REal-time data monitoring for **S**hared, **A**daptive, **M**ulti-domain and **P**ersonalised prediction and decision making for **L**ong-term Pulmonary care **E**cosystems

D1.2: Data management plan

Dissemination level:	Public
Document type:	ORDP
Version:	3.0
Date:	26-02-2024

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 965315. This result reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains.

Document Details

Reference No.	965315
Project title	**RE-SAMPLE - RE**al-time data monitoring for **S**hared, **A**daptive, **M**ulti-domain and **P**ersonalised prediction and decision making for **L**ong-term Pulmonary care **E**cosystems
Title of deliverable	Data management plan
Due date deliverable	31-08-2021
Work Package	WP1
Document type	ORDP: Open Research Data Pilot
Dissemination Level	Public
Approved by	Coordinator
Author(s)	Sofya Kopelyan (UT), Serge Autexier (DFKI)
Reviewer(s)	Serge Autexier (DFKI), César Mediavilla Martínez (ATOS), Costas Lambrinoudakis (UPRC), Anke Lenferink (UT), Christiane Grünloh (RRD), Marjolein Brusse-Keizer (MST)
Total No. of pages	22

Partners

Participant No	Participant organisation name (country)	Participant abbreviation
1 (Coordinator)	University of Twente (NL)	UT
2	Foundation Medisch Spectrum Twente (NL)	MST
3	University of Piraeus Research Center (GR)	UPRC
4	Foundation Tartu University Hospital (EE)	TUK
5	Foundation University Polyclinic Agostino Gemelli IRCCS (IT)	GEM
6	European Hospital and Healthcare Federation (BE)	HOPE
7	German Research Center for Artificial Intelligence GMBH (DE)	DFKI
8	ATOS IT Solutions and Services Iberia SL (ES)	ATOS
9	Roessingh Research and Development BV (NL)	RRD
10	Innovation Sprint (BE)	iSPRINT

Abstract

This document reports the third version of the data management plan for the RE-SAMPLE project. It analyses the main aspects of data management policy with the help of the template provided by the Horizon 2020 Online manual. Accordingly, it presents a summary of the data collected and generated in RE-SAMPLE, explores how the data can be made FAIR (Findable, Accessible, Interoperable, and Re-usable), evaluates the allocation of resources and data security, and addresses ethical and other issues. This data management plan will be updated one more time before the end of the project on the basis of latest developments.

Corrections

v2.0 Updated the following sections:

- Abstract
- 1. Introduction (the last paragraph)
- 4. FAIR data (the opening paragraph, 4.2.3, 4.2.4, 4.3.3, 4.3.4, 4.4.3, and 4.4.4).

v3.0 Updated the following sections:

- Abstract
- 1. Introduction (the first and last paragraphs)
- 3. Data summary (data collection purposes, data types and storage)
- 4. FAIR data (4.1.4, 4.2.4, 4.3.4)
- 5. Allocation of resources, 6. Data security, and 8. Other issues (the storing of blood samples is no longer applicable)
- 7. Ethical aspects (deliverables' numbers)

Contents

List of Figures

Definitions, abbreviations, and acronyms

API	Application Programming Interface
CCC	Complex Chronic Condition
CERN	Conseil Européen pour la Recherche Nucléaire (European Council for Nuclear Research)
COPD	Chronic Obstructive Pulmonary Disease
CORDIS	Community Research and Development Information Service
D	Deliverable
DANS	Data Archiving and Networked Services
Data repository	Online archive for data
DCMI	Dublin Core Metadata Initiative
DMP	Data Management Plan
DOI	Digital Object Identifier
DPO	Data Protection Officer
EHR	Electronic Health Record
EU	European Union
FAIR	Findable, Accessible, Interoperable, Reusable
FHIR HL7	Fast Healthcare Interoperability Resources Health Level 7
GDPR	General Data Protection Regulation
H2020	Horizon 2020
IPR	Intellectual Property Rights
ISBN	International Standard Book Number
M	Month
Metadata	"Data about data" – structured data describing other data
OpenAIRE	Open Access Infrastructure for Research in Europe
OpenDOAR	Directory of Open Access Repositories
ORDP	Open Research Data Pilot
PDF	Portable Document Format
PMO	Project Management Office
REST API	Representational State Transfer Application Program Interface
ROAR	Registry of Open Access Repositories
RWD	Real-World Data
T	Task
V	Version
VCP	Virtual Companionship Programme
WP	Work Package

