DREAMS project

[D8.1] Data Management Plan (DMP)

Project Founded by the European Commission		
DISSEMINATION LEVEL		
PU	Public	☒
SEN	Sensitive	☐

DREAMS project. Funded by the European Union under 101080229-2. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union (EU) or European Research Executive Agency (REA). Neither the EU nor REA can be held responsible for them.

Funded by the European Union

DELIVERABLE INFORMATION	
Deliverable N°	[D8.2]
Deliverable Title	[Data Management Plan (DMP)]
WP	[WP8 Project Management]
Lead Beneficiary	CECS
Contributing partners	Zabala
Type	[DMP — Data Management Plan]
Due Submission Date	04/2024
Actual Submission Date	29/04/2024

DOCUMENT HISTORY			
DELIVERABLE RESPONSIBLE		[CECS]	
Contributors	**Organization**	**Reviewers**	**Organization**
Laura Brullé-Soumaré	CECS	Ane Garcia	Zabala
Nicolas Maignan	Kantify	Segolene Martin	Kantify

VERSION HISTORY		
VERSION	**DATE**	**COMMENT**
1	10/04/2024	First draft by CECS
2	18/04/2024	Review 1 by Zabala
3	25/04/2024	Review 2 by Kantify
4	26/04/2024	Final

DREAMS project. Funded by the European Union under 101080229-2. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union (EU) or European Research Executive Agency (REA). Neither the EU nor REA can be held responsible for them

Funded by the European Union

Table of Contents

List of Tables

List of Abbreviations

TERM	DEFINITION
ADMET	Absorption, Distribution, Metabolism, Excretion, and Toxicity
DMP	Data Management Plan
DOI	Digital Object Identifiers
EU	European Union
GDPR	General Data Protection Regulation
GEO	Gene Expression Omnibus
IP	Intellectual Property
iPS	Induced Pluripotent Stem Cells
MIAME	Minimum Information About a Microarray Experiment
MIAPE	Minimum Information About a Proteomic Experiment
MIQE	Minimum Information for Publication of Quantitative Real-Time PCR Experiments
ML	Machine Learning
MoA	Mechanism of action
NMD	Neuromuscular Disorders
NME	New Molecular Entity
PAB	Patient Advisory Board
SOPs	Standard Operating Procedures
WP	Work Package

DREAMS project. Funded by the European Union under 101080229-2. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union (EU) or European Research Executive Agency (REA). Neither the EU nor REA can be held responsible for them

Funded by the European Union

1 Executive summary

The present document represents DREAMS projects' Deliverable 8.2 – Data Management Plan (DMP).

Set to span five years, the DREAMS consortium is paving the way for a novel approach to treating neuromuscular diseases. This transformative project merges the power of artificial intelligence, pluripotent stem cells, and cutting-edge pharmacological screening techniques to seek out therapeutic solutions for five debilitating conditions affecting more than 400 000 people worldwide.

The DREAMS consortium unites a highly skilled and complementary consortium of nine European partners, encompassing clinicians, patients, academic researchers, and private stakeholders. These organizations include: I-Stem (France), Kantify (Belgium), The Institute of Myology (France), Center for Neuroscience and Cell Biology - University of Coimbra (Portugal), The Technion - Israel Institute of Technology (Israel), Samsara Therapeutics (UK), Assistance Publique - Hôpitaux de Paris (France), AFM-Telethon (France) and Zabala Innovation (Spain)

2 Data generation and reusing strategy

DREAMs partners will both generate high valued new data and reuse existing datasets for achieving the project objectives.

Generating new insights

The project is set to generate a comprehensive array of data types, including imaging, numerical, transcriptomic, proteomic, tabular and textual data. These datasets are crucial for dissecting the pathology of Neuromuscular Disorders (NMDs), facilitating the identification of disease mechanisms at molecular and cellular levels as well as the validation of new drug candidates, and innovating clinical trial designs.

Leveraging Existing Datasets

A significant aspect of the DREAMS project is the strategic reuse of existing datasets, notably for the training of machine learning (ML) models using extensive databases such as PubChem, Dailymed, etc. This approach is aimed at enhancing the project's analytical capabilities from its start and accelerating the discovery process. Public omics data from previously published studies will also be integrated selectively, based on strict criteria to ensure relevance and quality.

