D3.1 – System Requirements Specification (including definition of data acquisition, transmission, integration, processing, analysics and governance requirements).

Mobilise-D

Connecting digital mobility assessment to clinical outcomes for regulatory and clinical endorsement

Grant Agreement No. 820820

[WP3 – Data Management Platform]

Lead contributor	P20 – UKER - Heiko Gaßner
Other contributors	P20 - UKER - Jochen Klucken, P9 - UCD - Brian Caulfield, David Singleton, P1 – UNEW - Silvia Del Din, Hugo Hiden, P8 – Fau - Felix Kluge
Reviewer	P18 - IXS

Due date (project	M12
Actual delivery date	M13
Document version	V1.5
Deliverable type	R
Dissemination level	CO

Document History

Version	Date	Description	Contributors
V1.1	15.04.2020	First Draft	Heiko Gaßner, Jochen Klucken
V1.2	22.04.2020	Second Draft	Brian Caulfield
V1.3	27.04.2020	Third Draft	Heiko Gaßner
V1.4	30.04.2020	Fourth Draft	Brian Caulfield
V1.5	30.04.2020	Submitted version	

Table of Contents

1 Publishable Summary

The main objectives of Mobilise-D are threefold: to deliver a valid solution (consisting of sensor, algorithms, data analytics, outcomes) for real-world digital mobility assessment; to validate digital outcomes in predicting clinical outcome in chronic obstructive pulmonary disease, Parkinson's disease, multiple sclerosis, and proximal femoral fracture recovery; and, to obtain key regulatory and health stakeholder approval for digital mobility assessment.

Work package 3 is responsible for development and implementation of the data management platform for Mobilise-D. The data management platform for Mobilise-D must cater for robust and reliable processes for end-to-end capture, ingestion, integration, storage, analysis, access, and sharing of all data that is generated during the activities of Mobilise-D. This platform must also adhere to all relevant legal and ethical frameworks and facilitate Mobilise-D's alignment to the FAIR and ALCOA+ principles regarding data access and integrity.

In the first 12 months of Mobilise-D, the WP3 team has worked closely with all relevant stakeholders to detail the specific requirements for the different elements of the Mobilise-D data management platform. This deliverable describes these requirements, in as much as their detailed specification has been agreed at the time of writing. The deliverable focuses on the requirements for data capture, ingestion, and integration in the Technical Validation Study (TVS) and Clinical Validation Study (CVS) in Mobilise-D. The requirements for data storage, access, analysis and sharing/presentation are still under discussion, and details of these will be presented at a later date.

This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No. 820820. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme, and the European Federation of Pharmaceutical Industries and Associations (EFPIA).

2 Introduction

2.1 Context

Mobilise-D is an ambitious undertaking that will involve multimodal data capture over a number of years from 16 different sites in different countries. The data that will be captured as part of the Mobilise-D programme must meet the highest standards of data integrity, facilitate multi-party sharing and access, and adhere to appropriate EU data protection and security standards. Therefore, the data management platform that sits at the heart of Mobilise-D, and the processes that underpin it, must be robust, reliable, secure and fit for purpose. In the first 12 months of Mobilise-D, the WP3 team has worked closely with relevant stakeholders to flesh out the requirements for this data management platform. The current deliverable sets out our understanding of these requirements, while describing progress towards addressing them. It focuses on the requirements related to data capture, ingestion and integration as this has been the main focus of the first 12 months. Requirements related to data storage, analysis, governance and access are addressed. However, as discussions regarding these elements of the platform are ongoing, detailed and agreed requirements will not be presented in this document.

2.2 Process

Throughout the first 12 months of the Mobilise-D programme, the WP3 partners have adopted a flexible and collaborative approach to working with other WPs to specify the requirements for the Mobilise-D data management platform. Such an approach has been necessary in the context of the specifics of the TVS and CVS protocols not being locked down at the outset of the programme. Indeed, at the time of completing this deliverable, the protocols for the CVS have not been fully agreed and signed off. The WP3 approach has been to work with partners to identify high level requirements and work towards a platform specification that is flexible enough to meet high level needs and adapt to specific needs as they are identified and agreed.

The work of WP3 began in May 2019 with a one-day requirements gathering workshop that was hosted by Universitätsklinikum Erlangen. This was followed by a period of weekly teleconferences in which representatives of WP2, WP4 and WP6 worked with WP3 partners to identify high level requirements related to the TVS, CVS and data analysis activities in Mobilise-D. Having held these meetings through June and July 2019 there was a change in strategy whereby relevant WP3 partners implemented a programme of participation in activities of WP2 and WP4 to work with relevant partners to advance our requirement gathering work. A further 1-day workshop was held in Newcastle in October 2019, in which WP3 presented its view of platform requirements as well as the planned approach to meeting these requirements to relevant partners from WP2 and WP4. As the work progressed, WP3 researchers participated in regular meetings with WP2 and WP4, and in WP4 (Stuttgart, Dec 2019) and WP2 (Sheffield, Jan 2020) workshops, to present their understanding of requirements and planned approach. A 2nd workshop was planned for April 2020 to present WP3 understanding of platform requirements and secure agreement/sign off on implementation plans. However, due to covid-19 control measures implemented across Europe, this workshop was cancelled.

3 Overview

3.1 Brief description of data requirements for Mobilise-D Research Programme

In order to achieve the D3.1, the team generated a data requirement concept tailored to the distinct characteristics of the Technical Validation Study (TVS) of WP2 and the Clinical Validation Study (CVS) of WP4, as well as acknowledging the requirements for data analysis (WP6). Therefore, the requirement specifications were separated into two different main aspects derived from A) the clinical + sensor-derived data requirements, and B) the data handling requirements. Within these two main aspects, several subcategories were specified:

A) Clinical + sensor-derived data requirements for TVS and CVS

3.1.1 Data Requirements derived from TVS and CVS protocols

3.1.2 Data Pipeline

B) Data handling requirements (Acquisition, Transmission, Integration, Processing, Analytics, Governance requirements) and presentation for analysis

3.1.3 Legal Frameworks

3.1.4 Data Quality

3.1.5 Data Accessibility

3.1.6 Data Security

As a final step of D3.1 the data requirement dictionary will be handed over to Task 3.2, 3.3 in order to implement the data assessment, transfer, and evaluation requirements (part of D3.2-3.4)

The following sections will describe the individual aspects and subcategories of the data requirement deliverable:

3.1.1 Data Requirements derived from TVS (WP2) and CVS (WP4) protocols

The research programme of Mobilise-D incorporates technical and clinical validation phases:

Technical Validation Phase:

For the purposes of this deliverable, we can further divide the technical validation phase into a retrospective analysis of legacy digital mobility datasets into existing algorithms, and into a prospective TVS.

