# PROJECT DELIVERABLE REPORT

| | |
|---|---|
| DELIVERABLE NUMBER | D5.1 |
| TITLE | ETHICS |
| AUTHOR(S) | M.ALLIGIER, W.PINXTEN, L.STEVNER, M.PINART, C.ROUSSEAU, F.MATTIVI, A.HODGE, A.ASSPOLLU, R.CANALI, J. BOUWMAN, M.LAVILLE |
| WORK PACKAGE | WP 5 |
| TASK | TASK 5.1 |
| WP LEADER | M.LAVILLE |
| BENEFICIARIES CONTRIBUTING TO THE DELIVERABLE | CRNH, TNO, UHASSELT, UCPH, MCD, FEM, ULG, BIOCC, CRA-NUT |
| STATUS – VERSION | FINAL - VERSION 1.0 |
| DELIVERY DATE (MONTH) | M24 |
| SUBMISSION DATE | M30 |
| DISSEMINATION LEVEL – SECURITY* | PU |
| DELIVERABLE TYPE** | O |

* Security: PU – *Public*; PP – *Restricted to other programme participants (including JPI Services)*: RE – *Restricted to a group specified by the consortium (including JPI Services)*; CO – *Confidential, only for members of the consortium (including JPI Services)*

** Type: R – *Report*; P – *Prototype*; D – *Demonstrator*; - O - *Other*

## CONTENTS

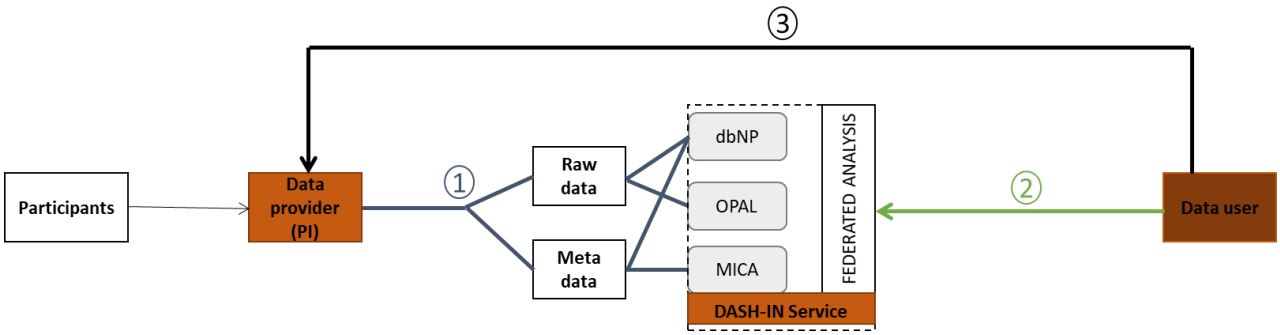
## 1- INTRODUCTION

ENPADASI aims to develop a database (WP3) based on the reuse of data obtained from different nutritional & clinical studies in Europe. However, the sharing of raw, meta or aggregated data raises several regulatory questions. Therefore, the scope of the WP5 is to define general rules required to share and reuse data in accordance with legal and ethical aspects of the EU participating countries and with respect to national policies. Four different topics are concerned in the WP5: ethics, data protection, data sharing polices and intellectual property.

Therefore, the deliverables of work package 5 aim to provide a set of rules and tools applicable to secondary use of nutritional data in the framework of ENPADASI project. Indeed, all the data users/providers have an obligation to operate in conformity with the requirements of their institution, and fulfill all necessary regulatory and ethical requirements imposed by their own national legislation.

Through several conference calls and one meeting in Paris, the consortium of the WP5, which gathers together several ethical & data protection experts, has identified the main ethical and data protection requirements to data sharing and has also proposed solutions and tools in order to help future data providers and users of the ENPADASI infrastructure to share and reuse their data in accordance with the current legislation.