1. Introduction

RE-SAMPLE participates in the Open Research Data Pilot (ORDP) in Horizon 2020 (H2020) and aims to provide open access to the scientific results, through publications, to the data that will be used during the research, as well as to the newly generated real-world data (RWD). The project is designed with open access and open science 'by default', wherefore it follows three principal regulations and initiatives: 1) free flow of non-personal data in the European Union (European Union, 2018); 2) FAIR Guiding Principles for scientific data management (European Commission, Guidelines on FAIR Data Management in Horizon 2020 (v.3, 26 July 2016), 2016)[1]; and 3) European Open Science Cloud EU Node[2]. In accordance with the conditions of ORDP, RE-SAMPLE has developed this data management plan (DMP) and intends to keep it up-to-date. We will also consider providing access to data and metadata needed to validate results in scientific publications and depositing the project's data that does not fall under privacy or IPR protection in a research data repository to ensure third parties can freely access, mine, exploit, reproduce and disseminate them.

This DMP describes the data management life cycle for the data to be collected, processed, generated, preserved, and re-used both during and after the end of the RE-SAMPLE project. During the project, the focus of data management will be on complying with the General Data Protection Regulation (GDPR (European Union, 2016)) for privacy-abiding treatment of health data. Towards the end of the project, priority will be given to making the data as openly accessible as possible, but at the same time as closed as strictly necessary. For example, if the privacy of the participants is at risk, or the data pertains to those project outcomes that fall under secrecy clauses in the consortium agreement or in the exploitation agreements, the dataset might be stored under a restricted license or kept completely closed.

The DMP is an "umbrella" document created for all overarching data management matters which are applicable to the whole project. It is prepared based on the template provided by H2020 Online manual[3]. Parts of this document are based on the Best Practice document developed by the Project Management Office (PMO) at the coordinating partner, University of Twente (UT). Through the use of the Best Practice document, the PMO aims to maintain the highest quality of all processes and matters pertaining to data management.

While being one of the outcomes of Task T1.5 "Data and ethics management" in Work Package (WP) 1, the DMP is closely connected to and complemented by the work on security measures for organisational, legal, and technical security and privacy requirements in WP4 and on protection of personal data in WP9.

This DMP is a "living" document that will be updated during the course of the project to include renewed insights in the data management. Version 1.0 of this DMP was released at an earlier stage of the project (Month 6). Version 2.0 was released in August 2022 (M18). The current version 3.0 provides an update for the end of the second reporting period (M36). However, some of the information it contains could still be amended, and we will continue appending new data to the DMP as soon as they are available. We foresee one last update before the end of the project with regard to the final project developments.

[1] https://www.go-fair.org/fair-principles.

[2] https://open-science-cloud.ec.europa.eu.

[3] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm#A1-template.

2. Objectives

The objectives of this DMP are to:

1) provide guidelines on data management to all project partners; and

2) create awareness on the FAIR principles to help the partners plan for the FAIR use of data at the earlier stages of the project.

In what follows, this deliverable describes the types of data that will be collected and generated in the RE-SAMPLE project, the adherence to the FAIR standards, the allocation of resources, data security, and ethical and other aspects of data management.

3. Data summary

This DMP encompasses both quantitative and qualitative data resulting from the RWD cohort, model validation, and prototyping and evaluation of the virtual companionship programme (VCP). Data will be gathered through sensor devices, diaries (questionnaires), interviews, and workshops.

There are three purposes of data collection in relation to the objectives of the RE-SAMPLE project:

1) Retrospective data collection to gain detailed insight into possible important predictors and parameters from retrospective data for patients with Chronic Obstructive Pulmonary Disease (COPD) only and with COPD and complex chronic conditions (CCC) in order to identify COPD and CCC progression and multimorbid CCC exacerbations;
2) User Requirements Elicitation and end-user testing for the VCP;
3) Prospective data collection to design, implement, and evaluate the VCP in order to support patients with COPD and CCC and healthcare professionals in proactive and tailor-made care that reduces the societal and economic burden of these CCC.