Analysis of Legacy Datasets:

In this task, existing digital mobility datasets that have been sourced by members of the Mobilise-D consortium have to be integrated and analysed to facilitate a first level of validation and performance comparison of existing algorithms. The aim of the analysis on the legacy datasets is to perform a comparative evaluation for identifying the best candidates and combination of algorithm(s) for evaluation of selected digital mobility outcomes (DMOs) to be then optimised and tested in the TVS.

From 52 potential datasets made available by the consortium partners, six datasets were identified to be included in the comparative analysis. The choice of datasets was based on a minimum set of requirements: permission (fully anonymised and sharable dataset), protocol (i.e. presence of lab-based and/or free-living/ structured activities gold standards), sensor position (availability of sensor on lower back), algorithms (availability of at least accelerometry and gyroscope data) and population (e.g. people with Parkinson's disease, people with Multiple Sclerosis).
The IMUs and gold standard data of the six selected datasets have been fully standardised, they are available and readily accessible by all Mobilise-D partners through the Mobilise-D data management platform (Table 1). The analytical workflow that has been put in place to process legacy datasets (Figure 1) is outlined in Appendix 1.

Partner	Dataset name	Type of test	Gold standard	PD	MS	CTRL	Other
UNISS	UNISS_UNIGE	Laboratory: Tests	Instrumented Walkway	10		10	10 Stroke, 10 Chorea
CAU	KC validation-walking tests	Laboratory: Tests	Stereo-photogrammetric			9	5 stroke, 1 Dementia, 1 lymphoma
CAU	KC validation-semistructured activities	Laboratory: Semi-structured activities	Stereo-photogrammetric	5	1	5	2 stroke
UNEW	ICICLE	Laboratory: Tests	Instrumented Walkway	119		186	
USFD	MS Project	Laboratory: Tests	Multiple IMUs		59	24	
USFD	Gait in Lab and real-life settings	Free-living and Laboratory: Tests	Multiple IMUs			20	

Table 1: List of selected datasets. CTRL: Controls, MS: Multiple Sclerosis, PD: People with Parkinson's disease.

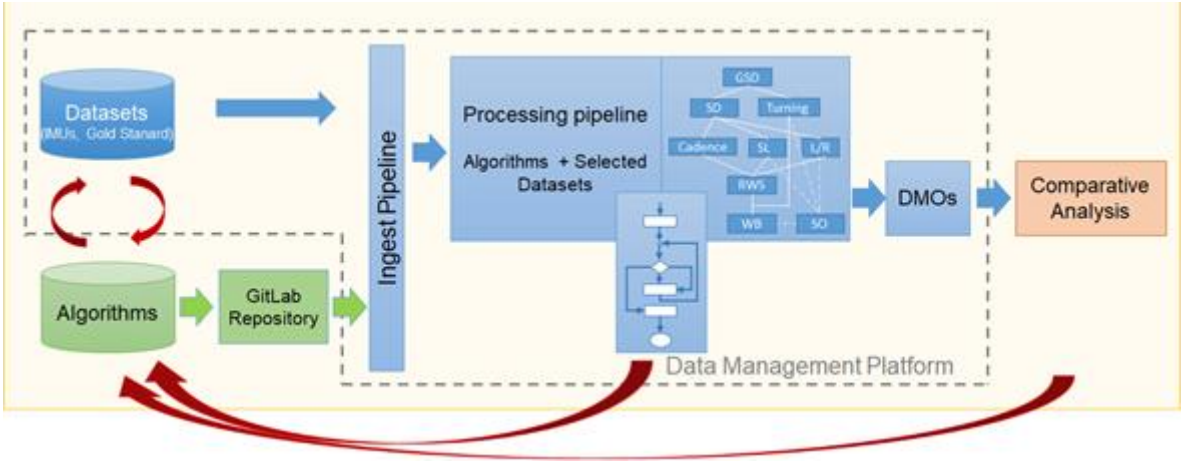

Figure 1: Framework for analysis of legacy datasets: dataset selection, algorithms repository (GitLab) and packaging, analytical/ processing pipeline for evaluation of DMOs and comparative analysis. Red arrows depict iterative/ cyclic steps. DMOs: Digital Mobility Outcomes, GSD: Gait Sequence Detection, L/R: Left/Right Step detection, RWS: Real Walking Speed, SD: Stride detection, SL: Stride Length, SO: Secondary Outcomes, WB: Walking Bout Assembly.

Prospective Technical Validation Study:

In the case of the prospective technical validation study, the task is to collect a range of data, as detailed in Figure 2, from 125 study participants in 5 cohort groups across 5 clinical sites.

Clinical Validation Phase:

In the clinical validation phase, the task is to collect a large range of data from 2400 participants in 4 cohort groups over 5 time points in 15 clinical sites. The study flowchart, detailing the data that will be generated at each time point, is outlined in Figure 3 below.

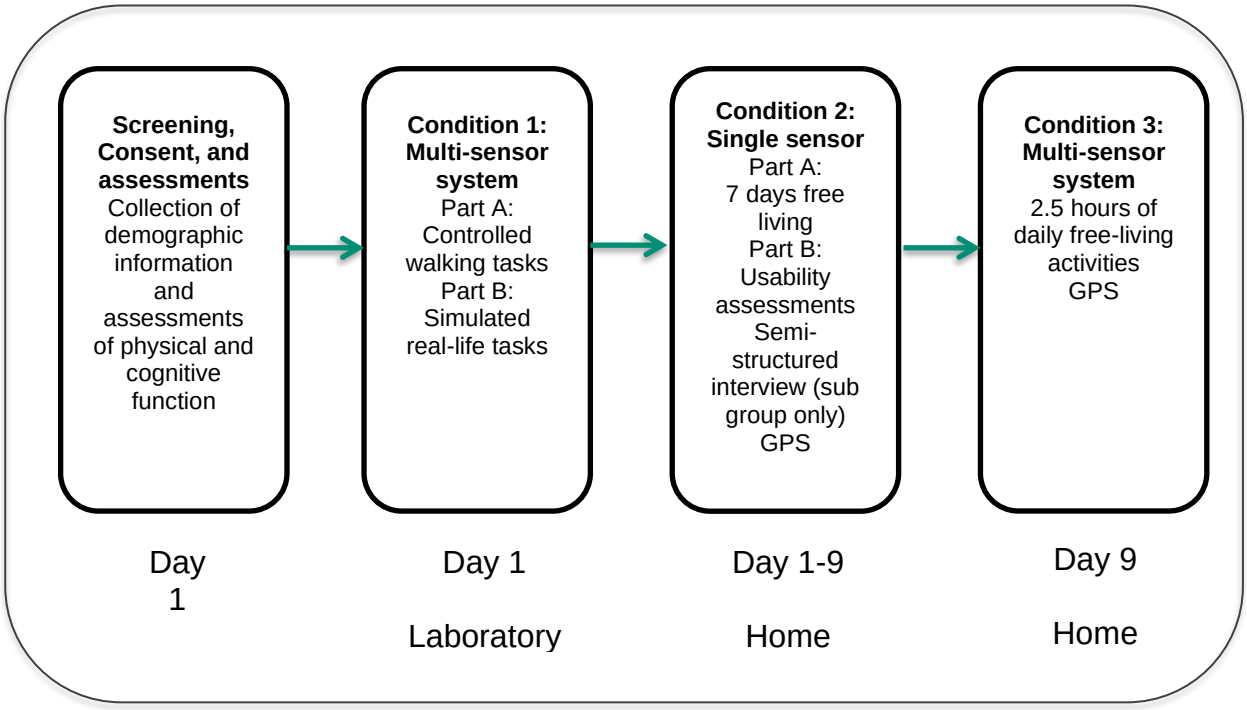

Figure 2. Overview of Activities and Datasets for Prospective Technical Validation Study.

