iPROLEPSIS

D1.2 / Data management plan (initial version)

Editor	Contractual delivery date	Actual delivery date
Dimitropoulos Kosmas (CERTH)	June 2023	June 2023
Deliverable type	**Dissemination level**	**Version - date**
R - Document, report	PU – Public	1.0 - 29/06/2023

Funded by the
European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency. Neither the European Union nor the European Health and Digital Executive Agency can be held responsible for them.

Deliverable ID

Project acronym	iPROLEPSIS
Project full title	Psoriatic arthritis inflammation explained through multi-source data analysis guiding a novel personalised digital care ecosystem
Grant Agreement ID	101095697

Deliverable number	D1.2
Deliverable title	Data management plan (initial version)
Work package	WP1 – Management and coordination
Deliverable type	R - Document, report
Dissemination level	PU – Public
Version - date	1.0 - 29/06/2023

Contractual delivery date	June 2023
Actual delivery date	June 2023

Lead partner	CERTH
Editor	Kosmas Dimitropoulos (CERTH)
Contributors	Dimitrios Konstantinidis (CERTH), Ilias Papastratis (CERTH), Vasilis Charisis (AUTH), Georgios Apostolidis (AUTH), Jolanda Luime (EMC), Andreas Raptopoulos (WCS), Christos Chatzichristos (AING), Juliana Valle (TUM), Sofia Dias (FMH-ULISBOA), Silvia Reis (PLUX), Laura Coates (UOXF), Catia Gonsalves (SPR), Prince Ofori (HUJI)
Reviewed by	Andreas Raptopoulos (WCS) Christos Chatzichristos (AING)
Approved by	Leontios Hadjileontiadis (AUTH, Project Coordinator)

Keywords	Data management; Dataset; FAIR principles; Data security

Document history

Version	Date	Contributors	Action / status
0.1	20/03/2023	CERTH	Document structure (table of contents) ready
0.2	01/06/2023	ALL	Received individual DMPs from partners
0.3	13/06/2023	CERTH	Ready for internal review
0.4	23/06/2023	WCS	Reviewed by Andreas Raptopoulos (WCS)
0.5	24/06/2023	AING	Reviewed by Christos Chatzichristos (AING)
0.6	28/06/2023	CERTH	Document revised
0.6	29/06/2023	AUTH	Approved by the Project Coordinator
1.0	29/06/2023	AUTH	Submitted to the EC by the Project Coordinator

Contents

List of abbreviations

DMP	Data Management Plan
FAIR	Findable, Accessible, Interoperable, and Re-usable
WP	Work Package
PsA	Psoriatic Arthritis
DOI	Digital Object Identifier
PsO	Psoriasis
RDM	Research Data Management
EHR	Electronic Health Record
OMOP	Observational Medical Outcomes Partnership
CSV	Comma-separated values
JSON	JavaScript Object Notation
PNG	Portable Network Graphics
EOSC	European Open Science Cloud
BS	Bower sound
PDPID	PsA digital phenotyping and inflammation drivers
EGG	Electrogastrogram

Executive summary

This document aims to serve as the first version of the Data Management Plan (DMP), formulated by the partners of the iPROLEPSIS project (EU Horizon Europe RIA program under Grant Agreement No 101095697). The current document describes the data that will be gathered and processed during the project, the data adherence to the Findable, Accessible, Interoperable, and Re-usable (FAIR) principles and the data storage and access protocols. In addition, an estimation of the required resources for making data FAIR compliant is reported, while legal and ethical issues related with the generated data are raised. This document will be a live document, meaning that it will be updated throughout the lifecycle of the project using the European Open Science Cloud (EOSC) ARGOS service in order to better serve the needs of the project.

1 Introduction

This document is the first iteration of the Data Management Plan (DMP) developed within the scope of the Horizon Europe iPROLEPSIS project. The primary objective of the DMP is to describe the type, size, and structure of data that will be generated during the project. In addition, the DMP aims to ensure that the data collected and managed during the project are Findable, Accessible, Interoperable, and Re-usable (FAIR). Finally, the DMP aims to outline the protocols and methodologies for handling the data effectively throughout the project's lifecycle from collection to potential reuse after the end of the project, as well as describe the protection against unauthorized or improper use.

The specific objectives of this first version of the DMP include:

1. Provide a comprehensive description of the data management lifecycle for the data to be collected and/or generated.
2. Present the methodology that will be employed to ensure FAIR principles are adhered.
3. Offer detailed and descriptive information about the types of data that will be collected and/or generated and how they will be managed, both during and after the project.
4. Describe the strategies by which the collected/generated data will be made openly accessible and searchable to stakeholders.
5. Estimate the required resources for making data FAIR compliant.