The following data types will be collected for purpose 1 and 3: identification / contact data; retrospective and prospective demographic and clinical data; quantitative and qualitative data related to user requirements and continuous end-user involvement; retrospective RWD (e.g., environmental data) and prospective RWD (sensor data and diaries). Additionally, metadata on project outputs (research publications, deliverables, datasets, etc.) will be created during working on RE-SAMPLE. Finally, textual and graphical data such as deliverables, publications, presentations, logo’s, etc., will be produced while delivering this project.

The existing data that will be re-used is specified in D2.3 *Overview rapport on the contents and quality of existing RWD, EHR and environmental data.* The re-used data comprises Electronic Health Records (EHR), RWD, and environmental data. These data will be selected from various sources (e.g., data from previous clinical studies stored in databases at the RE-SAMPLE clinical sites, environmental data from national or regional institutions) following a strict validation process in terms of data quality and completeness. As of yet, the size of the prospective dataset (prospective RE-SAMPLE cohort in WP5) is 175 (GEM: n=84, MST n=48, TUK n=43). In WP6, for the patient clustering analyses, we used a retrospective dataset of n=1980 patients. The resulting data might be useful, for example, to researchers working in the field of personalised e-Health technology and machine learning engineers, to healthcare professionals for more accurate predictions and treatment of COPD with CCC, and to patients with COPD and CCC actively engaged in the management of their health condition.

For purpose 2, we collected qualitative and quantitative data, e.g., audio recordings during workshops, interviews and end-user tests; and questionnaires. For studies carried out by RRD, the data is stored locally. In sum, we store 11.7 GB raw data for WP2 (i.e., user requirements and service model), and 9.73 GB raw data for WP5 (end-user tests and extra studies carried out by students at RRD). For studies carried out by pilot sites by themselves (e.g., interviews, user tests), the data is stored at the respective pilot site (GEM, TUK, MST).

Open access to research data is a rule in article 29.3 of the Grant Agreement. In making the decisions about providing open access to the data, the consortium will follow the open access decision graph from the “Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020 (v.3.2, 21 March 2017)” (European Commission, 2017) (Figure 1).

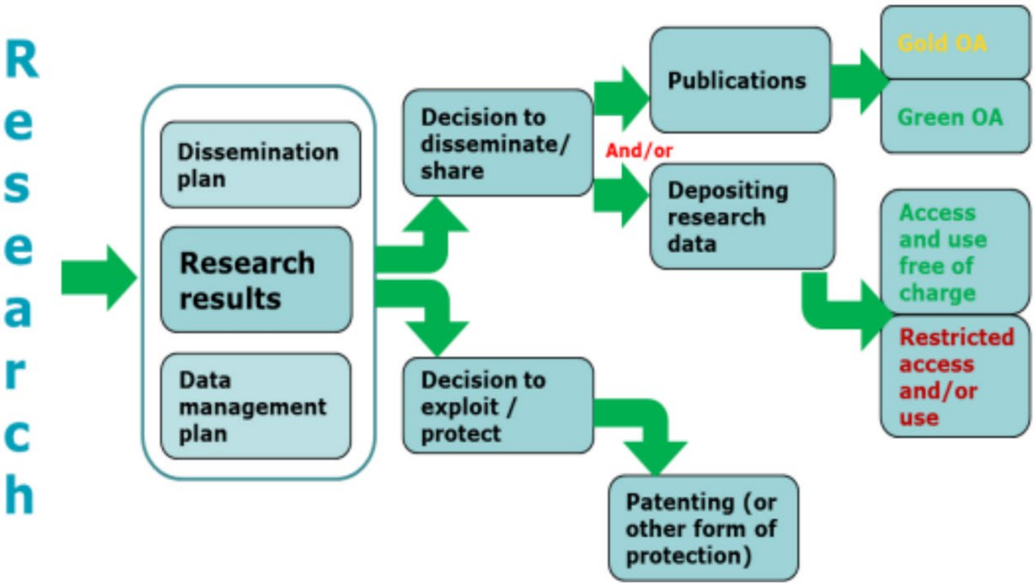
Research
Dissemination plan
Research results
Data management plan
Decision to disseminate/ share
Decision to exploit / protect
And/or
Publications
Depositing research data
Patenting (or other form of protection)
Gold OA
Green OA
Access and use free of charge
Restricted access and/or use