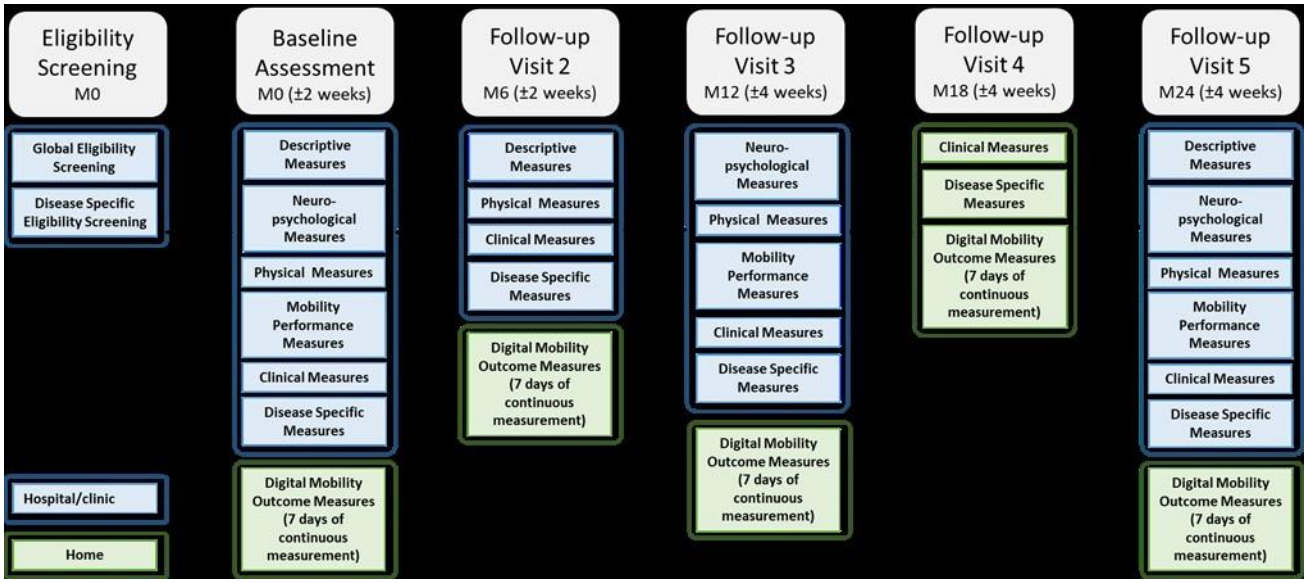

Figure 3. Overview of Activities and Datasets of the Clinical Validation Study.

3.1.2 Data Pipeline

Data flow is defined for the technical validation study (TVS) and clinical validation study (CVS) between the sources (patient- and investigator generated data as well as gold standard data) and the e-Science Central platform. For each type of data in this study a standardised and secure data transfer pipeline will be required, to ensure that all data can be integrated on the Mobilise-D data management platform. The data transfer requirements for the TVS and CVS are outlined in table 2 and 3 below. Data flow summaries for both studies are illustrated in figures 4 and 5 below.

Table 2: TVS Data Transfer Summary

Data Type	Pipeline
Clinical & Demographic Data	Data capture at 'patient-side' via e-form directly on Mobilise-D platform (eSC - HTTPS site will be accessible in 5 sites). In the absence of internet access we can capture data offline using template (.csv) files and upload as source when internet access available. As a last resort data can be captured on paper then copied / signed x2 (4 eyes principle), scanned and uploaded. Original paper forms kept at site as per local guidelines.
Standardised Clinical Outcome Assessments	*2 different strategies in this case:* **For COAs that are available on ERT platform:** All disease-specific scales (apart from LLFDI) are captured on ERT platform which has a dual-login for patient and clinician; ePRO - CAT, KCCQ, Pain VAS eClinRO - MS-UPDRS Part III, EDSS ERT-supplied tablet will be used to capture and transfer data to ERT platform. Data automatically transferred to e-SC via API. Once transferred to e-SC data will be removed from ERT server. **For COAs that are not available on ERT platform:** Data capture at 'patient-side' using electronic means this can be either e-form (completed on e-SC portal) or template (.csv uploaded to e-SC via portal). Where neither of these options are suitable (e.g. MoCA) paper will be used and the paper-based data source will be copied and signed x2 then scanned and uploaded to eSC. Original paper forms maintained at local site for duration as per local guidelines.
Motion data from Optoelectronic Motion Capture system	Files directly uploaded from SP PC to e-SC via web portal.
Motion data from McRoberts device	There are 2 types of data collected from the Dynaport MM+. 1. Data from short discrete RW walking tests which are transferred from sensor to PC via McRoberts application are manually uploaded to e-SC. 2. The second type of data is raw data for 7-day continuous assessment and is transferred from sensor to PC via McRoberts application and uploaded to McRoberts servers then on toe-SCvia API. Once transferred toe-SCall data will be removed from McRoberts server (before the end of the study).
Motion data from INDIP system	Data from sensors copied to INDIP application on study provisioned laptop. Transferred directly to e-SC via web portal.
Images from digital video capture device	Files reviewed at local site on an 'as needed' basis and relevant annotations captured in custom designed web-form for upload to e-SC. In the absence of a web form we can use a template file which can be uploaded to e-SC via the portal. Original video files maintained at local site for duration as per local guidelines.
Annotation data from USFD mobile app	Data transferred from phone to USFD servers. Processed data (annotations that describe walking aid use, indoor/outdoor location, walking up/down hill, and frequently visited location types) transferred from USFD servers to e-SC as required (to contextualise walking bouts). Once transferred to e-SC, raw data will be maintained in USFD for duration as per local guidelines.
Patient perceptions and feedback data from semi-structured interviews	Usability Scales: Capture data in template file on laptop (offline) and upload to e-SC via portal when back at site. Interviews: Audio files transcribed, cleaned (to remove any potential identifier remarks), and translated at local site. Transcriptions uploaded to e-SC via web portal. Once transferred to e-SC, audio files will be maintained at local site for duration as per local guidelines. '4 ears' principle used to check accuracy of audio transcription at local site by means of interviewer listening back to recording and checking against transcription. *Note: The preference would be to complete an e-form at the home/workplace but internet access may be a problem so the best alternative is completing a form offline and uploading when internet access available.*

