D3.5 Data Management Platform Performance Evaluation

Mobilise-D

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

Grant Agreement No. 820820

[WP3 – Data Management Platform]

Lead contributor	Brian Caulfield
Other contributors	Hugo Hiden, Dave Singleton, Paul Watson
Quality reviewer	Mike Jackson

Due date (project month)	M42
Actual delivery date	27 March 2023
Document version	V1.0
Deliverable type	R[1]
Dissemination level	CO

[1] **Use one of the following codes:**
R: Document, report (excluding the periodic and final reports)
DEM: Demonstrator, pilot, prototype, plan designs
DEC: Websites, patents filing, press & media actions, videos, etc.
OTHER: Software, technical diagram, etc.

Document History

Version	Date	Description	Contributors
V0.1	22 March 2023	First Draft for review	Brian Caulfield, Dave Singleton, Hugo Hiden, Paul Watson
V1.0	24 March 2023	Final version for submission	Brian Caulfield, Dave Singleton, Hugo Hiden, Paul Watson

Table of Contents

1 Abstract

The Mobilise-D Data Management Platform (DMP) has been designed to facilitate the multimodal data capture, ingestion, integration and standardisation, processing, and analysis needs of the Mobilise-D Technical Validation Study (TVS) and Clinical Validation Study (CVS). This document describes the performance of the DMP to date with due regard to the system requirements that were specified in Deliverable D3.1. Specifically, it describes the range of challenges and associated performance thresholds that were necessary to meet the requirement of the TVS and CVS, the solutions that were implemented to meet these challenges, and the performance to date.

2 Context & Overview

The DMP for Mobilise-D was designed to meet the needs of the technical and clinical validation phases of the study. In meeting the needs of these studies the platform needed to operate within a constrained context and consider the following requirements:

1. *Adhere to ALCOA+ data integrity standards*
2. *Maintain compliance with GDPR and Data Privacy standards*
3. *Implement an appropriate governance and access control model within the consortium*
4. *Make provisions for safe and secure release of data and code to the wider scientific community*
5. *Meet the requirements of the TVS and CVS within the context of dynamic and evolving study protocols and data management requirements*

The DMP has been implemented in 2 instances to accommodate the needs of the TVS and CVS:

I - Technical Validation Study:

Study Cohorts:
PD (n=20), MS (n-19), COPD (n=17), PFF (n=19), CHF (n=12), HC. (n=20)

Data:
The TVS study incorporated collection of clinical and demographic data to characterise the cohort, as well as mobility data across 3 different conditions (Fig 1).
Protocol Paper: Mazzà, C., Alcock, L., Aminian, K., Becker, C., Bertuletti, S., Bonci, T., Brown, P., Brozgol, M., Buckley, E., Carsin, A.E. and Caruso, M., 2021. Technical validation of real-world monitoring of gait: a multicentric observational study. BMJ open, 11(12), p.e050785.