## 2- DATA FLOW WITHIN THE ENPADASI INFRASTRUCTURE AND IDENTIFICATION OF THE DATA PROTECTION & ETHICAL ISSUES




**Overview of the database architecture/data flow**

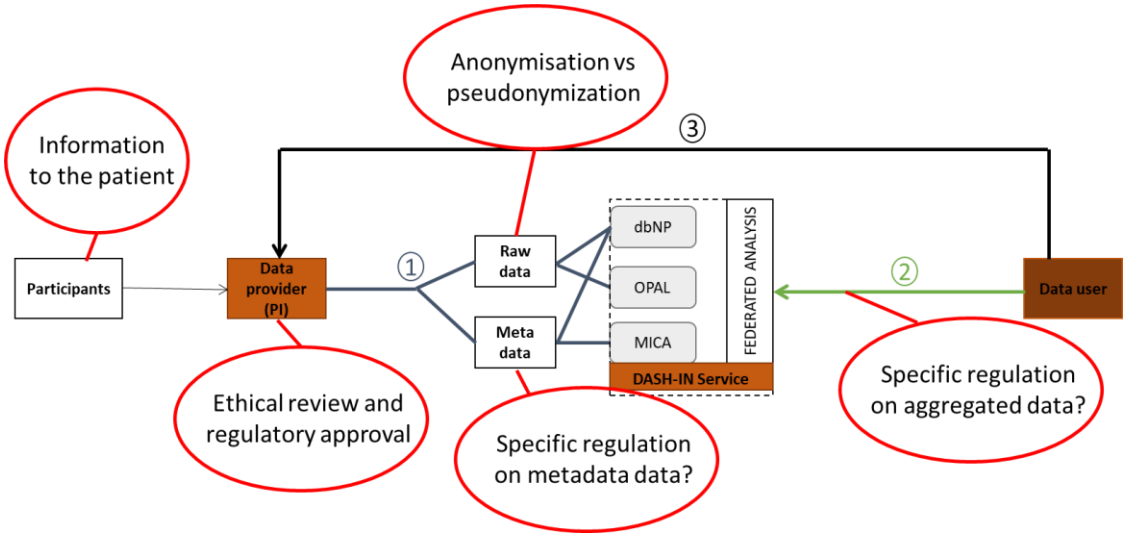
1. It will be mandatory for the Data provider (see for definitions below in the definitions paragraph) to upload his metadata either in the MICA server or in the Phenotype database (www.dbnp.org). The data provider also has the possibility to upload his raw data on the Phenotype database or in an OPAL service. The metadata will be accessible within the ENPADASI consortium, thus each partner will be able to see which kind of study/data could be re-used. The data provider has also the possibility to upload the raw data of clinical studies (both interventional and observational) on the Phenotype database or the OPAL system. But contrary to the metadata, the raw data will not necessarily be

disclosed to the ENPADASI consortium. Each data provider will have the choice to restrict (even to only one data user) or open access to its raw data.

2. The data user can query the MICA server to identify studies of interest according to his scientific hypothesis. Thanks to the FEDERATED ANALYSIS tool, it will be possible to conduct a preliminary statistical analysis in order to validate or disprove the scientific hypothesis. The results of these statistical analyses are named aggregated data. The use of the FEDERATED analysis tool solves several ethical and data protection issues, allowing the future data user to combine the data and obtain statistical results without access to raw data, encountering ethics-related data-sharing concerns.

3. However, in order to go further and to publish the research from the combined analysis, access to the raw data will be mandatory. Two possibilities: the data user will contact the data provider directly, or the data user can have an access by the dbNP or OPAL server. The raw data access will raise several ethical and data protection issues that should be solved before use.

## 3-ETHICAL AND DATA PROTECTION ISSUES RAISED BY THE DIFFERENT DATA SHARING




The scheme above sets out the main ethical and data protection issues facing the future data users and providers and to which we will respond in the present document.

