IDEA-FAST

Identifying Digital Endpoints to Assess FAtigue, Sleep and acTivities in daily living in Neurodegenerative disorders and Immune-mediated inflammatory diseases.

Grant Agreement No. 853981

WP5 – Data Management

D5.1: Data Management Plan

Lead contributor	P1 - UNEW
Other contributors	P9 – ICL P35 - Janssen

Due date	30 Apr 2020
Delivery date	30 Apr 2020
Deliverable type	R
Dissemination level	PU

Document History

Version	Date	Description
V0.1	21 Apr 2020	First draft
V0.2	23 Apr 2020	Additional content
V0.3	30 Apr 2020	Additional content
V0.4	30 Apr 2020	Content transferred to project deliverable template
V1.0	30 Apr 2020	Final version

Table of Contents

1 Abstract

This document details how the data collected, processed and generated by the IDEA-FAST project will be managed during and after the end of the project. It explains the purpose of the data collection and describes the different types of data and the associated data standards. Plans for the preservation, sharing and re-use of the data are also outlined.

2 Data Summary

2.1 Purpose of data collection

The ultimate goal of the project is to identify digital endpoints that provide reliable, objective and sensitive evaluation of activities of daily living (ADL)/disability/ health-related quality of life (HRQoL) relevant for the following neurodegenerative disorders: Parkinson's disease (PD), Huntington's disease (HD) and the following Immune mediated inflammatory disease (IMID): Rheumatoid arthritis (RA), systemic lupus erythematosus (SLE), primary Sjögren's syndrome (PSS), and inflammatory bowel disease (IBD).

The main purpose of the data collection during the project is to select a set of digital endpoints and corresponding digital technologies for the evaluation of fatigue and sleep disturbances for clinical validation and to seek qualification advice from regulatory authorities (such as the European Medicine Agency and the US Food and Drug Agency), and to explore digital correlates of other ADLs by profiling activity, biological, neurocognitive and behavioural aspects in patients with IBD, RA, SLE, PSS, PD and HD.

The large integrated clinical and device dataset established by the project will provide an invaluable resource for clinical practice, well-being of the patients, drug development, healthcare provision, health and wealth of the society and research and innovation.

2.2 Data types and formats

The project will generate and receive a large volume of very diverse data including:

- Pseudonymised participant demographic, clinical and patient-reported outcome data
- Sensor data from participants during the periods of sensor use
- Data generated from analyses of one or both of the above datasets.

The IDEA-FAST datasets will be derived from several sources:

a) the IDEA-FAST Feasibility Study and Clinical Validation Study and other activities from the Description of Actions (DoA) of the IDEA-FAST project;

b) relevant extant datasets generated by the academic and EFPIA partners;

c) Relevant datasets from other ongoing studies of the EFPIA partners as “contribution in kind”.

The estimated size of data collected during the project is less than 10GB per participant, including clinical data and device data.

All data generated in IDEA-FAST will be integrated and stored using a secure and robust process that is fully compliant with European data privacy and security legislation and standards, in particular the GDPR.

During the Feasibility Study and Clinical Validation Study, identifiable personal data will also be collected, but such data will only be held at the local recruitment centres. These identifiable personal data will not be transferred to the IDEA-FAST data management platform.

2.2.1 Clinical data

Clinical data will include demographic data and relevant clinical assessment including participant-reported outcome data. There will be some disease-specific clinical assessments and documentation of patient-reported outcomes that are tailored to the individual disease. Clinical study personnel will enter all data onto a clinical database that is managed by UCAM during the clinical studies. Some of the participant-reported outcome measures will be paper-based. The physical copy of these assessments (pseudonymised) will be stored in a locked room or cabinet at each local site for up to 25 years. We will also collect data of participants' experience in their participation of the clinical study and their use of the digital devices/technologies. Some of those data will be audio-recorded and transcribed by the researcher with all identifiable information deleted.

During the Feasibility Study and Clinical Validation Study, identifiable personal data will also be collected. Subject codes will be generated to pseudonymise the data. Both the identifiable personal data and the keys will only be held at the local recruitment centres according to the security standards required by the centres and ethics committees of the corresponding countries. Neither the identifiable personal data nor the keys to the subject codes will not be transferred to the IDEA-FAST data management platform.