Figure 1: Open access decision graph

In terms of data storage, the RE-SAMPLE project uses Microsoft Teams and SharePoint for internal documentation, confidential reports, presentations, graphic materials, etc. Public deliverables and open access publications are published and stored on the project's website https://www.re-sample.eu. GitLab – an online tool that provides a repository manager for software development and IT operations and is hosted securely by the University of Twente (UT) – is used to store software. Research data is stored in the three hospitals that participate in the project: Foundation Medisch Spectrum Twente (MST), Foundation Tartu University Hospital (TUK), and Foundation University Polyclinic Agostino Gemelli IRCCS (GEM), as well as in Healthentia, an eClinical platform that enables the collection of RWD through a mobile application, provided by Innovation Sprint (iSPRINT); and at Roessingh Research and Development (RRD).

In what follows, this DMP will concentrate on the four types of data generated in the project that can be shared open access:

- research publications,
- project deliverables,
- software, and
- research datasets and predictive models.

4. FAIR data

RE-SAMPLE is committed to allow the reuse of research data according to the FAIR principles to make the data findable, accessible, interoperable, and reusable (Wilkinson, et al., 2016). The project applies this to all four types of research results produced or generated during the runtime of the project, and in the following, it details the measures already setup or to be considered for making data findable, accessible, interoperable, and reusable depending on whether the data are publishable or not. Furthermore, we will consider making connection to one of the GO FAIR Implementation Networks[4] (e.g., to Personal Health Train).

4.1 Making data findable, including provisions for metadata

4.1.1. Research Publications

Consortium members are encouraged to publish in outlets that have structured metadata and issue a DOI (Digital Object Identifier). In case structured metadata is not available, it is recommended to consult DataCite Metadata Schema for the Publication and Citation of Research Data[5]. Authors will employ appropriate search keywords during the publication process. These may include keywords like AI or Artificial Intelligence, digital health or eHealth, personalised care or personalised medicine, RWD or real-world data, COPD or Chronic Obstructive Pulmonary Disease, and so on.

The authors must provide the disseminating partner (HOPE) and the Coordinator with the following metadata per publication:

- DOI
- Type of publication
- Repository Link
- Link to publication
- Title
- Authors
- Title of the Journal/Proceedings/Books series/Book (for book chapters)
- Number, date, or frequency of the Journal/Proceedings/Book
- Relevant Pages
- ISBN (International Standard Book Number)
- Publisher
- Place of publication
- Year of publication
- Availability in Open Access (Gold /Green)
- If Gold, the costs should be stated.
- Peer reviewed publication (yes/no)
- Joint public/private publication (yes/no)

The bibliographic metadata in the repositories should additionally include the following:

- The terms “European Union (EU)” and “Horizon 2020”,
- The name of the action, acronym, and grant number,
- The publication date and length of the embargo period, if applicable.

4.1.2. Project Deliverables

The metadata for project deliverables routinely include grant number, project title, the title of deliverable, the due date of the deliverable, which work package it pertains to, the type of the document, its dissemination level, approving authority, internal authors’ and reviewers’ names, and the total number of pages. Each deliverable also contains an abstract to allow potential readers to gain awareness of the contents and concepts and evaluate their suitability for their use.

[4] https://www.go-fair.org/implementation-networks/overview

[5] https://schema.datacite.org/meta/kernel-4.1/doc/DataCite-MetadataKernel_v4.1.pdf

Versioning has been outlined in D1.1 *Quality, risk, and IPR management plan* (Kopelyan, Garel, Kyriazakos, & Gonzalez, 2021). To reiterate, while in draft, all documents' titles follow the same format:

<Project Name> <Number> <Title> <Version> <Participant>.