The iPROLEPSIS project, due to its nature, involves human participants, and as such, will generate and manage personal and sensitive data. The project is committed to strictly adhering to all relevant national, EU and international ethics for conducting research with human participants and their personal data.

1.1 Document scope

This document comprises the initial version of the DMP and it is called deliverable D1.2 "Data management plan (initial version)" that is submitted in M6. It is followed by deliverable D1.6 "Data management plan (final version)" that will be submitted in M48 to outline the final version of the DMP. Both deliverables are parts of Task 1.4 "Ethics, legal and data management" of WP1 "Management and coordination".

The scope of this document is to develop the initial version of DMP providing details on the handling, storing, and sharing of data generated within the iPROLEPSIS project. Overall, the DMP covers all aspects of data management that will be implemented throughout the project's lifecycle and includes guidelines for:

- Data Collection: This involves the methods and tools used for gathering data, the types of data collected, and the standards followed during the collection process.
- Data Storage: This pertains to the storage solutions used for keeping the data secure and accessible, including details about the data backup and recovery processes.
- Data Sharing: This includes the strategies for sharing data within the project team and with external stakeholders, including the public, where applicable. It also covers the mechanisms for ensuring data privacy and compliance with relevant regulations.
- Data Preservation: This relates to the long-term preservation of data, including the strategies for maintaining data integrity and accessibility over time.

The DMP is a living document and will be updated at various stages of the project to reflect any changes in data management practices or requirements. Although part of WP1, this deliverable receives feedback of all project partners and is directly related with other work

packages that are responsible for generating or processing data, such as WP2, WP3, WP4 and WP5.

1.2 Document structure

This document is organised in seven main sections and an appendix. Section 1 gives a brief introduction to the data management plan and its scope. Section 2 describes the principles that the DMP of the iPROLEPSIS project will follow. Section 3 provides details of the data that will be generated during the project, along with information on how they are structured, their size and file types. Section 4 illustrates the strategy that will be followed so that the generated data adhere to the FAIR principles. Section 5 provides information on the allocation of resources for the collection of data, while Section 6 describes the steps that will be followed so that the data remain secure and protected from unauthorised users. Finally, Section 7 describes ethical and legal aspects related to the data. This document ends with an Appendix that contains the individual DMP template that was used to collect feedback from the partners of the iPROLEPSIS project in order to develop the initial version of the DMP.

2 Principles of the iPROLEPSIS data management plan

The Data Management Plan constitutes a guide regarding the anonymisation, exchange and release of data gathered during the project. Datasets produced within the iPROLEPSIS project span users' demographic and clinical data to data from sensing devices, and user generated inputs. These data may be of value to the research community for benchmarking AI algorithms with respect to several aspects of Psoriasis (PsO) and Psoriatic Arthritis (PsA) detection and progression, as well as the transition from PsO to PsA. In addition, the collected data may contribute to the establishment of a common basis for augmented policy decision making. Since iPROLESIS data collection phases involve human participants, data collected include, in cases, sensitive, personal information, and, as a result, the focus is placed on possible ethical issues and access restrictions regarding personal data, so that no regulations on sensitive information are violated.

In this context, this version of the DMP describes in detail the datasets that will be collected during the project by the different technical and clinical partners of the project. Individual data collection mechanisms are treated as individual cases and therefore individual data management plans are foreseen.

Regarding data storage and maintenance, a dedicated data management system will be built on the Hetzner[1] cloud infrastructure to serve as a trusted repository for the collected data. Hetzner is a public cloud provider, whose infrastructure supports backup, has data security mechanisms in place, follows repository and metadata standards and complies with the GDPR requirements. For clinical data collected from the pilot studies, electronic consent forms will be given to all patient participants, asking them for their approval to collect and process their data. Then, the data will be centralized using the OpenClinica platform[2], allowing access only to clinical personnel involved in the respective studies. All data collection procedures will receive approval by the research ethics committee or the administration of each clinical site (PSR, UOXF, EMC, AUTH), where the participant recruitment and collection take place, before

[1] Hetzner: https://www.hetzner.com/
[2] OpenClinica: https://www.openclinica.com/

the clinical studies are conducted. The 4 countries that have been selected for pilot studies are: Netherlands, UK, Portugal and Greece.

Additional data sources will be gathered from previous European projects of the partners (e.g., PROTEIN) and retrospective observational cohort studies, such as DEPAR[3] and MONITOR[4]. These existing data will complement the data collected from the pilot studies and assist in the pre-training of the AI models that will be developed in the framework of the iPROLEPSIS project.

Data access will depend on the sensitivity level of the data, with some datasets becoming publicly available after de-identification in well-known repositories (e.g., Zenodo[5]) and other datasets remaining protected or confidential with limited or no access to third parties.