The IDEA-FAST clinical database will be a relational multi-user secure database system with a public interface via which participants, sites and study administrators can enter clinical and other study related data. The database system will have the following components:

- secure public https website *https://ideafast.medschl.cam.ac.uk* hosted on University of Cambridge Clinical School MINTS domain Windows Server me-cctu IIS web server
- SQL Server database IDEAFAST on MINTS domain Windows Server me-cctu
- The public https website is the front end data entry interface for the SQL Server database. The website allows logged in users in appropriate roles to insert data.
- SQL Server database IDEAFAST on MINTS domain Windows Server me-cctu-sdhs
- me-cctu-sdhs server is located within the Clinical School secure data hosting service (SDHS). There is no internet connectivity directly to/from SDHS; secure access is achieved using 3 factor authentication and remote desktop connections.
- Data inserted into IDEAFAST database on me-cctu server is encrypted as soon as possible, transferred securely to SDHS, inserted into the copy of IDEAFAST database on SDHS, then nullified on me-cctu so that data is then only available within SDHS.
- Exceptions to the nullification of data may be specified and retained on me-cctu.

- me-cctu SQL Server insert update and delete triggers insert rows in AuditTrail table whenever study data is inserted, updated or deleted using the web front end or any other method.
- me-cctu IDEAFAST and me-cctu-sdhs IDEAFAST databases are backed up daily to networked group drives using automated methods. If a failure occurs, database administrators receive an email alert.
- The backup of me-cctu IDEAFAST database is encrypted and transferred to SDHS group drive
- Secure internal https website hosted on me-cctu-sdhs IIS web server. The website is only visible within the SDHS environment. Website is front end to me-cctu-sdhs IDEAFAST SQL Server database. Website allows logged in users in appropriate roles to insert, update and delete data.
- me-cctu-sdhs SQL Server insert update and delete triggers insert rows in AuditTrail table whenever study data is inserted, updated or deleted using the web front end or any other method
- Web based interface for IDEAFAST administrator to create and manage website users and roles on me-cctu
- *https://ideafast.medschl.cam.ac.uk* uptime constantly monitored and downtime immediately alerted to database programmers via email
- SQL Server jobs set up to send email alerts to database programmers whenever a job fails
- SQL Servers configured to log all database errors in ErrorLog table and send email alert to database programmers
- websites configured to log all web server errors to Elmah database and email alerts to database programmers

The IDEAFAST database system will include the following security features:

1. The web interface only permits usage over https encrypted protocol
2. Users can only register by invitation
3. Membership of site-team is verified by independent means (site password)
4. A full audit trail of all user access is maintained.
5. Data will be stored in the University of Cambridge Clinical School Computing Service Secure Data Hosting Environment (https://www.medschl.cam.ac.uk/research/information-governance/sdhs-security-policy/).

6. Database backups will be stored on networked drives which are themselves backed up.

7. A SQL Server scheduled job on each of the study servers will execute a full backup of all systems and study databases on each SQL Server each night.

8. *IDEAFAST* databases on both me-cctu and me-cctu-sdhs servers will implement a full backup model including transaction log backups every 5 minutes. This is intended to limit potential data loss in disaster recovery situations to no longer than 10 minutes.

2.2.2 Device data

During the clinical studies, the project will focus on capturing a full picture of the original data available through each device / application, in order to allow for a rich investigation into mapping measurements to the concepts of interest. Therefore, the device and application data will be stored in a number of different data formats as encrypted file containers together with a small number of meta-data fields (study location, participant ID, device ID, start and end date/time for recording) that will be stored in JSON or XML. Together with WP5, we will work to develop the format of these containers. For example, it may be wrapped in a 7z container with local encryption through a command-line script or small helper-tool before transfer. The original file formats as exported from the devices are MATLAB file format for MoveMonitor, CSV files for Fibion, SQLite, dat and wav files for the VTT bed sensor, custom binary dat files for AX6, etc.

In short: the device data will be in several file formats, for example CSV. The intention will be to package this unprocessed data into an encrypted container (current candidate format is *.7z with secure AES256 encryption) together with associated JSON metadata. These packages will be uploaded to the WP5 data management portal via secure HTTPS transport. Device data, once successfully uploaded, will be wiped off the host devices, and these will be reset back to pre-participant settings.

2.2.3 Extant data

Extant datasets are collected for several applications:

- Datasets provided by device manufacturers are used to compare the raw signal quality, or extracted features, metrics and endpoints with clinical gold standards when available (ex: sleep measurement headband compared with polysomnography), and to assess their variability and signal to noise ratio. Such properties are studied within WP4 to evaluate device performance and to inform WP7 in order to model the error, noise and variabilities in statistical power studies and design.