Summary of DREAMS data

Below is a detailed table summarising the data generation efforts across the lifespan of the DREAMS consortium. This table delineates the nature, purpose, and estimated volume of the data generated or reused in each work package. It serves as a testament to the project's comprehensive approach to understanding and treating neuromuscular diseases through innovative data-driven strategies.

DREAMS project. Funded by the European Union under 101080229-2. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union (EU) or European Research Executive Agency (REA). Neither the EU nor REA can be held responsible for them

Funded by the European Union

Table 1. Data generated in the DREAMS project.

WP	Data	Description (source, type, format)	Generating partner	Size
WP1	Imaging (Electron Microscopy)	Thin-section transmission electron microscopy and electron tomography for autophagy analysis in myoblasts and myotubes derived from iPS cells. Formats: digital images (.tif, .jpeg), proprietary formats (.emd, .ser).	SAMSARA, TECHNION, INSERM	1 TB
WP1	Numerical Data (Autophagy and Phenotypical Analysis)	Bioassays for autophagy markers, desmin function/aggregation, and additional phenotypic commonalities. Data in proprietary formats, summarised in Excel (.xlsx) and statistical outputs (.csv, .sav).	SAMSARA, TECHNION, INSERM, UC, APHP	50 GB
WP1	Transcriptomics & Proteomics	RNA sequencing for gene expression and Protein expression analysis via LC-MS/MS in cellular models (.fastq, .raw, mzML, .csv, .xlsx, .pptx).	CECS, UC	1 TB
WP1	High Throughput Screening Data	Data from screenings of repurposable drugs on autophagy-related defects, including raw data and processed data for hit identification. Formats: screening software formats, Excel summaries (.xlsx).	CECS, SAMSARA	50 GB
WP1	Textual Data (Documentation)	Ethical approvals, consents, meeting minutes, and reports. Formats: text files (.txt), Word documents (.docx), PDF files (.pdf), Excel spreadsheets (.xlsx).	All partners involved in WP1	30 GB
WP2	Software Development	Generated software (ETL, Machine Learning & statistical analysis pipelines), which is data in itself, will be generated and stored using a software management platform (issue tracking, time tracking, reporting, file locking, versioning, etc.).	Kantify, CECS	1 TB
WP2	Training Data	Data collected through ETL pipelines aggregating databases like UniProt, Protein Databank, Pubchem, PubMed, DailyMed, etc. (.sql). and for model training and statistical analysis purposes. This data feeds hit, ADMET and indication prediction models.	Kantify	2 TB
WP2	ML Model Weights	Data from training ML models for hit, ADMET, target & indication prediction (.csv, h5, .parquet). Various pre-trained and/or finetuned embedding models (disease, indication, drug, protein representations).	Kantify	1 TB
WP2	Omics Data Analysis	Analysis of omics data from WP1 for differential expression in NMDs, pathways analysis to identify shared drug targets (.csv, .xlsx, .pptx).	Kantify, CECS	20 GB
WP2	Target Discovery Analysis	AI-based therapeutic target identification from screenings and OMICS data (.csv, .xlsx, .pptx).	Kantify	100 GB
WP2	Indication Discovery Data	Indication prediction data from ML algorithm trained on drug/indication pairs (.csv, .xlsx, .pptx).	Kantify	100 GB
WP2	Predictive Analysis for Drug Repurposing	Predictive analysis data for identifying additional indications for repurposable drugs. Utilises newly fine-tuned Zepto.cure model to predict novel indications based on phenotypic rescue efficacy (.csv, .xlsx, .pptx).	Kantify	100 GB
WP3	Proteomic and Transcriptomic Data	RNA sequencing for gene expression and Protein expression analysis via LC-MS/MS in cellular models of additional diseases proposed in WP2 (.fastq, .raw, mzML, .csv, .xlsx, .pptx).	CECS, UC	500 GB
WP3	Biomarkers Validation Data	Data validating shared biomarkers between the additional diseases and the 5 NMDs. Includes results from immunofluorescence, western blot, qRT-PCR, RNASeq, and proteomic analyses (.csv, .xlsx, .pptx, .tiff, .jpeg).	CECS, UC	500 GB
WP3	Dose-Response Data	Results from dose-response experiments for phenotype rescue using repurposed drugs. Includes data from immunofluorescence, western blot, and qRT-PCR analyses to investigate the mechanism of action of drugs (.csv, .xlsx, .pptx, .tiff, .jpeg).	UC, CECS	100 GB
WP4	NME Selection Data	Results from virtual high throughput screening of ~12 million synthesizable compounds for binding properties, off-target effects prediction, and ADMET property evaluation (.csv, .xlsx, .pptx).	KANTIFY	100 GB
WP4	ADMET Properties Data	Data from analysis using ML algorithms (Zepto.ward) to evaluate ADMET properties of selected repurposable drugs and NMEs (.csv, .xlsx, .pptx).	KANTIFY, UC, CECS	50 GB