Finally, since iPROLEPSIS favors relevant EOSC platforms, the DMP will be developed and updated using the EOSC ARGOS[6] service, thus simplifying its creation and maintenance, and facilitating compliance of data sharing with FAIR principles and the GDPR.

3 Data Summary

The iPROLEPSIS project aims at enhancing PsA diagnosis and care by unravelling the process of health transitioning into PsA. More specifically, the project seeks to obtain an array of medical, lifestyle and environmental data from clinical studies and wearable sensors / smartphones in order to assess inflammatory symptoms of Psoriasis (PsO) and PsA, as well as gain insights on the progression of the illness and the transition from PsO to PsA. To this end, explainable AI (xAI)-based personalized and data-driven models will be developed for the prognosis, diagnosis and analysis of the severity and progression of the illness through the processing and fusion of personal, nutritional, clinical, environmental, and imaging features, as well as biological markers. The project also involves the design, implementation, and validation of innovative personalized interventions aimed at relieving PsA symptoms and optimizing treatments. The final outcome will be a validated, privacy-conscious, xAI-based toolkit that provides healthcare professionals with quantifiable, understandable evidence for disease screening, monitoring, and treatment optimization. Furthermore, it will empower individuals who have or are at risk of PsA with tailored insights to enable informed health management.

The data used within the iPROLEPSIS project encompasses an extensive range of sources to facilitate a comprehensive understanding of PsA and its progression. This multifaceted data set includes:

- **Medical and Clinical Data:** Detailed patient electronic health records, medical history, clinical observations, and related health data will be utilized. This information plays a crucial role in identifying key drivers of PsA and formulating personalized prognosis models.
- **Well-being, Lifestyle, Environmental, and Occupational Data:** A holistic perspective is ensured by including data on patients' overall well-being, lifestyle habits, job-related conditions, nutrition, and physical activity that might affect the development or progression of PsA.

[3] DEPAR study: https://ciceroreumatologie.nl/depar
[4] MONITOR study: https://www.ndorms.ox.ac.uk/octru/trials-portfolio/trials-open-to-recruitment-2/monitor
[5] Zenodo: https://zenodo.org/
[6] EOSC Argos: https://argos.openaire.eu/splash/

- **Biological markers:** Genomic and microbiome data, along with hair cortisol, will be measured and leveraged to improve the understanding of PsA at a molecular level, and these biomarkers will be incorporated into the personalized prognosis models.
- **Imaging Data:** Non-invasive skin microvascular and joint imaging data (obtained through Optoacoustics) will be used, enhancing the project's capability to detect PsA.
- **Digital Phenotyping Data:** Passive and active sensing with smart devices, like smartphones and smartwatches, will be employed to track key risk and progression markers of PsA in daily life.
- **Intervention Data:** Information from the implementation and validation of personalized interventions will be collected, aiming to offer insights into the optimization of treatments and symptom relief.

The collected data will be pseudo-anonymized to respect privacy norms and will contribute to the development of an xAI-based toolkit for disease screening, monitoring, and treatment optimization.

3.1 List of datasets

In this section, we present a detailed description of the datasets that will be collected or employed during the project, along with information on the types of files used, size of dataset and responsible partner. The size of the collected datasets cannot be determined accurately and thus it is usually presented in ranges.

Dataset #1: Clinical assessment and evaluation data		
Description: A large array of clinical data will be collected during the iPROLEPSIS PsA digital phenotyping and inflammation drivers (PDPID) study that will include patients' user profile data (e.g., age, sex, years of disease, medication, comorbidities, EHRs, etc.), as well as clinical evaluation data (e.g., joint count for swelling and tenderness, tendon count for enthesitis, body surface area, BMI, abdominal circumference, etc.), blood data (i.e., DNA, inflammatory blood-marker CRP), hair data (i.e., cortisol) and gut microbiome data. These clinical data will be used for the assessment of drivers of inflammation exacerbation and the transition from PsO to PsA.		
File types: .csv	**Size:** > 1 TB	**Responsible partner(s):** EMC

Dataset #2: Skin data		
Description: As part of the PDPID study, skin biopsies will be conducted on adult patients with plaque-type PsO in order to assess the role of mast cells in the transition from PsO to PsA.		
File types: .xlsx, .png	**Size:** ~ 50 GB	**Responsible partner(s):** HUJI

Dataset #3: Questionnaire data		
Description: During the PDPID study, patients with PsO and PsA will be provided with questionnaires, in which they will complete data regarding demographics, job title, health care usage, care activities, pain, work productivity, etc. Additionally, daily questionnaires will be administered to patients via the iPROLEPSIS mobile app or the Reuma.pt website, in which they will report the severity of pain, morning stiffness and tiredness in a Likert scale.		
File types: .csv, .xml	**Size:** < 10 GB	**Responsible partner(s):** CICERO, UOXF, PSR