Table 3. CVS Data Transfer Summary

Data Type	Pipeline
Clinical & Demographic Data	Data capture at 'patient-side' via e-form directly on Mobilise-D platform (e-SC - HTTPS site will be accessible in 5 sites). In the absence of internet access we can capture data offline using template (.csv) files and upload as source when internet access available. As a last resort data can be captured on paper then copied / signed x2 (4 eyes principle), scanned and uploaded. Original paper forms kept at site as per local guidelines.
Standardised Clinical Outcome Assessments	*2 different strategies in this case:* **For COAs that are available on ERT platform:** All disease-specific scales are captured on ERT platform which has a dual-login for patients and clinicians as appropriate. Full list of ePROs and eClinROs to be determined. ERT-supplied tablet will be used to capture and transfer data to ERT platform. Data automatically transferred to e-SC via API. Once successfully transferred to e-SC data will be removed from ERT server (before the end of the study). **For COAs that are not deployed on ERT platform:** Where practical we can use electronic forms directly on the e-SC platform. In cases where electronic data capture is not appropriate due to practical considerations paper will be used and the paper-based data source will be copied and signed x2 then scanned and uploaded to e-SC. Original paper forms maintained at local site for duration as per local guidelines.
Motion data from McRoberts device	Raw data for 7-day continuous assessment is transferred from sensor to PC via McRoberts application and uploaded to McRoberts servers then on to e-SC via API. Once successfully transferred to e-SC all data will be removed from McRoberts server (before the end of the study).
Patient perceptions and feedback data from semi-structured interviews	Usability Scales: Capture data in template file on laptop (offline) and upload toe-SC via portal when back at site. Interviews: Audio files transcribed, cleaned (to remove any potential identifier remarks), and translated at local site. Transcriptions uploaded toe-SC via web portal. Once transferred to e-SC, audio files will be maintained at local site for duration as per local guidelines. '4 ears' principle used to check accuracy of audio transcription at local site by means of interviewer listening back to recording and checking against transcription. *Note: The preference would be to complete an e-form at the home/workplace but internet access may be a problem so the best alternative is completing a form offline and uploading when internet access available.*