Example: RE-SAMPLE D1.2 Data management plan v0.1 SK.

When finalised, the format changes to:

<Project Name> <Number><Title><Version>Final.

Example: RE-SAMPLE D1.2 Data management plan v1.0 Final.

All documents start with version 0.1 and increment to 0.2 only on the authority of the author. All other participants who edit the deliverable increment the version to the next available number behind the version number, e.g.: v0.1.1 Deliverables submitted to the European Commission (EC) for the first time have version v1.0. In short, a version is numbered X.Y.Z, where X can be edited by the coordinator, Y by the author, and Z by anyone contributing.

4.1.3. Software

All software artefacts developed in RE-SAMPLE are going to be maintained in the GitLab offered by the coordinating institution (UT) or be part of the infrastructures of the Healthentia system. This provides a central place to find all necessary artefacts or connectors and Application Program Interfaces (APIs) to components in the Healthentia system. The GitLab system provides a version control system that allows to version each individual artefact.

4.1.4. Datasets and Predictive Models

All retrospective and prospective datasets provided by the clinical partners will only be accessible to the data processing partners (Data Processors), following the rules and conditions defined in the respective data processing agreements. Retrospective datasets will be identified through the pilot site partner name and the name of that dataset according to its internal naming. No version control is necessary for these datasets, and keywords available for these datasets will be reused. Prospective datasets are also identified through the pilot sites partner names (GEM, MST, TUK) and the entire prospective data consisting of all pilot site prospective datasets will be named RE-SAMPLE. All four will have a suffix indicating if the unencrypted (PLAIN) or encrypted (ENC) version is referred to, and a version according to a versioning scheme that will be defined in order to align datasets and predictive models.

Predictive models trained in RE-SAMPLE are considered as research data. For predictive models trained on the available datasets, the naming scheme will consist of the name of the dataset specific target variables and predictor variable configuration. In addition, they will contain the suffix if they are encrypted or unencrypted (as for datasets) and a version according to the to-be-defined version scheme aligning datasets and predictive models.

4.2 Making data openly accessible

4.2.1 Research Publications

As has already been indicated in D1.1 *Quality, risk, and IPR management plan*, Consortium members are encouraged to publish in repositories contained/registered in the Registry of Open Access Repositories (ROAR[6]) and/or in the global Directory of Open Access Repositories (OpenDOAR[7]). They are likewise advised to target conferences and journals that have publications listed in standard citation databases like

[6] http://roar.eprints.org
[7] https://v2.sherpa.ac.uk/opendoar

Scopus, PubMed, or the DBLP Computer Science Bibliography. It is recommended to publish preprints in open access repositories such as Zenodo[8] and arXiv[9]. A complete list of publications with links to full text will be available on the project's webpage[10].

All publications should be accessible with free or standard software (Microsoft Office, Adobe Acrobat, etc.).

4.2.2 Project Deliverables

RE-SAMPLE public deliverables are made openly available through the project's website[11] and through CORDIS (Community Research and Development Information Service), the European Commission's public repository and portal to all EU-funded research projects results[12]. All deliverables are accessible with Adobe Acrobat software.

4.2.3 Software

Although the Architecture of the RE-SAMPLE platform has been defined and all its components have been identified (D2.6, M12), partners involved in developing these components are still engaged in technical discussions on various details. Therefore, the list of components that shall be made available open source as of M36 could still be changed. When asked to identify if, and possibly under which license models, they plan to make their contributions accessible, the partners have reported the following components:

- Health Data Hub (ATOS) - Apache License, Version 2
- Implementation Guide (ATOS) - Copyright
- Data Manager (DFKI) - Apache License, Version 2
- Training Manager (DFKI) - Apache License, Version 2
- Model Manager (DFKI) - Apache License, Version 2
- Results Manager (DFKI) - Apache License, Version 2
- Environmental Data Manager (DFKI) - Apache License, Version 2
- Orchestrator Federated Learning Coordination (DFKI) - Apache License, Version 2
- Local Data Connector (iSPRINT) - MIT license
- A prototype (proof of concept) that enables the training and inference of medical machine learning models under homomorphic encryption (UT) - MIT license

For those components that shall be eventually made available open source, a publication on some open and durable GitLab[13] repository may be considered. It will be linked from the project website. For sustainable open access, a publication and linking from the Zenodo online repository created through the European Commission's OpenAIRE programme and hosted at CERN is an option to be agreed on by the partners. This equally applies to RE-SAMPLE software components that may be released in binary format, to be reusable by third parties.