DREAMS project. Funded by the European Union under 101080229-2. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union (EU) or European Research Executive Agency (REA). Neither the EU nor REA can be held responsible for them

Funded by the European Union

WP4	Combination Therapy Analysis Data	Data from evaluating combination therapies using hit prediction algorithms (Zepto.hit) for off-target effects and potential synergy of drug actions (.csv, .xlsx, .pptx).	KANTIFY, UC, CECS	50 GB
WP4	In Vitro Compound Efficacy Data	Data from screening 200 compounds (repurposable drugs and NMEs) for efficacy and cell viability in skMC-iPSC lines. Includes quantitative data on primary, secondary readouts, and toxicity tests (.csv, .xlsx, .pptx, .tiff, .jpeg).	CECS, UC, KANTIFY	100 GB
WP4	Drug Ranking and MOA Investigation Data	Ranking data based on clinical, commercial, and regulatory potential of screened drugs. Mechanism of action (MOA) data for top candidates, including data from autophagy, desmin, and related phenotype analyses (.csv, .xlsx, .pptx).	KANTIFY, INSERM, TECHNION, SAMSARA	10 GB
WP5	Toxicity and Safety Data	Data from preliminary analysis of potential renal, hepatic, and pancreatic toxicity using serum biomarkers and histopathology (.csv, .xlsx, .pptx, .tiff, .jpeg).	INSERM, TECHNION	100 GB
WP5	Autophagy Analysis Data	Data from testing the effect of drug candidates on autophagy in mouse models, using immunoblot, immunofluorescence staining, and electron microscopy. Includes images for immunofluorescence and electron micrographs (.tiff, .jpeg), Western blot data files, and quantitative analysis of autophagic markers (.csv, .xlsx, .pptx).	INSERM, TECHNION	100 GB
WP5	Muscle Function and Structure Data	Data on skeletal muscle function and structure from in vivo tests, including body composition, muscle strength, physical performance, and muscle force generation. Formats include quantitative data files (.csv, .xlsx, .pptx) for performance tests, and digital images (.tiff, .jpeg) for muscle fibre architecture (histology, electron microscopy).	INSERM, TECHNION	100 GB
WP6	Patient Engagement Feedback	Feedback and review documents from the Patient Advisory Board (PAG), including ethical consent forms, patient-reported outcome measures, and clinical trial protocol reviews (.docx, .pdf)	AFM-TELETHON, APHP	10 GB
WP6	Clinical Trial Protocol Development	Development documents and feedback for the innovative clinical trial protocol, including statistical analysis, seamless and adaptive design discussions (.docx, .pdf, .xlsx)	AFM-TELETHON, APHP, KANTIFY	10 GB
WP6	Clinical Trial Outcome Prediction Tool Data	Data from the AI-based tool developed by KANTIFY for evaluating clinical trial design parameters, including data analyses and predictions on successful trial chances (.csv, .xlsx)	KANTIFY	100 GB
WP6	Regulatory and Pharmaceutical Feedback	Documentation and communications with regulatory agencies (EMA) and pharmaceutical representatives (EFPIA), including meeting summaries, feedback, and recommendations (.docx, .pdf, .xlsx)	AFM-TELETHON, APHP	10 GB
WP6	Final Clinical Trial Protocol	The final version of the clinical trial protocol, including all stakeholder reviews, recommendations, and validations (.docx, .pdf)	AFM-TELETHON, APHP, KANTIFY	10 GB
WP7	Communication and Dissemination Plan	Dissemination and communication strategy documents including impact-enabling approach, activities, channels, tools, and timing (.docx, .pdf)	Zabala, All partners	25 GB
WP7	Visual Identity and Website Content	Project visual identity elements and website content including news updates, partner interviews, and webinar recordings (.jpg, .png, .mp4, .html)	Zabala, All partners	15 GB
WP7	Social Media Content	Content for social media platforms including posts, videos, and interviews (.jpg, .png, .mp4)	Zabala	20 GB
WP7	Event and Webinar Materials	Materials from events and webinars including presentations, recordings, and promotional materials (.pptx, .docx, .mp4)	Zabala, All partners	30 GB
WP7	Stakeholder Engagement Reports	Reports on stakeholder dialogue and outreach, including joint dissemination activities, workshops, and knowledge exchange (.docx, .pdf)	Zabala, All partners	40 GB
WP7	Networking Activity Reports	Reports and communications from networking activities with related projects and initiatives, regulatory bodies, and industrial stakeholders (.docx, .pdf)	CESC, AFM, KANTIFY, All partners	1 GB