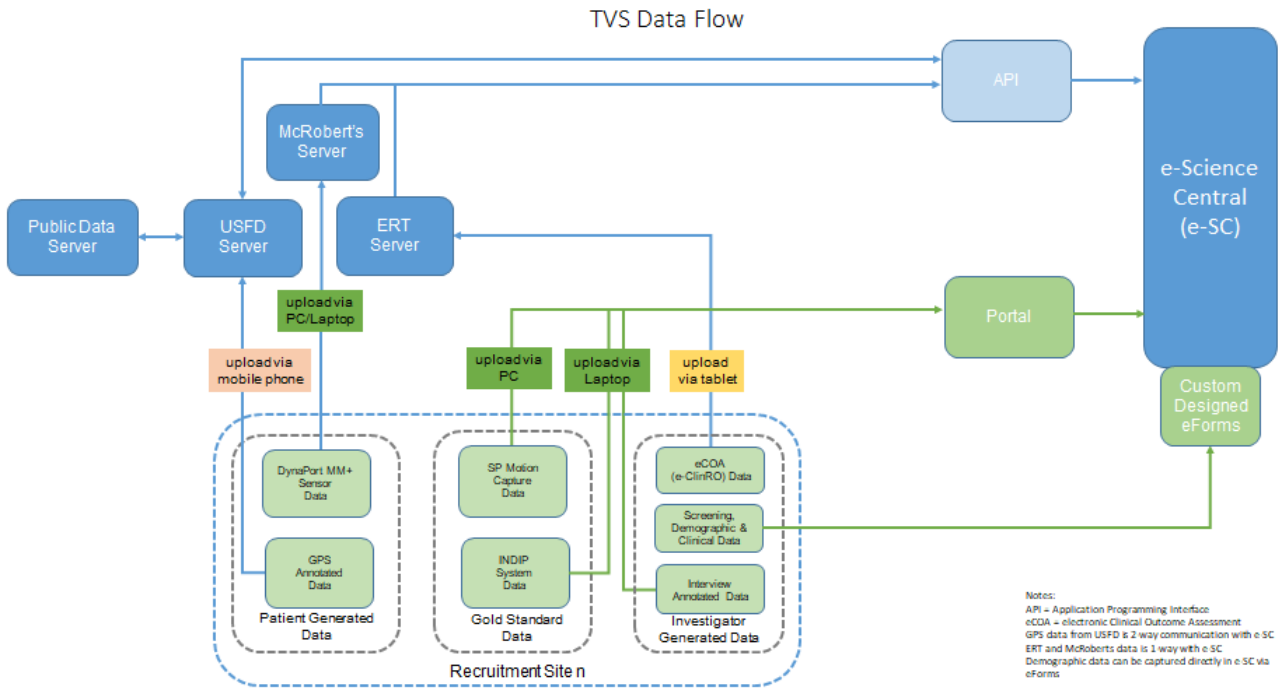

Figure 4: Data Flow for TVS

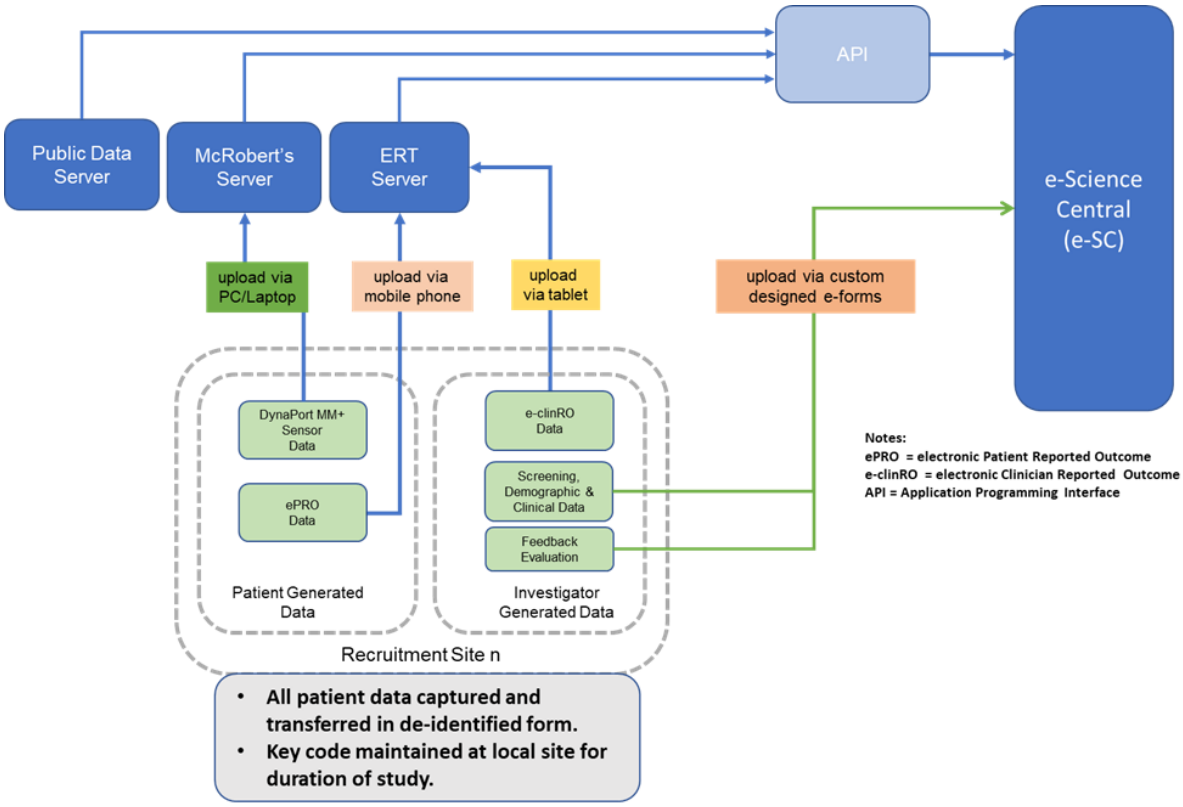

Figure 5: Data Flow for the CVS

Transferring Data to the e-Science Central Platform

e-Science Central (e-SC) is a generic cloud hosted data storage and processing platform, which has been adopted as the central component of the Mobilise-D Data Management Platform (DMP).

As part of its study management capabilities, e-SC can be used to manage patient data files and observations. This information can either be uploaded directly via the e-SC website or via one of the supported Application Programming Interfaces (APIs).

The platform has been used in a variety of research and commercial applications ranging from drug design modelling to the analysis of data from tens of thousands of connected consumer devices across the EU.

In order to support data ingestion from McRoberts and other potential partners, additional APIs have been provided which allow data and observations (forms etc.) to be uploaded and associated with patients which can be assigned to different studies (which represent the various sites and cohorts).

This API is restricted in that it can only be used to upload information and verify that this data has been successfully processed. Partners have no facility to download patient data, or browse through the patient cohorts etc.

All of the APIs to e-SC are secured using standard encryption methods and access is granted via the use of JSON Web Tokens which can be created and revoked centrally to control access to the platform.

The service itself is hosted in the Amazon Cloud and physically located in Frankfurt. There are two types of data storage associated with the platform – a SQL database contains the algorithm data, cohort information, file structures etc. and the raw data files themselves are stored in Amazon S3 storage. The database itself is an Amazon hosted PostgreSQL database which is encrypted and regularly backed up, whilst the S3 data is also encrypted at rest within the platform.

McRoberts have been provided with details of the upload API that they will use to send data to the platform and have managed to send data and validate its upload successfully. Raw data from McRoberts sensors (and, potentially, other sources) can be processed using the e-SC workflow engine. This allows code to be deployed to the platform and used to manipulate and analyze data files. The platform already provides basic file management / data pre-processing tools and specific algorithms are being provided by WP-2 in the form of Docker images that contain compiled Matlab code that has been benchmarked against various sets of Gold Standard movement data.

Data from patients (forms, metadata etc.) is currently stored as semi-structured data in the e-Science Central database and will be extracted, transformed and loaded into a data warehouse database when the structure for this has been finalized. The process can be repeated if needed if the structure of the data warehouse needs to change as the design progresses.

Data Integration

Data will be integrated on the platform by means of implementation of a standardised file nomenclature system. At point of capture, each file will be labelled in standardised format, including information on: Centre, Participant unique ID, Data source/ task, and Time point i.e. *Center#-Patient#-Source-Taskxx*-ddmmyyyy. For example, Centre 01, participant 1000, StereoPhotogrammetric Data, Standing Posture on April 1st 2020 will be listed as 01-1000-SP-

Standing01-01042020. In addition, the e-SC platform manages the assignment of data and files to individual patients assigned to studies automatically.

3.1.3 Legal Frameworks

All management of data in the Mobilise-D research programme has to adhere to the relevant EU legal frameworks. In particular, Mobilise-D has to be fully compliant with the following:

- General Data Protection Regulation (GDPR)
- Directive 2006/24/EC of 15 March 2006 on the retention of data generated or processed in connection with the provision of publicly available electronic communication services or of public communications networks

3.1.4 Data Quality.

Data monitoring will be implemented to ensure data quality of each instance of data created, used, and stored. A study monitoring plan has been created which outlines the monitoring responsibilities. Central monitoring of recruitment and data quality will be undertaken by individuals at Newcastle University (UNEW) and Norwegian University of Science and Technology (NTNU). The monitoring will involve accessing pre-configured reports from the Mobilise-D platform (e-Science Central). A named individual at both sites will have access to this facility. On-site monitoring training will be replaced by automated data validation and assessor training. The hands-on training sessions will be led by researchers at University of Sheffield (USFD) and University of Sassari (UNISS). Attendance will be documented via training certificates and in a training log filed in each site's Investigator Site File (ISF). On-site monitoring visits may be triggered if serious issues are identified at a site level. With regard to quality of data, this project adheres to ALOCA+ principles as research programme as this is ultimately serving a submission to regulatory authorities. ALCOA is the abbreviation for **A**ttributable, **L**egible, **C**ontemporaneous, **O**riginal, and **A**ccurate. The + describes additional quality criteria: **C**omplete, **C**onsistent, **E**nduring, and **A**vailable. In the following, the ALCOA+ and 4-eyes/ears principles are described in detail.

Attributable

When creating a record, you must record the identity of the person or computer system that collected or generated the data. It's also important to record the date of the collection or generation.

Legible

Ensuring data is legible is about more than being able to clearly read the data, although that is important in situations where manual data record-keeping takes place. Being able to make out the words is much less of a problem with electronic data, though.

That said, ensuring data is legible still has relevance. This is because it must be possible for data to be read and understood years and even decades after it's recorded. This can have as much relevance to digitally recorded data as it does to data recorded in notebooks.

In other words, it's important to avoid using clichés and unusual phraseology as this may be difficult to decipher in the future without getting clarification from the originator of the data, a person who may no longer be available.

Using consistent, straightforward language throughout an entire organisation, regardless of locality, is the best approach.

Contemporaneous

It's essential that individuals or systems make a record of an activity at the time it takes place. With electronic data, this is normal practice, so this is another point that has more relevance to manual record-keeping.

Original

Records should be original rather than copies or transcriptions. Again, this applies mostly to manual record-keeping. For example, you should not write a record on a scrap of paper with the intention of completing the main record later, as this can result in errors.