Dataset #4: Bowel sound and Electrogastography data

Description: PLUX aims to employ its smartbelt to unobtrusively capture bowel sounds (BS) and electrogastrogram (EGG) in order to develop objective digital indicators (biomarkers) for the presence and severity of inflammatory symptoms associated with PsA.		
File types: .txt	**Size:** ~ GB range	**Responsible partner(s):** PLUX

Dataset #5: Environmental data		
Description: Retrospective environmental pollution data will be collected from governmental agencies, while prospective data can be collected through clinical trials, wearables and questionnaires. These data will contain information on temperature, humidity and air pollution and the aim is to investigate correlations and the long-term influence of environmental data to PsA evolution.		
File types: .csv	**Size:** ~ GB range	**Responsible partner(s):** AING

Dataset #6: Optoacoustic data		
Description: Macroscopic (joints) and mesoscopic (skin) optoacoustic tomograms will be obtained from healthy volunteers and PsA patients to investigate the effect of PsA on the joints and skin microvasculature. In addition, novel imaging-based markers will be extracted from the optoacoustic data for PsO/PsA detection and progression.		
File types: .MSOT, .US, .mat, .xlsx	**Size:** ~ 1 TB	**Responsible partner(s):** TUM

Dataset #7: Foot, hand, and nail images		
Description: Images of feet and hands will be collected from smartphone devices during the PDIPD study or employed from retrospective studies. The purpose is to develop AI models that assess the swelling of toes and fingers, as well as the condition/deformation of nails that are evident to patients suffering from PsO/PsA.		
File types: .jpeg, .png	**Size:** < 10 GB	**Responsible partner(s):** AUTH, CERTH

Dataset #8: Body/hand movement assessment videos		
Description: Videos of hand and body movements from patients will be collected through the smartphone device and compared against movements of normal cases in order to assess the severity of dyskinesia of a PsA patient.		
File types: .avi, .mp4	**Size:** 10-100 GB	**Responsible partner(s):** CERTH

Dataset #9: Wearable data		
Description: Retrospective data from publicly available repositories (e.g., PhysioNet[7]), as well as prospective data, collected from smartwatches and smartphones during the PDPID study, will be employed in order to derive appropriate digital biomarkers for assessing drivers of inflammation exacerbation in PsA patients in daily living. These wearable data include, among others, accelerometer, gyroscope and physical activity (e.g., steps, distance, intensity, etc.) data, as well as heart rate, heartrate beat-to-beat interval, sleeping time and screen time. Important biomarkers that can be extracted from these data include heart rate variability, sleep quality and stress that are important factors for PsA progression and symptom exaggeration.		
File types: .csv	**Size:** ~ GB range	**Responsible partner(s):** AUTH, WCS, AING

[7] PhysioNet repository: https://physionet.org/

Dataset #10: Keystroke dynamics data		
Description: Typing events data, e.g., key hold time, press latency and release latency timeseries will be collected from smartphones during the PDIPD study and used as indicators of mood, hand function, presence of flare and disease progression.		
File types: .csv	**Size:** ~ GB range	**Responsible partner(s):** AUTH, WCS

Dataset #11: Nutritional data		
Description: Retrospective (from the EU-funded Horizon 2020 PROTEIN[8] project) and prospective nutritional data (e.g., meal composition, energy intake, macronutrients, micronutrients, etc.) will be collected and used by the AI recommendation engine of the iPROLEPSIS project in order to provide personalized meal recommendations to users. A weekly meal plan specifically built based on the patient's user profile will be sent to the iPROLEPSIS patient mobile app to inform the patient on their nutritional needs.		
File types: .csv, .xlsx	**Size:** < 1 GB	**Responsible partner(s):** CERTH

Dataset #12: Serious game data		
Description: Questionnaire data, as well as serious game data will be collected from PsA patients that perform specific physical exercises based on how well they perform the exercises. The purpose of these data are to inform health care professionals on a patient's physical abilities, as well as improve and/or sustain the wellness of patients and prevent health deterioration.		
File types: .csv, .xml, .json	**Size:** 5 - 10 GB	**Responsible partner(s):** FMH

3.2 File type description

Common file types, such as Comma-separated values (CSV[9]), JavaScript Object Notation (JSON[10]) and Portable Network Graphics (PNG[11]), will be used for all datasets collected during the iPROLEPSIS project to ensure maximum accessibility, interoperability and reusability according to the FAIR principles. In addition, the use of common file types enables full compatibility with popular software, thus allowing the use and processing of the collected data during the lifecycle of the project and even after the end of the project.