### 3.1) METADATA SHARING AND ETHICAL & DATA PROTECTION CONCERNS

Metadata (see for definitions below) is defined as data that describes other data such as descriptive data on the clinical study: title, purpose, type of population, design of the study, funding, duration of the study, place of investigation, period for the recruitment, informed consent, scientific publications… and also on the raw data collected (help to describe the data collection): type of analysis, type of technical measurements, characteristic of the recruited population (BMI, specific pathology, exclusion/inclusion criteria…)

→ Usually, these types of metadata can be shared and used without any ethical and data protection restrictions since the data does not directly concern the volunteers. However, the future data providers must ensure that by overlapping metadata it does not become possible to identify a volunteer, this may be the case for example with rare diseases. Therefore, the data providers have to ensure that the metadata is totally anonymous. Although the metadata is not in the scope of data protection and ethical regulation, it is generated by the consortium of the project and thus must meet some legal requirements (see the D5.3) in order to share and access them in a legal manner.

### 3.2) RAW DATA SHARING AND ETHICAL & DATA PROTECTION CONCERNS

Personal data and sensitive data (including all the health data) are the two types of raw data mainly concerned by the ethical and data protection regulation.

The EU Data Protection Directive 95/46/EC, which will shortly be replaced by the General Data Protection Regulation (which will enter into force in May 2018), which provides the European regulation regarding the processing of all personal data collected from an EU citizen. Both in the Directive and the Regulation, the main data protection and ethical rule in which each European must adhere to is: **data sharing may only be considered permissible if data is unlinked anonymized or the data subject has given specific consent for the use of their (personal) data for the intended use.**

The notions of anonymized data, information to the patients and the initial purpose of the study are crucial and detailed in the present document.

### 3.3) AGGREGATED DATA SHARING AND ETHICAL & DATA PROTECTION CONCERNS

Thanks to the FEDERATED analysis tool developed in the WP3, the future data users will have access to aggregated data without any access to the raw data stored in the different tools by the data providers. The processing (access, storage, sharing) of aggregated data is not controlled by any data protection and ethical regulations.

## 4- ETHICS

### 4.1) ANONYMIZATION AND PSEUDONYMIZATION

Anonymization of data both concern the ethics and the data protection that are tightly linked. Within the different ENPADASI tools (dbNP and OPAL), it will be possible to share raw data either fully anonymized or pseudonymized. Anonymized data can be processed without any data protection constraints. Pseudonymized data can only be processed if the data subject (participant to the clinical study) has given his/her consent for the purpose of the processing and the processing is compliant with an applicable legal authorization. However, the anonymization and pseudonymization processes meet specific definitions. Each future data provider has to carefully apply this definition to be sure that his data comply with the anonymization and pseudonymization standards. The following parts detail the definition and the basic principles of the anonymization and pseudonymization. However,

for all matters/advices/help, it is strongly recommended to the future data providers to refer to its own national data protection agency (see deliverable 5.2).

### 4.1.1 DEFINITION

The principles of data protection (EU 2016/679) should apply to any information concerning an identified or identifiable natural person and personal data pseudonymized should be considered as information on an identifiable natural person. Therefore, data protection principles should be applied to pseudonymized data. On the contrary, the principles of data protection cannot be applied to anonymous information which does not relate to an identified or identifiable natural person. A specific pitfall is to consider pseudonymised data comparable to anonymised data. (article 29 data protection working party 0829/14/EN WP216).

Anonymisation is a technique applied to personal data in order to achieve irreversible de-identification. Pseudonymization is not a method of anonymisation, it simply reduces the possibility to link a dataset with the data subject. Different practices and techniques exist with variable degrees of robustness that guarantee the capacity to prevent the three major risks of failure of the anonymisation process i.e. 1) Singling out or identity disclosure (singling out of an individual within a data set) 2) Linkability or attribute disclosure (the unveiling of sensitive information of an individual, e.g., having a specific disease, without performing a singling out) 3) Inference or membership disclosure, that is the possibility to deduce, with significant probability, the value of an attribute from the values of a set of other attributes or the ability to determine whether an individual is contained in a dataset.