Instead, the original recording of the data should be on the main record, whether that record is on paper or on a digital system.

Accurate

All records should reflect the reality of what happened and should be error free. Also, there should be no editing of the original information that results in that information being lost.

If changes are necessary, those changes must be documented in a way that makes it possible to refer back to the original information. Nothing should be removed, blocked out, or deleted.

When recording data electronically, the system must have built-in accuracy checks and verification controls.

Complete

All recorded data requires an audit trail to show nothing has been deleted or lost.

Consistent

This primarily means ensuring data is chronological, i.e. has a date and time stamp that is in the expected sequence.

Enduring

While this is touched on in a previous principle, this principle of ALCOA+ places specific emphasis on ensuring data is available long after it is recorded – decades in some situations.

Available

This point follows on from the last point, i.e. data must not only exist, it must be accessible. The most efficient way of achieving this is normally by recording data electronically.

Four eyes/four ears principle

The four eyes/four ears principle is a requirement that two individuals approve some action before it can be taken and should guarantee that the process is done in a correct way.

3.1.5 Data Open Access.

The Mobilise-D consortium is committed to making outputs open and available for use by the wider research and innovation community, in so much as doing so is within the control of the consortium and is in compliance with appropriate legal and ethical frameworks. We have set out a series of

guiding principles to inform our approach to deliver on this commitment. In this Deliverable, we describe these guiding principles in general terms referring to D7.3 (Guidelines for Open Access and Data Sharing).

Consortium Commitment and Agreement:

The Mobilise-D consortium is committed to making data assets, algorithms, publications, and standards developed and/or captured during the execution of the research programme available to the wider scientific community. This has been captured and agreed in the Consortium Agreement (CA), which has been signed by all partners in the consortium. The CA, in Section 8.3, states that 3rd parties will have the right to request access rights to the results of Mobilise-D for research use for a period of 15 years following completion of the programme. Furthermore, the CA states that such access rights will be granted subject to a bilateral agreement that is framed in response to a written request. It is important to note that the commitment to making Mobilise-D outputs open must achieved with due regard to protecting the anonymity of study participants, as well as reserving the IP rights of contributing partners.

Ethical Approval:

In adhering to the appropriate legal frameworks and establishing the legal basis for processing and sharing data, we need to establish a strong foundation for doing so at the point of obtaining ethical approval for our research involving prospective data collection. In practical terms, this means that we need to make provisions for open data in our ethics applications by means of signaling our intentions to seek consent for data sharing in our application material. In the case of Mobilise-D, this is a complex process addressed by multiple ethics committees in both phases (5 ethical committees for the technical validation study and 15 for the clinical validation study). At the application stage, we will have to set out our plans for explicit data sharing consents as well as subsequent archiving and sharing proposals.

Explicit Informed Consent:

As study participants are prospectively enrolled in Mobilise-D studies, we will provide them with a clear and explicit understanding (in writing) of our intention to make their research data available for sharing as open data. We will subsequently obtain written consent from them to enable us to do so. This will be done in full compliance with the terms of our ethical approvals and appropriate legislation. Where appropriate, we will implement tiered consenting models to allow participants to opt in/out for sharing specific data sources. All data that will be shared in an open platform will be irrevocably anonymized.

Single Point of Access:

Sharing of data will conform to the findable, accessible, interoperable and re-usable (FAIR) principles (https://www.go-fair.org/fair-principles/). As such, we will endeavour to host all Mobilise-D data on a common platform so that to there is a single point of entry to Mobilise-D data assets, though some data may also be hosted on national data repositories if required.

Access Protocols:

We will develop and disseminate detailed access protocols for 3rd parties who wish to access Mobilise-D outputs. This will incorporate the documentation and processes that should be used when requesting access, the terms of license agreements that govern access and use, and the level of acknowledgment that will be required in subsequent research outputs. These access protocols will be detailed in a future WP3 Deliverable (D3.6 Sustainability/Extensibility plan), due for publication in M54.

3.1.6 Data Security

Data Protection

All data will be de-identified at point of capture. No identifiers will ever be stored alongside any volunteer participants' data. Upon entering the study, a unique code will be created for each study volunteer. The study key code, which has details of volunteer participants' names and study codes, will be maintained in paper format at each recruitment site. This key code will be kept under secure conditions at each site, as per local guidelines. We expect that the minimum level of security will entail the key code being stored in a locked filing cabinet in a single occupant office.

The key-code will be destroyed once database is locked and local guidelines permit. It will be maintained for defined period (e.g. 5 years) in sites where this is required.

Secure protocols will be used to capture and transfer data both manually (HTTPS POST) and through 3rd party APIs (JWT) to e-SC which is hosted in the cloud on AWS (Frankfurt). Stored data will be encrypted using Amazon Server-Side Encryption (SSE-S3) which uses AES-256 block cipher.

Only de-identified data will be ingested to Mobilise-D platform. All data will be integrated on the platform using the unique code that is created for each study volunteer and the file nomenclature system outlined above. We will implement a 'Privacy by Design' protocol on the Mobilise-D platform. This will incorporate application of technical anonymization protocols to render the data to anonymous prior to it being stored in the data warehouse. There will be a very well defined access and governance model in place to ensure that access to the source (de-identified data) is limited to a small core group. Wider access is only available for the anonymised dataset.

The different roles, and their access rights are as follows:

Study volunteer – Provision of data under principles of informed consent. No access to the data management platform. Has the right to access a copy of their own data upon request, or to have their data removed from the database. These rights terminate once the dataset is anonymised and the study key code has been destroyed.

Clinical Investigator – Responsible for management of informed consenting process and acquisition of data from study volunteers. All data is captured in de-identified manner. Once data is collected, they are responsible for uploading it to the database (via the relevant pathway). Once the data is uploaded to the platform they can no longer access it, though they can receive summary reports regarding the level of completeness of the dataset. Cannot access the full dataset.

Data Controller – This is the lead clinical investigator at each site who has overall responsibility for the execution of the study at that site. They are responsible for the chain of custody of the dataset that originates from their site. They can access the full dataset.

Platform Manager – Responsible for building and maintaining the data management platform, incorporating implementation of appropriate industry standard security protocols, and overseeing access and governance for the platform. Responsible for implementation of quality control. Responsible for implementation of data integration and data anonymisation protocols once data is ingested to the platform. Can access the full dataset. Formally considered a data processor.

Data Analyst – Responsible for implementation of data analysis protocols, including processing and mining tools, once the anonymised dataset is available in the data warehouse. Can access the anonymised dataset in the data warehouse, can implement data analytics tools within the platform, and can create new derived datasets and results files. Cannot access the data as it is ingested to the platform. Can only access the data once it has been subjected to anonymisation processes. Formally considered a data processor.