4.2.4 Datasets and Predictive Models

The data sets generated in RE-SAMPLE, combining retrospective data and prospective data collected by GEM, MST, and TUK directly or through Healthentia, could be useful to the research community and more specifically to people conducting medical scientific research. To this end, the procedure adopted by RE-SAMPLE in order to provide Open Research Data (ORD), is the following:

[8] https://zenodo.org

[9] https://arxiv.org

[10] https://www.re-sample.eu/resources/publications

[11] https://www.re-sample.eu/resources/deliverables

[12] https://cordis.europa.eu/project/id/965315/results

[13] https://about.gitlab.com

- Patients are requested to grant consent for the anonymisation of their data and the subsequent use of their anonymised data for specific purposes of processing (e.g., medical-scientific research), with clear information provided to them regarding these purposes;
- Hospitals are responsible for anonymising the data;
- The (future) use of the anonymised data should be linked to specific purposes of processing;
- Interested third parties are required to request access to and use of anonymised data from the Data Controllers (Hospitals), outlining the processing purpose for which they need the particular dataset;
- Data controllers evaluate each request, guided by an "Access Policy of ORD", primarily to confirm its conformity with the initial purpose of processing for which anonymisation was carried out. Upon acceptance of the request, access to the ORD is granted to the interested third party.

The previously mentioned "Access Policy of ORD" for evaluating and prioritising access requests from external entities, along with the criteria for granting or denying access to the ORD, should be noted as currently under development.

Regarding Open Access to RE-SAMPLE's prediction models, the procedure adopted by the consortium is the following:

- Models trained with real data will not be made available as Open Access;
- Upon request, models trained with anonymised data (the data sets offered as ORD), will be offered to third parties following the same Access Policy that was used for the ORD;
- Models trained with data sets that are statistically related to the actual set (e.g., generated data) and lack any connection to the actual data will be publicly open.

Finally, the project will explore the possibility to generate synthetic datasets that have comparable statistical properties as the original datasets, and if permission can be obtained from the pilot site partners. If that turns to be viable, these synthetic datasets could be published open access, for example, publication in a research data archive like Zenodo will be discussed. Furthermore, datasets may be stored in partners' repositories, and national trusted repositories with a Data Seal of Approval, such as DANS (Data Archiving and Networked Services[14]) in the Netherlands. These synthetic datasets may be also used to train non-encrypted predictive models, which, if they have adequate predictive quality, may also be provided with open access. This will have to be discussed then, and the data management plan will be refined accordingly.

4.3 Making data interoperable

4.3.1 Research Publications

To enable inter-disciplinary and international knowledge exchange, all publications will be published in English and made available in PDF format. Vocabularies and keywords used by the authors will depend on the subject matter, target audience, and the outlet of the publication. The Dublin Core Metadata Initiative (DCMI[15]) can be considered for best practices in increasing data interoperability[16].

4.3.2 Project Deliverables

Similarly, all public deliverables are produced in English and are made available in PDF format.

4.3.3 Software

The software components of the RE-SAMPLE platform are designed as part of the definition of the architecture. The decision, which of these components or parts of components will be made available, pertains to the partners developing them. Different programming languages will be used to implement the different components, e.g., Python and Java. Thus, as part of the architecture definition process in T2.5 "Architecture and technical specifications for RE-SAMPLE platform", a microservice architecture is

[14] https://dans.knaw.nl/en
[15] https://www.dublincore.org
[16] This applies to all data, not just publications.

foreseen to bridge between the different components, and the interfaces between components will be REST APIs (Representational State Transfer Application Program Interfaces) specified in some interoperable data format specification where sensible and possible, e.g., in JSON format. If legacy software of some partners is included in the RE-SAMPLE framework, corresponding connectors complying with that architecture and APIs will be implemented. As a result, the software components of RE-SAMPLE will be interoperable on an API level, which eases their reuse. The architecture, the components and their REST API specifications are described in D2.6 *Architecture and technical specifications*.