- Extant clinical +/- device datasets provided by our consortium partners for the specific diseases of interest or relevant publicly available datasets will be studied. They are typically used to assess the effect of the disease, and possibly of treatments, on the metrics which will eventually be used to compute the endpoints.

Once the device-related digital endpoints have been defined, a preliminary work on their variability and predictiveness will be conducted to compute a first, rough clinical trial power study. And initial modelling of the variability of expect longitudinal data which will be conducted during the feasibility and clinical validation studies.

The transfer of these extant datasets to the IDEA-FAST data management platform and the use of such datasets will be covered by appropriate Data Sharing Agreements between the contributor and the consortium. The ownership of such datasets remains under the contributor.

2.2.4 EFPIA Contribution-in-kind datasets

The consortium has set up an EFPIA Data Monitoring Board consists of representatives of the EFPIA partners. Its role is to monitor the progress of clinical data contribution as committed by EFPIA/Associated Partners, to serve the objectives of the project. These datasets are owned by the contributing EFPIA/Associated Partners. The intended use of these datasets includes, but not limited to, inform the design of the clinical validation studies, to provide additional relevant information for the selection of candidate digital endpoints and to assess the potential impact of the candidate digital endpoints in therapeutic development trials. The use of individual EFPIA partner datasets will be coordinated by the EFPIA Data Monitoring Board, and may involve the leads/co-leads of WP2, WP3, WP4, WP5, WP7 and WP8, as well as the representative of the EFPIA partners contributing the dataset.

2.3 Re-use of existing data

We will re-use relevant extant datasets to assess properties of the digital sensor data at the project start. For each digital endpoint, evaluate existing methods or developed matching state-of-the-art algorithms on existing public (large-scale) datasets on different tasks.

2.4 Data utility

The large integrated clinical and device dataset established by the project will provide an invaluable resource for clinical practice, well-being of the patients, drug development, healthcare provision, health and wealth of the society and research and innovation.

3 FAIR Data

3.1 Making data findable, including provisions for metadata

3.1.1 Participant data

For the clinical studies, participants can choose if they want to:

a) Provide their data only for IDEA-FAST

b) Provide their data for IDEA-FAST and alternative sleep/fatigue or disease-specific research.

Personal data (i.e. code table) and biosamples collected will be kept for up to 25 years.

3.1.2 Naming conventions

We will design and develop IDEA-FAST data standards to define standards for data integration, analysis, storage and sharing (including secondary use of data to maximise impact and exploitation). Both clinical and device-specific data standards dictionaries will be designed and developed during the project, consisting of metadata templates, control vocabularies and data. Datasets will be curated against these data standards before being integrated into the data management platform.

3.1.3 Search keywords

Keyword-based search function will be provided by the IDEA-FAST data management platform to make sure the datasets and metadata will be searchable for re-use.

3.1.4 Version numbers

Clear version numbers of the datasets will be provided through the IDEA-FAST data management platform. Semantic versioning will be used for all software developed during the project. Version control of all source code will be enabled by git/Github. Changelogs that contain list of notable changes for each version of the datasets/software will be automatedly generated to track changes.

3.1.5 Metadata

Data produced in the project will be integrated and stored in the IDEA-FAST data management platform. Metadata describing the clinical data and device data will be created during the project. The platform will allow metadata to be attached to the datasets, with user-friendly search functions to improve the findability of the data.

The project will define clinical and device-specific data standards consisting of metadata templates, control vocabularies and data dictionaries. We will also make available the bibliographic metadata (include a persistent identifier) that identifies the deposited publication via the IDEA-FAST scientific publication repository.

3.2 Making data openly accessible

3.2.1 Open Access

Data-sharing will be enshrined in the Consortium Agreement, identifying partner rights and timescales. It will be a fundamental premise that no data is accessible to any party, internal or external, without full ethical sanction and that only de-identified and pseudononymised data will be shared within the Consortium. The project will exploit a professional data-handling structure for the secure handling of all project data, and will secure, share, curate and exploit these data.

The consortium has decided to opt-out from the “Open Access to Data Pilot for IMI projects (no applicability of Art 29.3 of the Grant Agreement), and hence no mandatory open access to data obligation will be put in place (in the frame of the regular application of rule on Open Access to Research Data for IMI projects which did not opt out of the Open Access to Data Pilot).

An important subset of the data generated in the project may or will be needed for filing regulatory documents, so preliminary sharing of data outside of the consortium would hinder the exploitation of the project results and hence the overall objectives of IDEA-FAST. In addition, the publication of clinical data may cause issues under personal data laws.