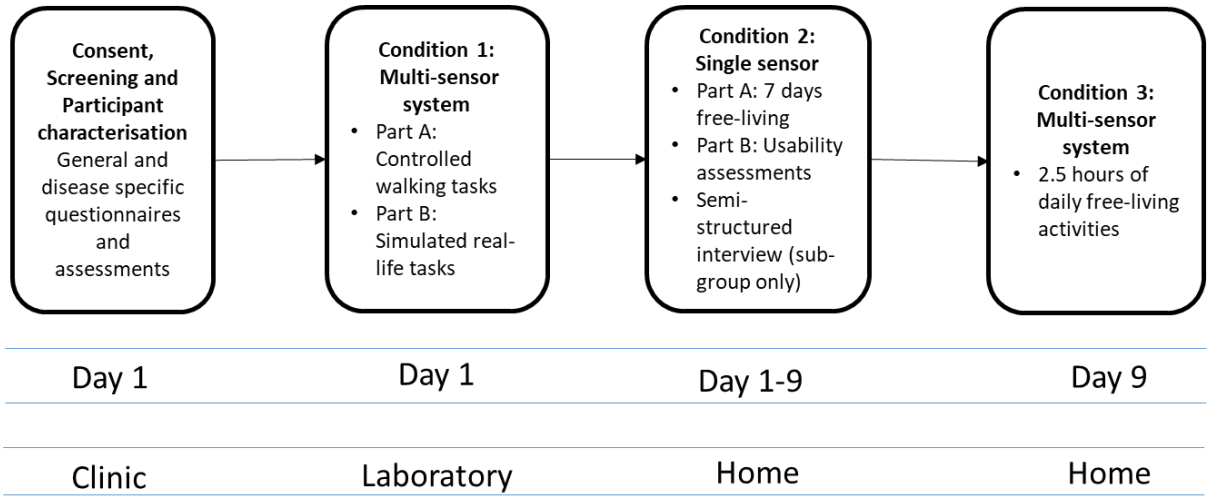


Fig 1 – TVS data capture

Demographic and Clinical Data:
All participants underwent a clinic/laboratory-based session to record generic and disease-specific characterisations. This included participant reported outcomes, assessments and medical notes review.
The following participant reported outcomes were collected: descriptive information (gender, year of birth, living arrangements, education), comorbidities, number of falls and injuries in the last 12-month, walking aid usage and current medication.
Cognitive status was recorded from the **MoCA** (completed during screening). Participants were asked to rate their general pain and pain during walking using a **Visual Analogue Scale** (0-10, from no pain to worst pain possible). Activities of Daily Living (ADL) function was assessed by use of the **Late-Life Function and Disability Instrument (LLFDI)**[5,6]. The LLFDI is a comprehensive assessment of function

and disability in community-dwelling older adults. The functional component (32 items) reflects a person's ability to perform specific actions or activities and the disability component (16 items) reflects a person's ability to perform socially defined life tasks within a typical sociocultural and physical environment.
Participants' height, weight, shoe size and waist circumference were recorded.
Participants then underwent additional disease-specific assessments, with the exception of healthy adults who did not complete additional assessments.

PD Cohort:
The MDS Unified Parkinson's Disease Rating Scale (MDS-UPDRS)[7] Part III (motor examination).

MS Cohort:
The Expanded Disability Status Scale (EDSS)[8] score

COPD Cohort:
The COPD Assessment Test (CAT)
Lung function – Spirometry.
Six minutes walking distance (6MWD) assessment (using a 20-meter walking path).

CHF Cohort:
The Kansas City Cardiomyopathy Questionnaire (KCCQ) score.
Six minutes walking distance (6MWD) assessment (using a 20-meter walking path).

PFF Cohort:
The Short Physical Performance Battery (SPPB). This consists of a static balance task, a five times chair-raise test, and a 4m walk test at preferred gait speed.

Condition 1
In condition 1 study participants performed a series of mobility tasks in a laboratory setting while data were collected from 3 different systems:

1. The 'gold standard' laboratory optical motion capture system
2. The 'silver standard' INDIP multi wearable sensor system
3. The McRoberts Dynaport MoveMonitor index device

The mobility tasks were taken from the following list, with participants' progression through the range of tests being based on individual capability.

1. **Timed Up and Go (TUG) Test** (3 meters)
2. **Straight Walking Test** (completed at three different walking speeds with two trials each).
3. **L Test**[9] (sit in a chair, stand up, walk straight, turn at a curved 90°, continue walking straight, turn again at a curved 180°, follow the circuit back to the chair and sit down).
4. **Surface Test** (walk the circuit by turning around the cones, completed twice and finish at the marked end point).
5. **Hallway Test** (including stair ascent and descent by step up and down off a step).
6. **Simulated Daily Activities*, level one (**a simplified version accommodating the participants who may find some of the tasks in Level two too physically demanding or uncomfortable).
7. **Simulated Daily Activities*, level two** (some complex movements that may not be suitable for all participants but will be useful for the development and validation of the single sensor system-algorithm pair).