Finally, a common data model that complies with the Observational Medical Outcomes Partnership (OMOP)[12] standard will be employed for the clinical data collected in the different clinical sites during the PDIPD study. In this way, the iPROLEPSIS project will ensure that the collected clinical data are standardized and easily accessible and reusable by all partners of the project.

[8] EU-funded Horizon 2020 PROTEIN project: https://protein-h2020.eu/
[9] CSV file format: https://dev.socrata.com/docs/formats/csv.html
[10] JSON file format: https://developer.mozilla.org/en-US/docs/Learn/JavaScript/Objects/JSON
[11] PNG file format: https://www.adobe.com/gr_en/creativecloud/file-types/image/raster/png-file.html
[12] OMOP standard: https://ohdsi.org/omop/#:~:text=The%20Observational%20Medical%20Outcomes%20Partnership,the%20effects%20of%20medical%20products

4 FAIR data

Proper Research Data Management (RDM)[13] is mandatory for any Horizon Europe project generating or reusing research data. It is a key part of Horizon Europe's open science requirements and states that beneficiaries must manage the digital research data generated in a project responsibly, in line with the FAIR principles, to stimulate knowledge acquisition and innovation.

The FAIR data strategy is encapsulated by its acronym:

- **F**indable data: Implementing straightforward naming and versioning systems for data and metadata, integrating search keywords and utilizing DOIs.
- **A**ccessible data: Outlining methods for data availability and specifying tools necessary for data access.
- **I**nteroperable data: Employing standards and vocabularies for metadata and data types.
- **R**eusable data: Detailing the period during which data will be available and setting licensing conditions for data.

For this project, all collected data from research participants will be pseudonymized, ensuring that no personal data relating to identifiable individuals or research participants is disclosed publicly.

Additionally, given that the project deals with sensitive data, it's important to evaluate the risks associated with each type of data before deciding on their public disclosure. Particularly, the feasibility of re-identifying research participants by integrating other data sets should be considered.

4.1 Findable data

The data collected in this project will be identified by a persistent identifier. This identifier will be a study id and a participant id for both clinical data and smartwatch/smartphone data. This procedure will ensure that all data are de-identified. Both study id and participant id will be registered in the miPROLEPSIS app at the time of participant recruitment, during which the participants download and install the app in their smartphones and fill in the study id and participant id that will be provided by the recruiter to a dedicated screen. The identifier will be applied to link the user with the data collected, but no other identifiers will be used or stored that may lead to user identification. The key tables for the de-identified data will be held by the coordinating team of each clinical site that may provide limited access to other users. Anonymization techniques may be implemented after the end of the project, but this is not guaranteed. Therefore, a persistent identifier, with access to some additional information, could identify the data.

Rich metadata will be provided to facilitate discovery, identification and detailed description of the data produced. The exact type and format of this metadata will be decided in the future. In addition, search keywords will be provided in the metadata to optimize the possibility for discovery and potential re-use. These keywords will be mostly related to data protection law. The metadata will be offered in such a way that it can be harvested and indexed. They will comply with the Open Archives Initiative's Protocol for Metadata Harvesting (OAI-PMH) and

[13] Horizon Europe mandate for RDM: https://www.openaire.eu/how-to-comply-with-horizon-europe-mandate-for-rdm

will be searchable via the OpenAIRE catalogue and Zenodo, both of which support web-based search. Finally, the aim is to ensure that the metadata is as accessible and usable as possible.

4.2 Accessible data

The project ensures that data is deposited in a platform built on the Hetzner infrastructure, a trusted public cloud provider. This platform supports backup, has data security mechanisms in place and follows repository and metadata standards. Clinical partners will use local infrastructures/servers for data collection and storage, while copies of the clinical data will be securely transferred to the Hetzner cloud infrastructure. Data from wearable sensors (i.e., smart watch, smartphone) will be collected via the iPROLEPSIS mobile (miPROLEPSIS) app and then transferred to the Hetzner cloud infrastructure. The data will be erased from the smartphone upon successful transfer or if the application is uninstalled. On the other hand, weather and pollution data, as well as photos of nails and joints will be transferred directly to the Hetzner cloud infrastructure. Moreover, identifiers will be assigned to the data, probably digital object identifiers (DOI). All data communication will be encrypted, and security measures are in place for data security at the platform. Data lineage is always clear and each of the partners that work with data have security measures in place that prohibited unauthorised use of the data.