DREAMS project. Funded by the European Union under 101080229-2. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union (EU) or European Research Executive Agency (REA). Neither the EU nor REA can be held responsible for them

Funded by the European Union

WP7	Exploitation and IPR Strategy	Intellectual Property Directory, Exploitation Plan, and IPR Management Plan including business modelling and market analysis (.docx, .pdf, .xlsx)	Kantify, All partners	1 GB

3 Data sharing and Intellectual Property

The consortium is committed to a nuanced data management approach, balancing the imperative of sharing findings with the scientific community against the protection of intellectual property. Data sharing will be preceded by thorough consultations among consortium partners, and specifically its Innovation and Exploitation Board and Project Coordinator, to determine the suitability of datasets for publication or internal use. This strategy ensures that while contributing valuable insights to the broader scientific community, the proprietary interests of the consortium members are safeguarded.

Data Utility Outside the Project

Research Community: The generated data holds immense value for scientists engaged in stem cell research, muscular diseases, and drug discovery. It offers a rich resource for further investigation or as a benchmark for comparative studies, facilitating new discoveries and enhancing the understanding of disease mechanisms.

Pharmaceutical Industry: Pharmaceutical companies, particularly those invested in drug development, can leverage the project's findings to inform drug repurposing initiatives or identify novel therapeutic targets. This could accelerate the development of effective treatments for NMDs and other related conditions.

AI and Computational Biology Fields: The comprehensive datasets produced by DREAMS are also instrumental for researchers and developers in AI and computational biology. The detailed biological data can be used to train or refine AI models, contributing to advancements in the analysis and interpretation of complex biological information.

The consortium will establish an **Innovation and Exploitation Board (IEB)** in charge of defining the protection and exploitation strategy of the generated data, and the DREAMS project results in general, as well as defining and managing the IP of the project results. This Board will be composed by one member from each project partner, and its main role will be to provide advice and support to General Assembly on matters related to publication, intellectual property, commercialization strategies, and market analysis.

4 FAIR data

The DREAMS consortium is dedicated to adhering to the FAIR data principles, ensuring that data generated and utilised throughout the project is *findable, accessible, interoperable, and reusable*. This commitment spans various data types and sources, including both partner-generated and shared datasets, as well as publicly released data. The consortium navigates complex landscapes of intellectual property (IP) rights, diverse data management systems, and distributed data generation and storage practices.

4.1 Making data findable, including provisions for metadata

The DREAMS consortium prioritises the findability of data across three distinct layers:

- At Individual Partners: Partners implement best practices within their own data management systems, ensuring local data is findable through detailed metadata and persistent identifiers where applicable.
- Consortium-level Shared Data: KANTIFY is responsible for the implementation of a centralised data storage system specifically designed for consortium-wide shared data. This system features advanced

DREAMS project. Funded by the European Union under 101080229-2. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union (EU) or European Research Executive Agency (REA). Neither the EU nor REA can be held responsible for them

APIs, access rights management, and automated duplication across service providers, supporting both the autonomy of individual partners and the collective utility of shared data. Data within this system is made findable through the assignment of standardised names and automatically generated metadata.

- Openly Shared Data: For data designated for public release, the project leverages trusted repositories like NCBI's GEO and ProteomeXchange. These platforms, familiar within the bioinformatics community, provide long-term preservation and accessibility, including the assignment of Digital Object Identifiers (DOIs) to datasets.