4.3.4 Datasets and Predictive Models

Retrospective datasets will only serve internal purposes of the project and will only be available to the partners designated as Data Processors (through the respective data processing agreements). Prospectively collected datasets will be mapped in a uniform format to allow interoperability in the project to develop predictive models from the data collected at the different pilot sites, as well as to allow to use the models for predictions. That uniform RE-SAMPLE data format is developed in T4.1 “Representation of multi-modal data incl. disease progression monitoring features” and adopted where possible. A FHIR HL7 (Fast Healthcare Interoperability Resources Health Level 7) representation and any necessary extensions will be designed in a FHIR HL7 compliant format. The FHIR HL7 format, being an international standard in the healthcare domain, will ensure a high level of interoperability of the datasets as well as the applicability of the developed predictive models. Moreover, if synthetic datasets may be obtainable from RE-SAMPLE, they will be made available in that FHIR HL7 compliant format for reuse. The final format has been published as part of D4.1 *Representation of Multi-Modal Data and Disease Progression Monitoring Features* (M18).

4.4 Increase data re-use (through clarifying licences)

4.4.1 Research Publications

Open Access publishing typically allows a retention of copyright by authors. Creative Commons[17] are the most frequently used licenses in this case, although some outlets may use custom open access licenses. The choice of the appropriate Creative Commons license is made by the copyright holder and depends on the following criteria:

- Attribution,
- Commercial use,
- Modification,
- Redistribution.

In general, publications will be made available for re-use immediately. In case the publication should be kept confidential for some period, this can be requested by the WP leader and will be discussed in the Project Management Team meeting. Clear reasons should be given why the publication will be kept confidential, and it should be specified how much longer the confidentiality should be put in place, but no longer than ninety (90) days, as stated in the Consortium Agreement. It should be noted that platforms like Zenodo and academic social networks like ResearchGate[18] may support an itemised access control in such case, whereby data owners are able to grant and revoke access to restricted content for particular users.

Open access publications will remain re-usable for an indefinite period of time. Their quality will be assured via peer review and moderation during the publication process.

4.4.2 Project Deliverables

Public deliverables are disseminated by RE-SAMPLE through CORDIS (for an indefinite period of time) and the project website (for the duration of the project and at least four years after its completion) according to the terms and conditions of the Grant Agreement which obligates to disclose all results to the public as

[17] https://creativecommons.org
[18] https://www.researchgate.net

soon as possible. Deliverables are made available for re-use on a royalty-free basis without any restrictions. Their quality is assured via an internal review by two different consortium partners.

4.4.3 Software

The provision of components for re-use depends on the exploitation plans of the individual partners. Although the architecture of RE-SAMPLE has been designed, the final list of components of the RE-SAMPLE framework is not completely fixed. See section 4.2.3 for information under which licenses the partners may make their components available.

4.4.4 Datasets and Predictive Models

Datasets and models developed in the project will be stored in appropriate and secured places by the partners (see also Section 6) to ensure good scientific practice if they have been used for scientific publications. They will be versioned according to the scheme described in Section 4.1.4 and kept up to 10 years. If during the project it is decided to produce synthetic data and predictive models based on these, or publish prospective datasets, then licensing and re-usability rules will be defined and agreed on among the relevant partners, and the data management plan will be refined accordingly.

5. Allocation of resources

Principal Investigators of each research group are responsible for the data management at their organisation. Work Package leaders are responsible for the data management of respective WP tasks. Decisions on research data storage will be taken by the General Assembly. The RE-SAMPLE Project Management Team is responsible for overall data management and for monitoring the implementation of this DMP.

Public deliverables are published at no cost. The costs of open access publications and open research data must be covered by partner organisations. The project will only use free repositories for potential publication and long-term preservation of synthetic data and predictive models. Non-synthetic datasets and models, and also software artefacts will be stored by partner organisations at no charge.