The iPROLEPSIS project will ensure restricted access to sensitive data, as well as deliverables with private or sensitive content, while the rest of the data and the produced content of the iPROLEPSIS project will be made publicly available to promote research and innovation. In cases where an embargo is applied to give time to publish or seek protection of intellectual property, the maximum duration of the embargo period is set to 6 months, to provide time to the user group that proposed the test to publish the results. After this period, the data will be processed, analyzed, and published as soon as possible. Access to the data will be possible through authentication mechanisms that ensure role-based access control, while a free and standardized access protocol will be provided, where possible. For private and sensitive data, access will be provided to authorized users only and user authentication mechanisms and interfaces for secure data exchange will be provided and put in place.

The metadata of an experimental campaign will be made openly available upon completion of the testing campaign, while non-sensitive research data itself will be made openly available as soon as possible. Metadata will be made openly available and licensed under a public domain dedication CC0, as per the Grant Agreement. This metadata will contain information to enable the user to access the data. The metadata will be available during the lifetime of the project and for a specified period after the end of the project, which will be determined by each responsible partner individually at a later stage of the project. For all data packets stored in the repository, a complete set of metadata will be provided and stored. The metadata will also accompany each data packet in the data repository where it is stored.

All data will be accessible by several open-source editors and the relevant documentation will be provided. This ensures that the data is not only accessible but also usable, as users will have the necessary tools and information to interpret and work with the data.

4.3 Interoperable data

The project will use various data and metadata vocabularies, standards, formats, and methodologies to ensure data interoperability. This will allow for data exchange and re-use within and across disciplines. After internal meetings, all iPROLEPSIS partners agreed on the development of a common data model based on the OMOP standard to describe the data and metadata. Thus, interoperability will be facilitated by adopting standard formats and units of

measure widely used in the community and by providing proper documentation in terms of Readme files. Data harmonization and standardization will be led by the management and coordination partners of project (Task 2.4) on the Hetzner cloud infrastructure.

In the case of data transfer from applications to the central cloud data management system, REST APIs will be utilized, with the exact design of the interfaces and standards applied currently under discussion. REST APIs allow for the efficient exchange of data between systems, making them ideal for this purpose.

The iPROLEPSIS project does not plan to use uncommon or generate project-specific ontologies or vocabularies. The iPROLEPSIS partners aim to amend the OMOP ontology and ATHENA[14] vocabulary via the participation of an iPROLEPSIS representative (AING) in the "Vocabulary workgroup" of OHDSI. However, the current plan is to use existing ontologies and vocabularies, when possible, which will ensure that the data is compatible with other datasets and can be easily understood and used by others in the field. The data produced in the project will include qualified references to other data. The data packets produced are self-contained and are fully described by the associated metadata. Proper citation of the data is ensured by the citation text, as well as a unique DOI number, provided for each data packet, which is also provided at the repository, in which the data resides. Some data packets will link to other data, such as HPOS data or Reuma.pt data, providing a context for the data and allowing for a more comprehensive understanding of the data.

4.4 Re-usable data

The iPROLEPSIS project will ensure that a comprehensive documentation of the data is provided to facilitate data re-use. This documentation will include Readme files, variable definitions, units of measurement and other relevant information. The exact form of documentation will be decided in the future, but it will likely include details on methodology, codebooks, data cleaning, analyses and more. Part of the documentation might also be written in the corresponding deliverable of each work package of the project.

The iPROLEPIS project has varying levels of data availability for re-use based on the nature of the data. Non-sensitive data will be made freely available in the public domain by depositing data and associated metadata to trusted repositories in order to permit the widest re-use possible. On the other hand, private and sensitive data will offer restricted assess and the data licencing approach will be decided at a later stage of the project. The type of license, if any, will be decided in accordance with the obligations set out in the Grant Agreement. The usability of the project's data by third parties, particularly after the end of the project, depends also on the nature of the data and the choices of the responsible partners. Some partners aim to provide free access to the data produced in the project, while others do not plan to make their data usable by third parties.

The project will thoroughly document the provenance of the data using appropriate standards. The data to be uploaded are recorded data and become known for the first time. The Hetzer platform, as well as the infrastructure provided by INTRA, ensure that versioning is supported. In this way, the origin and lifecycle of the data are clearly defined, enhancing the data re-usability.

Finally, all partners will process and publish data according to FAIR principles and best practice guidelines. After the project, periodic quality assessment procedures will take place

[14] ATHENA vocabulary: https://athena.ohdsi.org/search-terms/start

to ensure the ongoing integrity and usability of the data. These processes will ensure that the data remains reliable, accurate and suitable for re-use over time.