Metadata across these layers will include technical (format, file size, version, source software or tools), experimental (descriptions of experimental design, methodologies, cell lines, conditions, and treatments). and descriptive fields (title, source, date of creation or modification, authorship), adhering to standards such as MIAME and MIAPE for specific data types.

4.2 Making data securely accessible

Ensuring the accessibility of data generated by the DREAMS project encompasses a multifaceted approach, addressing the storage, management, and dissemination of data across different stages of the project. This strategy takes into account the varying needs for access, the imperative of adhering to intellectual property rights, and the commitment to open science principles.

Consortium-level shared data

KANTIFY will implement and manage a centralised data storage solution that significantly enhances the consortium's ability to manage and share data efficiently:

- Bucket Storage System: Utilising a bucket storage system, organised with standardised naming conventions and automatically generated metadata, facilitates a structured and easily navigable data environment. This system employs the AWS S3 protocol, ensuring robust and secure data exchange capabilities.
- Access Control: Managed by KANTIFY administrators, the system features sophisticated access controls, enabling precise management of who can view or manipulate different datasets. This level of control is crucial for addressing privacy, legal, and IP concerns within the consortium.
- Enhanced Data Security: The solution includes features like automated replication across multiple service providers, guaranteeing data integrity and mitigating the risk of data loss.

Repository and data accessibility

For data earmarked for public release, the DREAMS project leverages trusted repositories recognized within the bioinformatics community:

- Trusted Repositories: NCBI's Gene Expression Omnibus (GEO) for RNAseq data and ProteomeXchange for proteomics data are among the repositories selected for their robust mechanisms for long-term data preservation and accessibility.
- Persistent Identifiers: These repositories assign unique identifiers, such as GEO Series accession numbers and ProteomeXchange IDs, to each dataset. This ensures that datasets are not only easily accessible but also citable, enhancing their utility within the scientific community.
- Engagement with Repositories: Prior to data submission, the consortium will engage with these repositories to ensure compliance with data format, metadata standards, and submission guidelines. This proactive approach maximises the efficiency of the data deposition process.

Data Sharing and IP Considerations

The DREAMS project navigates complex landscapes of data sharing and intellectual property:

DREAMS project. Funded by the European Union under 101080229-2. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union (EU) or European Research Executive Agency (REA). Neither the EU nor REA can be held responsible for them

Funded by the European Union

- Open Accessibility: While the project aims for the broadest possible accessibility of data, certain datasets will have restricted access due to privacy, legal, or intellectual property considerations. This differentiation ensures that sensitive data is handled appropriately while promoting open access where feasible.
- Embargo Periods: To balance the interests of publication and IP protection with the principles of open science, embargo periods may be applied. Such periods will be kept as brief as practicable, typically not exceeding 12 months from the data collection phase's conclusion.
- Standardised Access Protocols: Data will be accessible through free and standardised access protocols provided by the repositories, with additional mechanisms in place for datasets under restricted access conditions. For sensitive or personal data, a data access committee may be established to evaluate and approve access requests, ensuring that data sharing complies with ethical standards and legal requirements.

4.3 Making data interoperable

Interoperability is a cornerstone of the DREAMS project, ensuring that the diverse datasets generated and utilised are not only accessible but also compatible across various scientific disciplines and research domains. This commitment facilitates seamless data exchange and re-use, enhancing collaborative research and innovation.

Standards and Best Practices for Interoperability

The DREAMS consortium is dedicated to adhering to widely recognized standards, vocabularies, and formats to maximise data interoperability:

- Adherence to Standards: For omics data, the project will follow community-endorsed standards such as MIAME for microarray experiments and MIAPE for proteomics data. These standards provide comprehensive guidelines for data reporting, ensuring that datasets are comprehensively described and easily interpretable across different research fields.
- Common Data Formats: Utilising common data formats such as FASTQ for sequencing data, CSV and Excel for numerical data, and TIFF and JPEG for imaging data, facilitates compatibility and ease of use across various software platforms and research communities.
- Methodological Consistency: The project will employ standardised methodologies for data generation and processing, ensuring that datasets generated by different partners or at different stages of the project can be integrated and analysed cohesively.

Project-specific Ontologies and Vocabularies

In instances where the generates unique ontologies or vocabularies specific to its research goals:

- Mappings to Common Ontologies: The project will provide mappings to more widely used ontologies, ensuring that the unique aspects of the project's data can be understood and integrated within broader scientific contexts.
- Publication of Ontologies: Any project-specific ontologies or vocabularies developed will be openly published, allowing for their reuse, refinement, or extension by the broader research community. This openness fosters collaborative improvements and the broad applicability of the project's innovations.