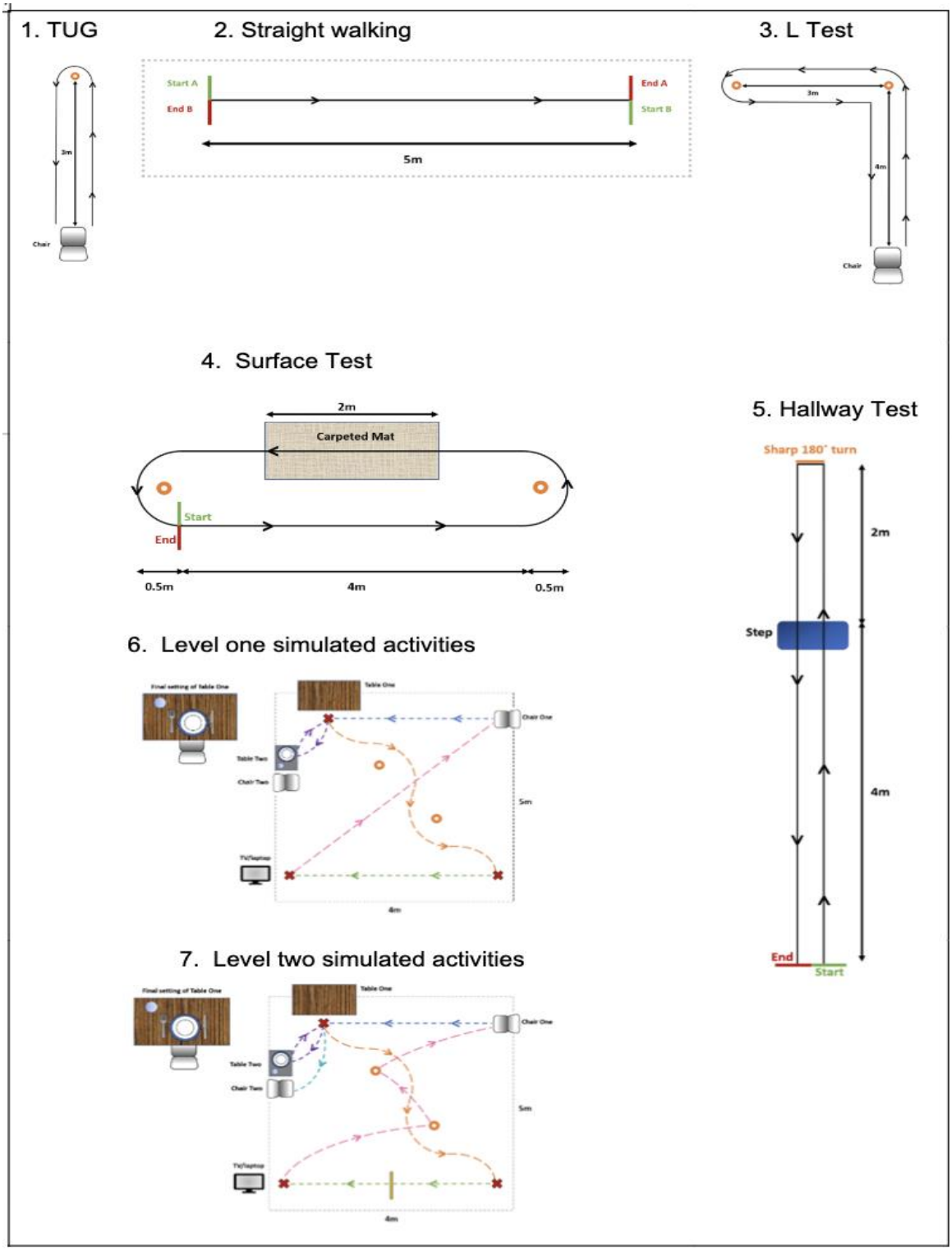


Figure 2. *Illustration of laboratory and procedures during all in-lab tasks.*

Condition 2

Mobility Data: McRoberts Dynaport MoveMonitor data were collected from all participants over the course of 7 days in which they went about their usual daily routines in their home and community.

Contextual Factors Data: Geolocation data and walking aid proximity data (if appropriate) were collected using a combination of a custom written Mobile app and Beacon proximity sensors. These data were subsequently processed to yield contextual labels relating to the following criteria:

Whether walking occurred inside or outside

Whether walking occurred with or without a walking aid

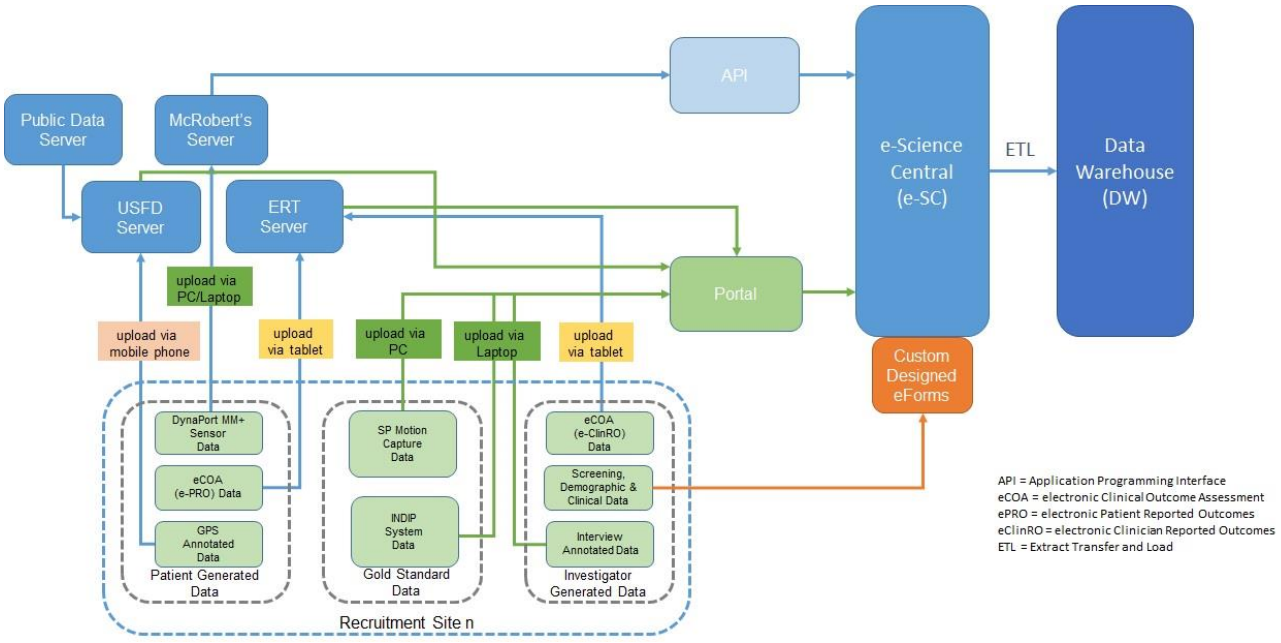
Information about staypoints and frequently traversed walking locations