5 Allocation of resources

Resources will be allocated to ensure that all data gathered during the lifetime of the iPROLEPSIS project abide by the FAIR principles. As a result, there will be direct and indirect costs related to:

- Data storage and infrastructure
- Data curation and processing
- Open access publications

A better view on the costs will be available in a later stage of the project, but all of them will be covered by the project funding. In addition, a cost estimate associated with data storage and infrastructure setup on the Hetzner cloud provider has been provided in accordance with the requirements of the Grant Agreement under the partner INTRA (ref. GA – Annex 2).

Regarding data curation, processing and storage, each partner that collects data will be responsible for assigning a Data Protection Officer (DPO) that will oversee that the gathered data are compliant to EU directives 2016/679[15] and EU 2016/680[16] on the processing of personal data.

6 Data security and protection

Any gathered data will be securely handled throughout the entire duration of iPROLEPSIS so as to protect it from loss and unauthorized access. All data will be pseudo-anonymized, meaning that each patient will be represented by a study id and a participant id, making it impossible to be identified and/or mapped to their data. The infrastructure that will be used for the data storage and analysis will be hosted by a trusted cloud provider (Hetzner) under a direct contract with INTRA. Clinical data will be collected via eCRF and stored in the local servers of the countries participating in the PDPID study prior to being encrypted for a secure transfer to the Hetzner cloud infrastructure. On the other hand, data from smartwatches will be first transferred to the smartphone via Bluetooth connection and then transferred to the Hetzner cloud infrastructure. Similarly, data collected through the use of the miPROLEPSIS app will be directly and securely transferred to the Hetzner cloud infrastructure. Finally, weather and air pollution data will be collected by utilizing an external public web-service and then sent to the Hetzner cloud infrastructure. All data communication will be encrypted and security measures are in place for data security at the Hetzner cloud infrastructure. Data stored in the smartphone will be sandboxed, ensuring that no information will be shared with other apps. The users will be provided with options to opt out from sharing the data with others as well as exercise their right for their data to be forgotten. Additionally, raw data will be locked to avoid (accidental) modifications of the raw data.

Data lineage is clear at all times and each of the partners that work with data have security measures in place that prohibited unauthorised use of the data. Data infrastructure provided by INTRA will offer access only to authorised users. User authentication mechanisms and interfaces for secure data exchange with the cloud-based back-end will be provided and ensure role-based access control (e.g., health personnel, scientific personnel, etc.). Data access to the Hetzner cloud infrastructure will be provided by the Data Steward (to be

[15] EU directive 2016/679: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679
[16] EU directive 2016/680: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016L0680

appointed), while access and use of data will be logged. The security measurements performed by INTRA at the VMs hosting the servers are:

- encryption at rest by encrypting the hard disks and making them readable only to authorized entities,
- firewall protection using the iptables/netfilter framework for allowing only specific connections with external servers,
- authenticated access using SSH public key authentication through cryptographic keys, while password-based authentication is disabled,
- encrypted and secure communication between the servers using HTTP over TLS (HTTPS), and
- adoption of policies for the creation of secure passwords.

In addition, physical security is provided by Hetzner, whose data centers are certified in accordance with ISO/IEC 27001. The iPROLEPSIS partners will ensure that the data remains protected under all necessary security controls (including backup policies and integrity checks) and access controls (identification, authentication, authorization). In the unfortunate event of personal data breach, each of the clinical sites will follow the procedure for data breaches as given by their institutions. Such a procedure includes the notification of the department head or supervisor, the notification of the data subject(s) that may have been affected by the breach and the creation of an internal report that documents the type and content of data that have been compromised.

Regarding long-term data preservation and curation, each partner producing data will be responsible to develop their own strategy for handing the data after the end of the project, since these data will be stored in institutional archives. PLUX aims to store raw anonymized physiological data for 5 years after the project ends, while WCS and UOXF aims to delete all data after the end of the project. On the other hand, HUJI and TUM aim to employ the RSU Dataverse for their data repositories. With this approach, HUJI and TUM will administer and update their data repositories on regular basis, thus ensuring data curation and long-term preservation for 20 and 10 years, respectively. Finally, since iPROLEPSIS collects medical data from human subjects, EMC and CICERO in Netherlands are legally obliged to preserve these data for at least 15 years.

7 Ethical and legal aspects

Since the iPROLEPSIS project is concerned with the collection of private and sensitive data, such as personal information, medical records, and images, from patients with PsO and PsA, ethical and legal aspects arise that will be dealt by the partners of the project accordingly. T1.4 “Ethics, legal and data management”, led by DBC, is dedicated to identifying and formulating the key ethical (for healthcare and clinical research) and legal (focusing on data protection, FAIR principles, and GDPR) requirements that are pertinent to the project and will govern its activities, including the development of the digital health ecosystem and the design and execution of clinical studies.