Furthermore, subsets of the IDEA-FAST dataset also include extant datasets from consortium partners or other sources, and “contribution-in-kind" datasets from EFPIA partners. The ownership of such datasets remains under the respective contributors which has their own data access policies.

While opting out of the Open access to Data Pilot, the consortium is prepared to offer an alternative access to data, outside of the regular framework of an Open Access to Data under Art 29.3 of the Grant Agreement.

During the life-time of the project, the IDEA-FAST steering committee will be responsible for determining which subsets of the IDEA-FAST will be Open Access and which subsets will be withheld from Open Access and the timescale and the reasons for such decisions. The governance structure of IDEA-FAST dataset beyond the lifetime of the project will be finalised as part of the Sustainability and Exploitation Plan.

3.2.2 Data repository

The IDEA-FAST data management platform (open-source) will provide secure large-scale data storage to host the integrated data as well as the raw device data and clinical/device metadata. The platform will also provide a web-based user interface with secure access control that allows data analysts to access the pseudonymised data for data analysis. It will be a fundamental premise that no data is accessible to any party, internal or external, without full ethical sanction and that only pseudonymised data will be shared within the Consortium.

Documentation and a user manual of the IDEA-FAST data management platform will be publicly accessible from the project Github repository, as will the source code for all software developed during the project. Two-factor authentication for access to the platform will be in place, with user-friendly password policies.

3.2.3 Data Access Sub-Committee

We will establish a Data Access Sub-Committee (DASC) to review requests for access of the IDEA-FAST datasets generated by the IDEA-FAST consortium as part of the Description of Action (largely (but not exclusively) related to the data from the clinical studies) which are not Open Access. The membership of the DASC will include different expertise and will include internal and external members as well as patients. The IDEA-FAST steering committee will be responsible for the appointment of the DASC. The key purpose of the DASC is to ensure the requests for access is scientifically justified, legally and ethically acceptable and have no relevant conflicts of interest. We will develop a standardised data request form for all requests for data access.

The recommendations of the DASC will then be considered by the IDEA-FAST steering committee. If the request is approved, appropriate Data Sharing Agreement will be arranged with the requesting party before data is released. Where appropriate, license fees may be charged to the requesting party.

The Data Management Team will develop robust mechanisms to control and monitor users of the IDEA-FAST datasets. The levels of access will be appropriate to the requirement of the users.

3.3 Making data interoperable

Data standards will be developed taking into account clinical knowledge combined with devices and measurements. Based on the data standards, data standardisation and integration pipeline will be built to curate the data against the defined standards. The IDEA-FAST data management platform will provide the project with a hub of curated data harmonised with clinical information together with the measurements based on a defined data standard.

Device data analysis pipelines include steps such as validation, feature extraction, and visualization and analysis. They will use as input the data (‘raw data’) coming from different devices according to the manufacturers’ own specifications. These data include various formats, such as, csv, json, text, and binary formats such as edf. The data analysis steps form a modular pipeline, generating intermediate and final output files from different processing steps. These intermediate files are in an open and fully documented format allowing the community to further use the results for additional research purposes. The routines for reading and writing the intermediate files will be fully documented and available with (Python) source code. Final results (graphs, reports) can be delivered both as text/csv files and pdf documents.

The data standards developed during the project will include metadata templates, control vocabularies and data dictionaries. The clinical data generated by the project will compliant with well-established standards such as Clinical Data Interchange Standards Consortium (CDISC). It is of relevance to have the clearance of Regulatory Agencies about ontology and data-format that are aligned to the current

standards. For this reason, the EPFIA, SME and academic teams will explore the inclusion of updated standard approaches by exploring opportunities from precompetitive consortia (e.g. MEDICS).

Most interoperability concerns will be addressed in post-processing, while the clinical studies focus on collecting original source data reliably and securely.

For the data generated during the clinical studies, the interoperability will be focused on the format of the JSON meta-data headers aligned with each encrypted data packet. Converters will then be developed in collaboration with WP4 to assure that further processing is facilitated. For these efforts, a large variety of formats need to be unified. Raw data in disparate formats must be processed into a single unified format. This is an ongoing concern between WP3, 4 and WP5 software solutions and will be a major part of our work. Responsibility for this is shared between WP3, 4 and WP5.

Successful development of these conventions will be in accordance with any database standards set in the data management platform, making use of established eHealth / mHealth standards as far as possible. It is important to note these headers will be annotated with pseudonyms only and sharing will remain limited to a small circle of authorised project members until further processing.