6. Data security

The partners are responsible for the protection of the data collected, processed, and stored within their organisations. All partners having access to personal data will comply with GDPR. Among other things, RE-SAMPLE will strictly follow privacy- and security-by-design principles and use highly innovative secure multiparty computation techniques to satisfy all security and privacy-related requirements imposed by the legal framework and ethical considerations of its users.

Furthermore, a security and data protection policy has been developed in T4.5 “Security & privacy measures, security & data protection policies”, including important topics such as organisational measures for data protection, procedures for acquiring / withdrawing user consent, satisfying users’ right (e.g. the right to be forgotten), data portability, use of portable devices, managing security and privacy incidents, managing log files, and internal audits.

Project management datasets which contain no personal data will be stored in the Microsoft Teams/SharePoint environment provided by the UT. They are stored on a secure UT server backed-up daily on two locations. Software will be stored in the GitLab repository hosted securely and backed up by the University of Twente (UT). The UT provides backups of the RE-SAMPLE website as well. The newsletter is distributed through MailChimp, which will provide own backups. All registered users can opt out of the newsletter at any time.

Sensitive data will be stored locally in partner organisations in secure environments. Thus, audio and video recordings will be stored at the local secure servers and will be destroyed once they have been transcribed. The data that can identify research subjects will be stored separately in either a local secure server in a password protected file or in a physical storage space under lock. Data storage in the Healthentia application, certified as a Class I Medical Device, is utilising the services of Microsoft Azure in servers of European Union member states (Microsoft acts as processor). Additionally, the Healthentia platform follows a rigorous Information Security Policy. Only pseudonymised data may be transferred strictly under joint controllership/data processing agreements between the partners. If a decision is made to upload some data to an open repository, they will be anonymised and/or encrypted before the upload.

7. Ethical aspects

Ethical aspects have already been treated separately in a number of dedicated deliverables: D1.4 *Ethics strategy*, D9.1 *H-Requirement No. 1* and D9.2 *POPD-Requirement No. 5*, as well as in the submitted study protocols, patient information and informed consent forms (D5.1 *First study subject approvals package RWD cohort* and D7.2 *First study subject approvals package cohort with companionship programme*). Furthermore, the questions of data management, security, and privacy have been resolved in WP4 and have been reported in iterative deliverables like D4.3, D4.10 & D4.11 *GDPR related and Security/Privacy Requirements*, D4.7, D4.13 & D4.14 *Measures for Organisational, Legal, and Technical Security and Privacy Requirements*, and in D4.8 *Security and Data Protection Policies*.

8. Other issues

While this DMP is an “umbrella” document covering data management matters which are applicable to the whole project, partner organisations are also encouraged to develop and use local plans for data management. Thus, the practices of data archiving (which data to archive and for how long) will vary per partner and per type of data. For example, most project management data will be archived through JOIN[19], the archiving system of the UT, and will be stored for 10 years.

[19] https://www.utwente.nl/en/service-portal/organisation-regulations-and-codes-of-conduct/archive/archiving-services/join-ut-document-management-system

9. Bibliography

European Commission. (2016). *Guidelines on FAIR Data Management in Horizon 2020 (v.3, 26 July 2016).* Retrieved from https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

European Commission. (2017). *Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020 (v.3.2, 21 March 2017).* Retrieved from https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

European Union. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC.* Retrieved from https://eur-lex.europa.eu/eli/reg/2016/679/oj

European Union. (2018). *Regulation (EU) 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a framework for the free flow of non-personal data in the European Union.* Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32018R1807

Force11. (2016). *The FAIR Data Principles.* Retrieved from https://www.force11.org/group/fairgroup/fairprinciples

Kopelyan, S., Garel, P., Kyriazakos, S., & Gonzalez, A. (2021). *Quality, risk, and IPR management plan.* Retrieved from https://www.re-sample.eu/.uc/faec539960102ab6f0802a2b7010254bbcd30f9ee2e9c00/RE-SAMPLE%20D1.1%20Quality%20Risk%20and%20IPR%20Management%20Plan%20v1.0%20Final.pdf

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., . . . Crosa. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*(160018).