Specifically, patients are asked by their attending physicians to participate during consultation. They receive a letter and the patient information leaflet and within 7 days they will be approached by the research team to ask whether they want to participate. If they agree they sign the consent form and they can receive a signed copy of the consent form. The consent form will be created in local languages relevant to each hospital location. Prior to the participation, each patient will be kindly asked to read the patient information sheet (PIF) that

includes information on data storage, privacy and participant rights and sign the informed consent form on whether they agree with data collection, processing and preservation and after given enough time to consider all given information. Contact information of the hospital and research team for questions or complaints will also be included. The information sheet and the consent form will be made available in this document as soon as possible.

The patients will be able to indicate what type of data they give permission to store and if they refuse to participate, their data will not be collected. Automatic data collection on digital parameters can be controlled from the iPROLEPSIS mobile app. Clinical, biological data are collected physically, and questionnaires are provided digitally, therefore patients have control on what data to provide. Patients are informed that they have the right to withdraw at any moment of data collection. When a participant withdraws consent, the data collected up to the moment of withdrawal will be used (according to GDPR/AVG Article 89(1)[17]), but no further data will be collected. Participants can request their own data by contacting the research team. Upon request, the identity of the person will be verified and the principal investigator will make a data extraction of the participants' data only and will send the data to the participant via an encrypted link (SURFfilesender[18]).

Additionally, iPROLEPSIS will ensure that the exposed metadata from the collected datasets will not record confidential or restricted information. Finally, the Hetzner cloud infrastructure that will store and manage the generated datasets of the iPROLEPSIS project offers a Data Processing Protection Agreement that can ensure the GDPR compliance, while in M47, before the final platform is released, a privacy statement will be drafted to confirm that data are handled within the appropriate legal framework (GDPR, etc.).

Conclusions

The key takeaways from D1.2 are:

- An initial version of a detailed DMP is formulated that describes the data that will be gathered and processed during the project.
- A presentation of the methodology employed to ensure data adherence to the FAIR principles is made.
- The types of data that will be collected and/or generated during the project are described.
- The required resources for making data FAIR compliant are estimated.
- Ethical and legal issues related with the generated data are raised.

This document will be available online using the EOSC ARGOS service that will enable the DMP to be continuously updated during the lifespan of the iPROLEPSIS project until the submission of the deliverable D1.6 "Data management plan (final version)" in M48. The next update and refinement of the DMP will take place internally using the EOSC ARGOS service in M30 (internal report) after the end of the first clinical study, during which a first version of the datasets will have been collected and processed and the iPROLEPSIS partners will have a good knowledge of the type, size and content of data collected.

[17] Art. 89 GDPR: https://gdpr-info.eu/art-89-gdpr/
[18] SURFfilesender service: https://www.surf.nl/en/surffilesender-send-large-files-securely-and-encrypted

Appendix I: Horizon Europe Dataset Description Template

Data Summary

Will you re-use any existing data and what will you re-use it for? State the reasons if re-use of any existing data has been considered but discarded.

What types and formats of data will the project generate or re-use?

What is the purpose of the data generation or re-use and its relation to the objectives of the project?

What is the expected size of the data that you intend to generate or re-use?

What is the origin/provenance of the data, either generated or re-used?

To whom might your data be useful ('data utility'), outside your project?

FAIR data

Making data findable, including provisions for metadata

Will data be identified by a persistent identifier?

Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?

Will metadata be offered in such a way that it can be harvested and indexed?

Making data accessible

Repository:

Will the data be deposited in a trusted repository?

Have you explored appropriate arrangements with the identified repository where your data will be deposited?

Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

Data:

Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.

If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g., patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Will the data be accessible through a free and standardized access protocol?

If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?

How will the identity of the person accessing the data be ascertained?

Is there a need for a data access committee (e.g., to evaluate/approve access requests to personal/sensitive data)?

Metadata:

Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?

How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?

Will documentation or reference about any software be needed to access or read the data be included? Will it be possible to include the relevant software (e.g., in open-source code)?

Making data interoperable

What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?

Will your data include qualified references1 to other data (e.g., other data from your project, or datasets from previous research)?

Increase data re-use

How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g., readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?

Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?

Will the data produced in the project be useable by third parties, in particular after the end of the project?

Will the provenance of the data be thoroughly documented using the appropriate standards?

Describe all relevant data quality assurance processes.

Allocation of resources

What will the costs be for making data or other research outputs FAIR in your project (e.g., direct and indirect costs related to storage, archiving, re-use, security, etc.)?

How will these be covered? Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions)

Who will be responsible for data management in your project?

How will long term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long)?

Data security

What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?

Will the data be safely stored in trusted repositories for long term preservation and curation?

Ethics

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?

Other issues

Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones (please list and briefly describe them)?