Data collected from the Feasibility Study and Clinical Validation Study will be curated against the defined standards, and then integrated across different domains (e.g. demographics, diagnosis, laboratory tests, medications), across studies and across multiple digital devices, on the data management platform for data sharing and analysis.

3.4 Increase data re-use (through clarifying licenses)

3.4.1 Plan for data re-use

As part of our Dissemination Plan, we will include plans to raise awareness of the availability of the IDEA-FAST datasets to all stakeholders. As described earlier, the IDEA-FAST steering committee will determine the conditions of data sharing with external users of the datasets including license fees.

For reasons described earlier, we will not make the IDEA-FAST datasets Open Access. The IDEA-FAST steering committee will be responsible for reviewing the accessibility of the IDEA-FAST datasets/subsets to third parties and the appropriate embargo periods. Such review will be carried out at least annually. Beyond the lifetime of the project, we will set up an IDEA-FAST governing body (details to be confirmed in our Sustainability and Exploitation Plan), which will continue to review the accessibility of the IDEA-FAST datasets/subsets as well as the reusability of the datasets.

3.4.2 Quality assurance processes

From a data management perspective, we will develop SOPs to guide clinical and sensor data acquisition and implement data quality control processes to assess data quality and discovery data inconsistency. From a data processing perspective, methods include biomedical signal processing methods, and statistical and probabilistic approaches (e.g. linear and non-linear filtering, matched filters, PCA, and ICA) that will be used to assess and control data quality. Advanced methods such as AI-based time-series analysis algorithms to deal with imperfect data will be employed.

4 Allocation of resources

The key resources and costs associated with maintaining the datasets will be established in the Finance and Sustainability Plan (developed as part of the Sustainability and Exploitation Plan led by WP9). The IDEA-FAST budget includes some funding for Open Access publications. WP5 will be responsible for data management in the project and a Data Management Team will oversee the

maintenance of the datasets beyond the project's lifetime. The Sustainability and Exploitation Plan will also include the establishment of a governing board of all IDEA-FAST assets including the IDEA-FAST datasets, which will have oversight on data sharing and preservation beyond the lifetime of the project.

5 Data security

Data generated during the project will be mirrored at all times to prevent accidental loss. Device data, once successfully uploaded, will be wiped off the host devices, and these will be reset back to pre-participant settings. The device data will be associated with various meta-data headers and encrypted before transport. All headers will only employ pseudonyms and unencrypted headers required for establishing secure transport will not contain any study-data or metadata. Data that is manually uploaded by study centre staff will also be synched to a fully encrypted local HDD as a backup precaution.

End-to-end encrypted data transfer and encrypted storage will be used for bespoke system development, with GDPR compliance. There will be two-factor authentication for platform access, with user-friendly password policies, and a limited number of administrators / super-users.

Curation methodology will be established during the project and will be used to inform the addition of datasets after the project ends. Long term storage solutions and data access will be determined during the lifetime of the project. In addition, suitable repositories for long-term data storage will be identified in the duration of the project (e.g. eTRIKS https://www.etriks.org/, Elixir BioStudies https://elixir-europe.org/platforms/data/elixir-deposition-databases).

6 Ethical aspects

The impact of ethical and legal issues on reuse of the IDEA-FAST datasets varies depending on the data subsets. As mentioned in earlier section, there are three main data subsets.

For extant datasets from academic and EFPIA partners and the "contribution-in-kind" datasets from the EFPIA partners, these data subsets are owned by the contributing partner, who are also responsible for the approval of the sharing of the data, and how the datasets will be shared. In addition, the extent to which the datasets can be shared for reuse will depend on the terms of the specific consents provided by the study participants. The relevant data protection legislations must also be taken into consideration.

For the datasets generated as part of the IDEA-FAST Description of Action, one of the tasks of WP8 is to develop appropriate mechanisms to maximise the potential for data sharing and reuse within the current ethical, legal and data protection framework. For instance, in the preparation of the Information Sheets for study participants, we will ensure the participants are provided with clear information on the plan for data sharing and reuse by putting together a model Informed Consent Form (Deliverable 8.1).

Written informed consent will be obtained from all participants of the clinical studies of the IDEA-FAST project. The Informed Consent will contain clear statements on long-term preservation of the data and the plan for future data sharing, including any data protection issues such as the potential to share data with countries with data protection legislation different from the European Union.

7 Conclusions

This Data Management Plan describes the overall approach to managing the data collected, processed and generated throughout the lifetime of the IDEA-FAST project. It is a working document and will

be updated whenever any changes arise.