There are two main different approaches to anonymised data: the first is based on randomization while the second is based on generalization. Randomization is a family of techniques that modify the veracity of the data in order to remove the strong link between the data and the individual. These techniques that include noise addition, permutation or differential privacy, make the data sufficiently undefined so that they cannot be longer related to a specific individual. Generalization is the second family of anonymisation methods. This approach consists of generalizing, or diluting, the attributes of data subjects by modifying the respective scale or order of magnitude. To reduce the reidentification risk, aggregation and K-anonymity technique is applied. L-diversity extends k-anonymity to ensure that deterministic inference attacks are no longer possible. Finally, T-closeness is a refinement of l-diversity.

Pseudonymisation consists of replacing one attribute (typically a unique attribute) in a record by another, it can be done in a traceable way; the natural person is therefore still likely to be identified indirectly. The most used pseudonymisation techniques are as follows: Encryption with secret key, Hash function, Keyed-hash function with stored key and Tokenization.

The table below provides an overview of the strengths and weakness of the techniques considered in terms of the three basic requirements (article 29 data protection working party 0829/14/EN WP216 )

| | Is Singling out still a risk? | Is Linkability still a risk? | Is Inference still a risk? |
|---|---|---|---|
| Pseudonymisation | Yes | Yes | Yes |
| Noise addition | Yes | May not | May not |
| Substitution | Yes | Yes | May not |
| Aggregation or K-anonymity | No | Yes | Yes |
| L-diversity | No | Yes | May not |
| Differential privacy | May not | May not | May not |
| Hashing/Tokenization | Yes | Yes | May not |

Table 6. Strengths and Weaknesses of the Techniques Considered

As shown in the table, the techniques described herein may not provide anonymisation by itself and should always be combined; When used alone, pseudonymisation does not reduce the three main risks of identification of data subjects and will not result in an anonymous dataset. The optimal solution should be decided on a case-by-case basis possibly by using a combination of different techniques (Opinion 05/2014 on anonymization techniques, 10/04/2014 (the working party on the protection of individuals with regard to the processing of personal data set up by directive 95/46/ec of the European parliament and of the council of 24 October 1995).

### 4.1.2 HOW TO CHOOSE BETWEEN ANONYMIZATION AND PSEUDONYMIZATION

Data anonymisation is an important measure to protect personal data but the application of the different techniques can significantly reduce the expressiveness, utility and quality of data. To attenuate these problems a number of complex algorithms have been proposed, which aim at increasing data quality or improving efficiency. However, it is difficult to decide which algorithm is the best suited to specific requirements (Prasser, F. Kohlmayer F. Kuhn, KK. A Benchmark of Globally-Optimal Anonymisation Methods for Biomedical Data. Proceedings of the 2014 IEEE 27th International Symposium on Computer-Based Medical Systems; page 66-71). Moreover, anonymisation does not allow any feedback on health information to volunteers or communication of incidental findings to the patients and deprive them of the possibilities to use their right to withdraw their consent, considering that what is anonymous today can be re-identifiable tomorrow. (Bahr A. and Schlunder I. Code of practice on secondary use of medical data in European scientific research projects, International Data Privacy Law, 2015)