Statistician – Responsible for implementation of pre-defined statistical analyses on the derived data and results files. Can only access these data and files. No capability for accessing the raw data. Formally considered a data processor.

Data storage

The full anonymised dataset will be maintained indefinitely as part of the Mobilise-D commitment to an open data policy. Once the study is completed, the fully anonymised dataset will be made available to the wider research community for secondary research purposes.

Source data will be maintained at local site of capture in de-identified manner for period of time stipulated by local ethics committee (normally 5 years). Once this period of time has elapsed the original de-identified dataset, and the study key code, will be destroyed.

3.2 Specification of Data Dictionary

A key enabler in Task 3.1 was the stakeholder requirements workshop, which was held on May 17th in Erlangen, Germany (first month of the project). This workshop included relevant stakeholders from WP2, WP4, and WP6. The output from this initial workshop was the first draft of an agreed project vocabulary and definitions, as well as a detailed outline of the functional requirements and system architecture for the data management platform in relation to the activities of WP2. Afterwards, we were continuously working on the data workflow and definitions of parameters that were specified with cohort leads or assigned representatives as well as contributors from WP6. In March 2020, we finalised a data dictionary for the WP2 trial defining the characteristics of about 350 single variables (see example in Figure 6). This data dictionary serves as basis for the WP4 trial. The compounds of this dictionary are described in the following:

Disease entity – includes parameters of the core data set (all entities) and disease specific parameters for PD, MS, CHF, PFF, COPD, HA (healthy adults)

Time of recording – describes different parts of recording e.g. general screening, characterization, descriptive measures, participant interviews

Parameter ID – a unique number for each parameter was defined

Parameter description – detailed explanation of the parameter

Data recorded by – identifies the data source, e.g. investigator, participant or participant asked by investigator

Data input – defines the structure of the data entered, e.g. Yes/No; concrete numbers or categories

Scale – describes the scale the parameters belongs to (nominal, ordinal, metric)

Categories – if applicable the categories are defined here. For gender: male, female, prefer not say

Additional information – if necessary, more information can be given here

Study phase – differentiates between technical validation study (WP2) or clinical validation study (WP4)

Recording device – describes the device where the data entered to

Patient- and investigator generated data were defined for general parameters (all entities) and for disease specific parameters. These data input requirements are finalised for WP2 and in preparation for WP4 (will be part of D3.4). At the current time point, the data structure of the core data set for the clinical validation study (WP4) is completed. The definition of disease specific parts of the WP4 protocol is in progress. Monthly stakeholder review meetings will be held in order to finalise the data dictionary for the WP4 study.

Disease entity	Time of recording	Parameter ID	Parameter Description	Data recorded by (person, measure)	Data input	Scale
ALL	General screening	G1	Participant has read PIS and had opportunity to ask questions	Investigator	Y/N	nominal
ALL	General screening	G2	Informed consent form completed / signed	Investigator -signed by participant	Y/N	nominal
ALL	General screening	G3	Year of Birth	Participant - asked by investigator	yyyy	metric
ALL	General screening	G4	Age	Participant - asked by investigator	xx years	metric
ALL	General screening	G5	Gender	Participant - asked by investigator	Please select	nominal
ALL	General screening	G6	Place of residence	Participant - asked by investigator	Please select	nominal
ALL	General screening	G7	Education	Participant - asked by investigator	Y/N	ordinal
ALL	General screening	GC1	Comorbidity check: primary diagnosis	Assessor (review medical notes)	diagnosis	nominal
ALL	General screening	GC1.1	If PFF: first hip fracture?	Assessor (review medical notes)	Y/N	nominal
ALL	General screening	GC2	Confirmed via	Assessor (review medical notes)	Please select	nominal
ALL	General screening	GC3	Secondary diagnosis	Assessor (review medical notes)	Y/N	nominal
ALL	General screening	GC4	Sondary diagnosis impacts gait	Assessor (review medical notes)	Y/N	nominal
ALL	General screening	G8	MoCA sum score (calculated by system)	Investigator	J/N	nominal

Figure 6: Example of data dictionary for the technical validation study

3.3 Data Analysis

At the time of completing this Deliverable document the detailed platform requirements for data processing, analysis, visualization and presentation for the full Mobilise-D work programme have not been specified as the CVS protocols and Data Analysis Plan have not been completed. However, in the case of phase 1 (analysis of legacy datasets and the TVS analysis plan) significant progress has been made.

Comparative analysis plan:

An objective methodology for comparing algorithms performances has been developed by USFD in collaboration with UNEW, ISGlobal and UCD. This approach will be used for analysing data from existing datasets and will be extended to the TVS.

For this aim, an *ad-hoc* decision matrix has been established to objectively rank the different algorithm performances. Such matrix is made of three components (i.e., the elements to evaluate, their relevance and the relevant scoring system), as proposed in established design processes.

Four domains (concurrent validity, human factors, data capture process and wearability and usability) and relevant criteria were identified through literature research (Figure 7). The criteria were shared

with the whole consortium in a questionnaire for final agreement relative to the ad-hoc matrix, with concurrent validity deemed the criterion with highest relevance (importance).

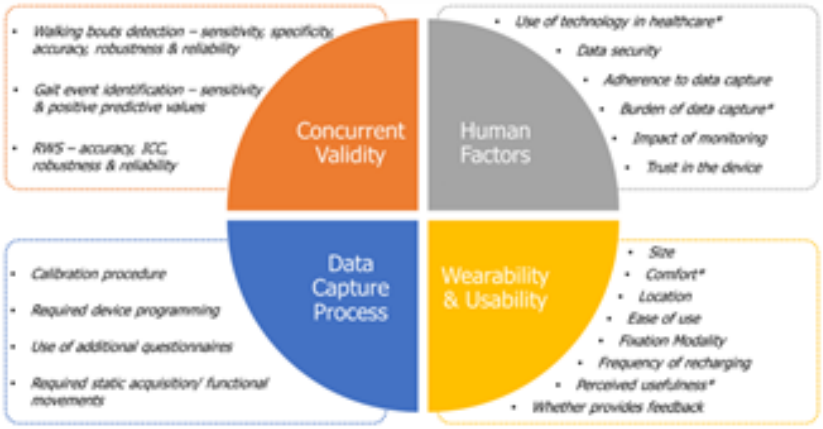

Figure 7: Domains and relevant criteria identified for comparing algorithms performances.

To quantify the concurrent validity domain, a comprehensive series of validity criteria for evaluating the performances of each of the algorithms has been developed by UNEW as follows.

- The algorithm performance evaluation for each analytical block will be based on results from:
- Stride Detection, Stride Length, Cadence: time events (e.g. initial contact) and DMOs (e.g. cadence, stride time, stride length).
- Gait Sequence Detection, Turning, Walking Bout assembly: start and end events
- Real Walking Speed: values averaged across each identified walking bout
- For each analytical block output, the following metrics will be quantified:
  -Error
- -Absolute and relative error (%) (for identified events and DMOs)
- -Specificity, sensitivity, F1 score, accuracy and positive predictive value
- Criterion validity:
- Concurrent validity: Interclass correlation coefficient (ICC) between DMOs and gold standard values.
- Reliability
- Repeatability: Test-retest of absolute and relative errors will be assessed within and between subjects for several gait trials.