Ensuring Clarity and Comprehension

Given the interdisciplinary nature of the DREAMS consortium, effective communication channels are essential:

DREAMS project. Funded by the European Union under 101080229-2. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union (EU) or European Research Executive Agency (REA). Neither the EU nor REA can be held responsible for them

Funded by the European Union

- Data Documentation and Metadata: Comprehensive documentation of data, including detailed metadata, is a critical expectation for all data generators within the consortium. This documentation will cover descriptive, technical, and experimental aspects of the datasets, ensuring clarity and facilitating understanding across disciplines.
- Communication Channels: Data transfers within the consortium will be accompanied by direct communication between partners. This interaction is vital for clarifying the nuances of the data, addressing any questions, and ensuring that datasets are utilised effectively and appropriately by receiving parties.
- Qualified References: Datasets generated or utilised by the project will include qualified references to other related data, whether from within the project or from external research. This practice enriches the context of the data, enhancing its utility and the insights that can be derived from it.

By implementing these strategies, the DREAMS project ensures that its data is not only interoperable within the consortium but also aligns with broader scientific efforts, facilitating advancements in understanding and treating neuromuscular diseases. This approach underscores the project's commitment to collaborative research and the open sharing of knowledge.

4.4 Increase data re-usability

Documentation to validate data analysis and facilitate data re-use

To maximize data re-use, we'll make multiple protocols public and publish detailed processes of analyses in papers. Data published in scientific papers will be freely available, ensuring usability by third parties during and after the project. Published Proteomics data will be accessible online via ProteomeXchange, and Transcriptomics data will be available on GEO under the GSExxx accession number. Programming pipelines will be made public on GitHub as often as possible, facilitating validation and re-use.

Data availability

The data produced will be published according to the usual characteristics of data dissemination in science (articles; conferences, etc.).

Data subject to intellectual property or specific expertise will not be shared or on a case-by-case basis after advice from the Innovation and Exploitation Board.

Provenance of the data

The provenance of our data will be thoroughly documented in alignment with the highest standards in the field. For bioinformatics analyses, including RNA sequencing (RNAseq) and proteomics, we will adhere to the Minimum Information for Publication of Quantitative Real-Time PCR Experiments (MIQE) guidelines and the Minimum Information About a Proteomics Experiment (MIAPE) standards, respectively. These frameworks ensure comprehensive traceability of our data and analyses, enhancing their reliability and reproducibility.

All biological experiments will strictly follow Standard Operating Procedures (SOPs) that we will make publicly available. These SOPs are designed to align with the FAIR principles, ensuring that our methodologies are fully transparent and can be accurately replicated or built upon by the broader scientific community. This meticulous approach to documenting the provenance and methodology of our data underpins the integrity of our research and facilitates its contribution to scientific advancement.

Data quality assurance processes

The transcriptomics data are subject to standardised quality control at the dedicated platform. The mRNAs in

DREAMS project. Funded by the European Union under 101080229-2. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union (EU) or European Research Executive Agency (REA). Neither the EU nor REA can be held responsible for them

Funded by the European Union

the samples are checked for their quality (RIN score >7; control of the level of degradation); the sequencing libraries produced are subject to quality control (DNA profile: fragment size, concentration, amplification level); sequencing is controlled internally according to the equipment used (proportion of bases sequenced, PHRED score, Q score (Q20 equivalent to a sequencing error every 100 bp).

Bioinformatics analysis undergoes quality control (fragment alignment percentage, ribosomal RNA, adapter) and encoding quality assessment. Data are filtered by quality and aligned to the target species' genome to identify potential contaminations. PCA mapping and a correlation matrix assess variation in biological replicates (reproducibility, variability). Software used includes FastQC (v0.11.2), Prinseq (v0.20.4), cutadapt (v1.16), and samtools (v0.1.19).

All equipment used is subject to calibration and regular maintenance carried out by the various managers to guarantee their proper functioning. Deviations during data acquisition are systematically documented in laboratory notebooks.