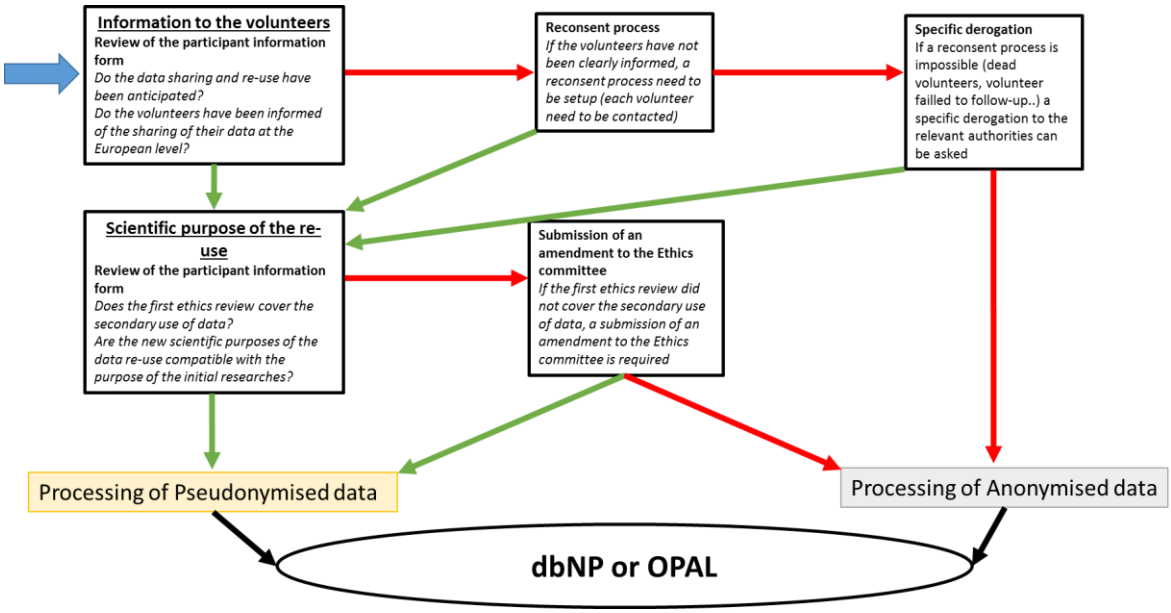
Even if the use of anonymised data requires no data protection constrain, certain types of research cannot undergo to a process of anonymisation. The main mechanism for accommodating such requirements is pseudonymisation that retain the possibilities to trace back to an individual but significantly reduce the risks associated with data processing, while maintains the data expressiveness and utility. Such pseudonymised data may or may not be considered as personal data depending on the circumstances and the member state – in particular, third parties without access to the key code may or may not be able to treat the data as anonymised depending on the test applied (Table 1). (A critique of the regulation of data science in healthcare research in the European Union John M. M. Rumbold and Barbara K. Pierscionek Rumbold and Pierscionek BMC Medical Ethics (2017) 18:27 )

The processing and secondary use of pseudonymised personal data can be anticipated in the informed consent procedure. In the informed consent process, participants can be informed about privacy risks during data-mining/sharing and stimulated to make explicit decisions about data sharing. In this respect, the participant information should contain an accurate description of the methods used for the pseudonymization and for the maintaining of their privacy protection.

If adequate provisions have been made to protect and maintain subject/patient anonymity and authorization has been asked on data sharing and informed consent has been signed by the participant, the data can be considered as eligible for data mining or input into larger data-bases or data sharing. The EU Regulation 2016/679 encourages researchers to adopt codes of conduct that promote pseudonymisation recognizing the ability of pseudonymisation to help protect the rights of individuals while also enabling data utility (https://iapp.org/news/a/top-10-operational-impacts-of-the-gdpr-part-8-pseudonymization/).

But, in cases where the volunteers have not been clearly informed and explicitly consented to data sharing, a reconsent process needs to be set up to share personal data; in case where the process is impossible a specific derogation to the relevant authorities should be asked. Finally, if specific informed consent has not been attained from subjects/patients in on going studies, the responsible investigators are advised to approach the applicable research and ethics committees(REC) in order to seek advice on amending their research and ethics provisions so that data sharing is anticipated for future inclusions.

.