Moreover, for a better understanding of RWS algorithms performances, an assessment of RWS outcomes based on walking bout length will be performed.
All outcomes will be analysed based on the properties of the walking bout of origin.
Thus, several groups of walking bouts will be structured for further analysis.

- -Group 1: ∀WB > WBmin (e.g. 4 strides as per agreed operational definition)
- -Group 2: WBmin < WB < WB (10s)
- -Group 3: WB (10s) < WB < WB (30s)
- -Group 4: WB > WB (30s)

This analysis plan adopted for the analysis of the legacy datasets will then be extended to the TVS and will include analysis of laboratory-based data, 2.5 hours of free-living simulation data and 7 days free-living data. Within WP2, a Statistical Analysis Plan (SAP) will be generated with the support of WP6 for the TVS (see below).
In order to prepare the data for statistical analysis, WP6 has supported the process of defining the data requirements and reviewed the data structure for the TVS thoroughly. The data flow from the e-Science platform to WP6 for statistical analysis is in progress. With regard to deliverable 3.4

(Integrated end to end data collection, management and analytics platform), the process of generating a database providing requirements for “cleaned data” in relation to e.g. impossible values, missing values has begun.

3.4 Conclusion

The Mobilise-D data management platform has to cater for both the technical validation (legacy and prospective datasets) and clinical validation (prospective datasets) phases of the research programme, while adhering to relevant legal and data security norms and facilitating open access for the wider scientific community. In this deliverable we have outlined the requirements of the data management platform in as much as they have been agreed during the first 12 months of the Mobilise-D programme. The work of the first year of Mobilise-D in this regard has focused on the data capture, ingestion and integration elements of the pipeline, while setting out the data security, privacy, access and governance principles/constraints that will guide platform development. The next phase of development will focus on specification of the requirements and design of the data warehouse application of processing, analysis and visualization tools on the platform.

APPENDIX 1

Analytical workflow/ pipeline for evaluation of DMOs

An extensive, iterative process and constant interactions between WP2 and WP3 was carried out to support identification of analytical workflow/ pipeline for comparison of algorithms and quantification of DMOs (in particular, real walking speed (RWS) and walking bouts (WB)).
Key analytical blocks and relative algorithms have been identified and implemented including: Gait sequence detection (GSD), Stride Detection (SD), Cadence estimation, Stride Length (SL) estimation, Left/Right stride detection, Turning detection, RWS estimation, Secondary Outcomes (SO), WB assembly (Figure 8).

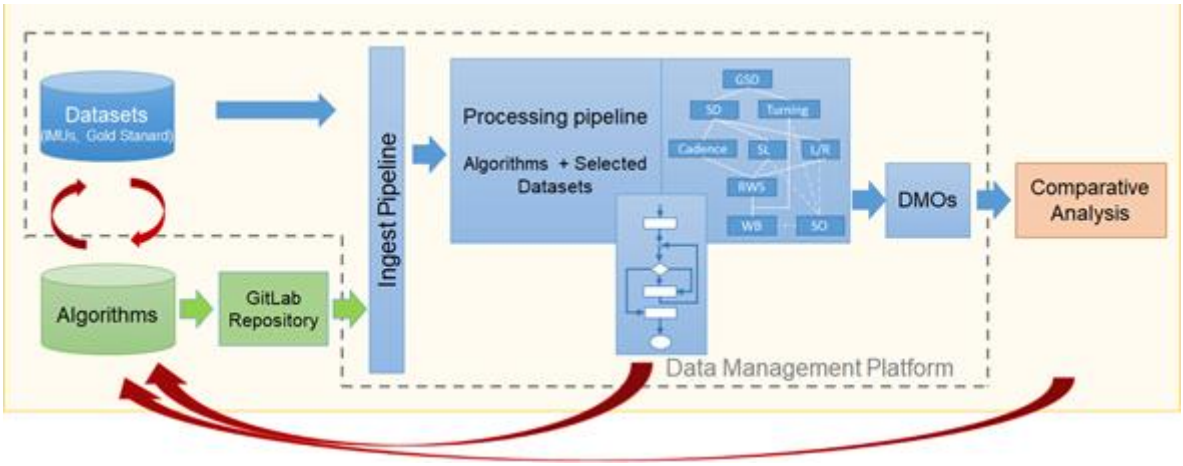

Figure 8: Framework for analysis of legacy datasets: dataset selection, algorithms repository (GitLab) and packaging, analytical/ processing pipeline for evaluation of DMOs and comparative analysis. Red arrows depict iterative/ cyclic steps. DMOs: Digital Mobility Outcomes, GSD: Gait Sequence Detection, L/R: Left/Right Step detection, RWS: Real Walking Speed, SD: Stride detection, SL: Stride Length, SO: Secondary Outcomes, WB: Walking Bout Assembly.

For each analytical block, a set of so-called Ground Truth algorithms have been developed, in order to allow for independent assessment of each of the blocks, independently on the added noise and errors from previous blocks. The Ground Truth algorithms replicate as an output the values of the gold standard so that the subsequent block will use gold standard data as an input to run the algorithms. For example, Ground Truth GSD will output the gait sequence detection obtained from the gold standard and will input it in the SD block, in this way it is possible to independently compare all the SD algorithms and find the one performing in the best way.
A GitLab repository was set-up by UCD, this allows secure storage and version control of all algorithms. All algorithms developers were given access to GitLab in order to upload available and ready algorithms to be packaged and run on the data management platform.
The numbers of algorithms currently uploaded for each of the blocks are the following: GSD: 8, SD: 18, Cadence: 17, SL: 19, Left/Right: 2, Turning: 2, RWS: 1, WB Assembly: 1, SO algorithms: 13.

Comparative analysis of algorithms on selected datasets
A multiple stage-approach was developed for the comparison of the algorithms based on a comprehensive set of validity criteria (see section 3.3 Data Analysis).

In stage 1 we set out to test the SD block, then GSD, Cadence, SL, and RWS. The comparative analysis started with lab-based datasets for testing of SD, Cadence, SL and RWS blocks. We then continued to extend the comparison to free-living like/ structured activities datasets for testing the entire analytical pipeline and the various combinations of algorithms.

A strategy has also been designed for machine learning-based Stride Length algorithms consisting in first using already trained models and then use an 80% training test and 20% testing set strategy for training and testing of models in each available dataset.

All tools for automatically testing the performance of the algorithms in each analytical block have now been implemented and used to quantify performances of the available SD algorithms on two available standardised datasets: ICICLE and UNISS-UNIGE.