5 Allocation of resources

Data management responsibilities and resources

CECS and Kantify will be in charge of the data management in the project (Task 8.3 of the Gant Agreement DoA). Resources for data management have been allocated in the project in the corresponding working packages to avoid any potential problem with the use and maintenance of data that are crucial to guarantee the success of the project. All partners have adequately estimated the necessary budget to generate and manage the data related to their tasks in the project.

Costs to making the data or other research outputs FAIR

As much as possible, RNAseq and proteomics data will be published along with the research paper in scientific journals, which is in most cases, free of charge in databases like GEO and ProteomeXChange. Eventually, indirect cost will be for resources for formatting, curating, and adding metadata before publishing. These databases are backed by entities like NCBI and EBI, ensuring a robust solution for data security.

Researchers who will be in charge of publishing in scientific journals will also have the following tasks regarding data included in the publication: preparation, formatting, and submission of data to repositories and journals.

Duration of data preservation and implemented resources

The expected duration for data preservation in GEO and ProteomeXchange is expected to data availability for an extended period (e.g., over 15 years for GEO). This long-term perspective facilitates future research and innovation by providing a reliable foundation for data re-use and meta-analyses.

6 Ethics and data security

Ethical considerations and adherence to legal frameworks are paramount in the DREAMS project, particularly when handling sensitive data such as patient information. These aspects are critically evaluated to ensure the project's data management practices comply with both ethical standards and legal requirements, including the General Data Protection Regulation (GDPR).

DREAMS project. Funded by the European Union under 101080229-2. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union (EU) or European Research Executive Agency (REA). Neither the EU nor REA can be held responsible for them

Funded by the European Union

Ethics and Legal Issues Impacting Data Sharing

- Patient Data: The collection, storage, and sharing of patient data are subject to stringent ethical and legal scrutiny. The project recognizes the sensitivity of health-related personal data and is committed to upholding the highest standards of privacy and confidentiality. This commitment extends to ensuring that all patient data is anonymized or de-identified to prevent the possibility of re-identification, following GDPR and other relevant data protection laws.
- To prevent any ethical issues associated with the processing of patients' personal data who provided their blood samples to Généthon, we requested that patient data be anonymized at the investigator centres. Therefore, these centres will label the blood sample tubes solely with the first letters of the patient's first and last names, along with the participation number, before sending them to Généthon.
- Genethon receives de-identified blood tubes and uses them to prepare PBMC cells, which are then shared with the CECS. The latter reprograms the PBMC cells into induced pluripotent stem cells (iPS). These cell lines, which will be shared with multiple partners, are appropriately de-identified.
- To guarantee the collection of all patient blood samples for the project in strict adherence to French and European legislation and the rights of the involved patients, we submitted the complete study dossier (comprising all necessary documents) through the research portal for studies involving the human body (SI RIPH 2G) on April 5, 2024, as part of a RIPH3-class study.
- Informed Consent: In scenarios involving the collection of personal data, such as blood samples, the DREAMS project will obtain informed consent from all participants. This consent will explicitly cover data sharing and long-term preservation, ensuring that participants are fully aware of how their data will be used and safeguarded. The consent process will also address the potential for data to be used in future research, underlining the measures in place to protect participant privacy and data integrity. Given that we have clearly stated in the consent form that patient data will be de-identified, meaning that the link between their name and this alphanumeric identifier will be retained in a secure computerised database accessible only to the principal investigator. Their medical data will thus be de-identified for all other researchers. It is also protected by medical confidentiality, to which the medical team is bound, as well as professional secrecy, to which the sponsor's staff and health authorities are subject. Therefore, their personal data will not be disclosed in any way during the publication of research findings.
- Legal Frameworks and GDPR Compliance: The project is committed to full compliance with GDPR and other applicable legal frameworks governing the collection, use, and sharing of personal data. This includes implementing robust data governance policies, ensuring data minimization, and securing data against unauthorised access. The project will also adhere to principles of transparency, providing clear information to participants about the use of their data and their rights under GDPR.

The consortium will establish an **Ethics Board** in charge of monitoring the ethical concerns of the project. This Board will be a consultative body, composed by an independent consultative expert, together with one member from each project partner, and its main role will be to assist the DREAMS consortium in ethical aspects of the project.

DREAMS project. Funded by the European Union under 101080229-2. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union (EU) or European Research Executive Agency (REA). Neither the EU nor REA can be held responsible for them

Funded by the European Union