**Decision tree to guide data providers in the required ethical and data protection steps**

## 4.2) INFORM CONSENT

The informed consent process plays an important role in anticipating and arranging data sharing. Depending on how informed consent has been or will be obtained, options for data sharing may vary. To anticipate data sharing in the informed consent process, a generic paragraph on data sharing that can be adopted in the participant information could be helpful. Within ENPADASI, such a generic paragraph was developed. In its design, we pursued full respect for known ethical and legal requirements. However, given the considerable diversity in national legal regulations involved, it is impossible to account for all aspects. Therefore, we emphasize that a positive review by the applicable ethics committee is still required before data collection or subject enrollment can start. The generic paragraph is added in annex.

## 4.3) THE INFORMATION RELATED TO ETHICS NEEDED TO BE IMPLEMENTED ON THE DASH-IN SERVICE

We provide a list of minimum information related to ethical issues covering the following domains: legal, data and intellectual property. The list was developed based on the templates circulated to identify both observational and intervention/mechanistic studies, coupled with the in-depth discussions regarding legal advice on issues related to data protection and ethics during the TCs held within the Work Package 5.

The first domain requests information related to the signed informed consents, Ethics Committee approval, potential limitations regarding secondary use of data and Data Transfer Agreements.

The second domain requests information related to the willingness of sharing data and metadata. In addition, the list contains two questions related to anonymization or pseudnonymization of the data for those partners willing to share raw data.

| Label | Unit | Comments |
|---|---|---|
| **Legal** | | |
| Did individuals providing data sign informed consent? | Yes/No | |
| Does the informed consent cover the secondary use of the data? | Yes/No | |
| Was the study approved by an Ethics Committee? | Yes/No | |
| Does the ethical review cover the secondary use of the data? | Yes/No | |

**Checklist of minimum information for ethics in the DASH-IN research infrastructure**

# 5- DEFINITIONS

**Anonymous/ised data:** Information which does not relate to an identified or identifiable natural person and personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable, including through cross-analysis by overlapping data.

**Data Access Committee (DAC):** Integral component of the DASH-IN Service for managing access to the data. The DAC is responsible for reviewing, approving or disapproving applications from potential users for a variety of restricted access cases.

**Data Provider:** The Provider (or 'Data Provider') is the individual researcher or investigator or body of researchers or investigators that makes data available for access and use within the context of the ENPADASI consortium and database. (It does not refer to the research participants.). The data provider should be the legal person or body that is responsible (owns) the data.

**Data User:** The 'Data User' is the individual researcher or investigator or body of researchers or investigators from either academia or industry that requests access to samples and/or data and use through the Data Sharing In Nutrition (DASH-IN) Service. The Data User is a 'data processor' in the meaning of the EU General Data Protection Regulation. The Data User may seek access outside of the context of the DASH-IN Service environment.

**Ethics Committee:** The term 'ethics committee' in this document refers to a committee which has given ethics approval for a study which has/intends to collect and use health data that will be subsequently made available by the Data Provider within the database and the DASH-IN Service. (It does not refer to the ENPADASI Data Access Committee.)

**Metadata:** Metadata is data describing other data. Metadata summarizes basic information about data, which can make finding and working with particular instances of data easier. Metadata can be created manually, or by automated information processing. Manual creation allow is the user to input any information they feel is relevant or needed to help describe the file, which is very relevant for example for the description of the study design. Automated metadata creation can display information such as file size, file extension, when the file was created, who created the file and can also include the logs of the machine used to generate the data. Metadata are highly aggregated and generalised data. Metadata is most often anonymised data.

**Personal Data:** Any information relating to an identified or identifiable natural person (data subject). An identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.

The use of the term 'personal data' in this document covers sensitive categories of personal information as defined within the EU General Data Protection regulation data such as health data, biological and clinical data and the use of wellbeing data. Such data are particularly protected under privacy rules and secured management and access processes.

