Data storage and data archiving is not the same. Data can be stored safely, but this does not mean that data will be findable and accessible in the future. Data archiving, on the other hand, is about the sustainable accessibility of the research data, where the data are provided with metadata, documentation and a persistent identifier. Subsequently, it is decided where the data and/or metadata will be archived for sustainable accessibility. This does not necessarily mean that the data is available to everyone, see Data Sharing. Global accessible data archives include: DANS, 4TUdatacentrum, SURFSARA, CLARIN (NWO), EMBL-EBI.

3.7. Data Sharing

3.7.1. Accessibility

In addition to open access to publications, accessibility of the underlying research data is stimulated, within the framework of the privacy legislation, the control rights of the study participants and the contractual agreements with third parties and funders. For each separate study, the data access policy will have to be determined and described in the Data Sharing Statement (one of the principles of Research Data Management). Agreements about sharing data with a specific third party (both during the research and after publishing the research results) are contractually written in a Data Transfer Agreement.

3.7.2. Trusted Third Party (TTP)

For linking research data and data from external parties at patient level, a TTP is used, or, if possible, similar technological or procedural solutions within the UMC Utrecht.

4. Abbreviations and definitions

Term	Definitie
Anonymise	The removal of directly and indirectly identifying characteristics, such that the data in the dataset can in no way be traced back to a person (see also definitions of directly and indirectly identifying data below).
Data	See 'Research data'
Data Archive	An indexed repository where data is stored permanently.
Data Manager	A data manager is someone who deals with one or more aspects of data management. The content of the tasks can therefore vary greatly depending on the data manager function. These processors are aware of the content of this policy, current laws and regulations, possibilities and facilities in the field of data management. See also Data Steward.
Data Management (or data processing)	Any action or a set of actions relating to data, including collecting, organizing, storing, updating, modifying, retrieving, consulting, using, providing by means of forwarding, dissemination or any other form of making available, bringing together, interrelating, as well as the protection, erasure or destruction of data.
Data Management plan	A formal document that describes how data is dealt with during and after the research and who is responsible for the data management.
Data Steward	A Data Steward is someone who is responsible for professional and careful care of the data during all phases of the research. The Data Steward ensures the long-term - and sustainable data management.
Direct identifying data	Data that have a unique relation to a person, are called "identifiers". Examples of identifiers are name, address and date of birth. These data are, in combination with each other, so unique to a specific person that it is used to distinguish people in daily life. We therefore speak of directly identifying data. (definition of Autoriteit Persoonsgegevens) (also see Indirectly identifying data).
DPIA	Data Privacy Impact Assessment
EPD	Electronic Patient Database
FAIR data	**F**indable: datasets are findable, provided with a persistent identifier and sufficient metadata , with which specific datasets can be found; **A**ccessible: datasets are accessible, either at data level or at metadata level, and clear user conditions have been drawn up; **I**nteroperable: datasets are interchangeable and standard terminologies and vocabularies are used wherever possible; **R**eusable: datasets are reusable, meet the above criteria and are sufficiently described, so that reuse is scientifically and ethically sound.
GCP	Good Clinical Practice
GDPR	General Data Protection Regulation

Human research	All scientific research in which people are involved as a study participant and, in whatever way, personal data is collected from the study participant.
Indirectly identifying data	Individuals can be identified on the basis of less direct identifiers. Think of external characteristics (length, posture and hair color), social and economic characteristics (profession, income or education) and online identifiers such as IP addresses. Although this data in itself usually does not allow us to identify a person, they can still lead to identification by their mutual connection or by linking to other data. We therefore speak of indirectly identifying data. (Definition Autoriteit Persoonsgegevens) (also see Direct identifying data).
KNMG	Koninklijke Nederlandsche Maatschappij tot bevordering der Geneeskunst
Metadata	Metadata provide the information needed to enable a broad scientific community to find, share and understand the content and context of a dataset.
Metadata standard	There are various metadata standards in circulation. The best known and most used is Dublin Core.
Personal data	Any information concerning an identified or identifiable natural person. A person is identifiable if his or her identity can be reasonably determined, without disproportionate effort. This refers to physical identity, not of deceased persons.
Pseudonimise	The replacement of direct identifying features with a code, in which the relationship between code and person is stored in a separate table and where, based on procedures, it is possible to go back to the identified person.
Raw data	See Source data
Research data	All data, with the exception of the data that can directly be traced back to the person, which are necessary to provide an answer to the research question.
Research Data Platform	UMC Utrecht Datawarehouse with data for research.
Research Data Governance model	A model that describes how the responsibilities for the various phases in Research Data Management are invested, which policy documents and procedures follow from them, who is responsible for them, which infrastructure has been developed and which support is offered.
Research Folder Structure	A folder structure provided by the Direction Information Technology, in which the different folders have differentiated authorizations for access, such that within one project separate folders are set up for the access to personal data.
Source Data	All raw, unprocessed data in original records and certified copies of original registration documents relating to findings, observations and other activities in a study necessary for the reconstruction and evaluation of the study. Source data are located in source documents or databases.
Study participants	A study participant is a patient or healthy participant who is involved in the research.
TTP	Trusted Third Party
UAVG	Uitvoeringswet Algemene verordening gegevensbescherming; Dutch implementation law on the GDPR
WBI	Wet Beoordeling Instrumentarium; Law on assesment of tools
WGBO	Wet op de Geneeskundige Behandelingovereenkomst; Medical

	Treatment Contracts Act
WMO-research	Research that falls under the scope of the 'Wet Medisch-Wetenschappelijk Onderzoek met Mensen (WMO)'.

5. Relevant links

Basisselectiedocument openbare en bijzondere universitaire medische centra

CCMO-Gedragscodes

Connect pagina AVG & Onderzoek

Gedragscode Gezondheidsonderzoek Federa

Handbook for Adequate Natural Datastewardship (HANDS), opgesteld door de Nederlandse Federatie van Universitaire medische centra (NFU).

ICH-GCP-richtsnoer;

Informatiebeveiliging van het UMC Utrecht

KNMG richtlijn 'Omgaan met medische gegevens'

KNMG richtlijn omgaan met medische gegevens, september 2016

'Kwaliteitsborging Mensgebonden onderzoek' NFU;

Nederlandse Gedragscode Wetenschappelijke integriteit.

Privacy Reglement UMC Utrecht

Privacy richtlijn - geheimhouding medische gegevens

Research Code van het UMC Utrecht

SOP Uitvoeren van WMO-plichtig onderzoek

Taichman DB, Sahni P, Pinborg A, Peiperl L, Laine C, James A, et al., 2017))

Uitvoeringswet Algemene Verordening Gegevensbescherming, zoals door de Tweede Kamer aangenomen op 13 maart 2018

Wet Medisch- Wetenschappelijk Onderzoek met Mensen (WMO)

Wet op de beroepen in de individuele gezondheidszorg (WGBO)

Wet op de Geneeskundige Behandelingsovereenkomst (WGBO)

Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship.
 Sci. Data 3:160018 doi: 10.1038/sdata.2016.18 (2016)

Documentkenmerken	
DocID	
Versie	3.0
Datum autorisatie	8-1-2019

Organisatiekenmerken	
Opdrachtgever	Raad van Bestuur
Doelgroep	UMC Utrecht

Totstandkoming en goedkeuring	Datum
Beheerder	Research Office
Status document	Definitief
Goedkeuringsstatus	Goedgekeurd