**Pseudonymisation:** The processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.

**Raw data:** Raw data refers to any data object that has not undergone thorough processing, either manually or through automated computer software. Raw data may be gathered from various

processes and IT resources. Most digital equipment does not record the raw data but immediately processes it through vendor-defined algorithms into a vendor-specific primary record while discarding the original signals recorded in the equipment. In this context, such primary record files will be seen as analogous to raw data.

Raw data is primarily unstructured or unformatted repository data. It can be in the form of files, visual images, database records or any other digital data. Raw data is extracted, analysed, processed and used by humans or purpose-built software applications to draw conclusions, make projections or extract meaningful information. The processed data takes the form of information. Raw data can include personal data in the meaning of Article 4 of the General Data Protection Regulation of the European Union (Regulation (EU) 2016/679 [1]). In such a case the respect of applicable personal data protection laws will need to be ensured by the data providers and the users of the DASH-IN service. Raw data shall be pseudonymised before any exchange in order to ensure appropriate data protection.

**Sensitive data**: data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, genetic data, data concerning health or data concerning a natural person's sex life or sexual orientation.

## ANNEX: TOOLS AND PRATICAL INFORMATIONS TO SHARE DATA IN THE FRAMEWORK OF ENPADASI

### BROAD INFORMED CONSENT

To anticipate data sharing in the informed consent process, a generic paragraph on data sharing that can be adopted in the participant information could be helpful. Within ENPADASI, such a generic paragraph was developed. In its design, we pursued full respect for known ethical and legal requirements. However, given the considerable diversity in national legal regulations involved, it is impossible to account for all aspects. Therefore, we emphasize that a positive review by the applicable ethics committee is still required before data collection or subject enrollment can start. The generic paragraph sounds as follows:

> *Within this project, we will collect and store personal data. We will make sure that all necessary precautions to protect your privacy are taken. We will collect and store all data respectful of [applicable national law] and the European General Data Protection Regulation. In published reports of the research results, your identity will never be revealed. Your data will be stored a maximum of [specify] after [your last visit/study termination].*
>
> *The collected Data will be [coded/anonymized], which means that [your name is removed from all documents and replaced by a unique code/data cannot be traced back to your person by any*

[1] http://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A32016R0679

*means]. The list with codes is stored securely and only accessible to [specify]/ your name is removed from all documents and data can by no means be traced to your person].*

*[In case of coded/pseunonimized data]: We hereby ask you permission to share the data collected in this study with other research teams across the world conducting research in the same conditions and diseases. Your identity will not be disclosed to other study teams. New projects in which your data might be used will need permission by an ethics committee according to applicable regulations. [If applicable: If your data is transferred to a data processor in a "third country" e.g USA, which do not complied with EU-legalization, we use a contract to ensure compliance with European legislation.]*

*Should you withdraw your consent to study participation or data storage, your personal data will be removed from our records. You can withdraw from the study at any time by [explain procedure, also how this can be done after termination of the study]. [If applicable: However, research results that already have been obtained from the analysis of a dataset including your data, will remain unaltered after your withdrawal.] [If applicable: Should your data already have been anonymized, it is technically no longer possible to delete them from our records].*

*Individuals study results will [/not] be communicated to [you/dr. XX, your GP] [, even when they might be of potential clinical relevance to you]. In the case that your data are used for other research projects during or after termination of this study, individuals study results will [/not] be communicated to [you/dr. XX, your GP] [, even when they might be of potential clinical relevance to you].*

*The law provides that occasionally, at any time during or after the study, staff from the sponsor or their designated representatives, the monitor (only applicable for medicinal studies) and the applicable authorities, will be granted direct access to your entire medical records/source data so that they can confirm that the information collected during the study is accurate. In these circumstances, your identity may be disclosed. However, anyone who has access to this data will be legally bound to keep the information confidential.*

The transfer of data to third countries requires specific attention as specified in Art. 49 of the GDPR.