Condition 3
Mobility Data: McRoberts Dynaport MoveMonitor data and INDIP data were collected from all participants over the course of 2.5 hours in which they were asked to attempt a range of mobility tasks in their home and community.
The participants were instructed to "Go about your day as you usually would but try to include tasks such as:" (see list below for details).

Examples of activities to be performed in a self-chosen order
Walking indoors, both short and longer distances (e.g. within one room and then moving around in the house, including corridors if possible)
Rise from a chair and walk to another room
Walk to the kitchen to get something to drink
Walking up and down stairs (if possible)
Walking outdoors (if possible)
If walking outdoors, try including walking up and down an inclined path
Note: Participants are asked later to indicate what activities they did during this period

The data flow diagram for the TVS is outlined below.

TVS Data Flow



II - Clinical Validation Study:

The clinical validation study dataset incorporates a range of demographic, clinical and mobility data from 4 clinical cohorts (PD, MS, COPD, PFF) over 5 measurement periods. Each cohort consists of 600 participants.
Protocol paper: https://www.medrxiv.org/content/10.1101/2022.05.25.22275598v1

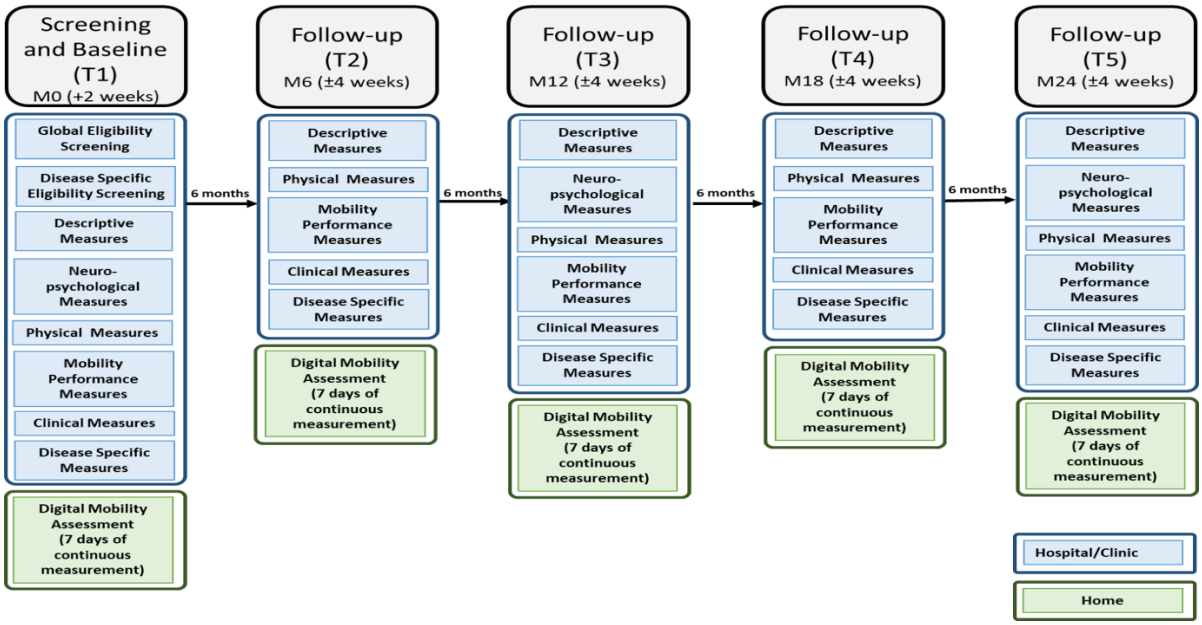
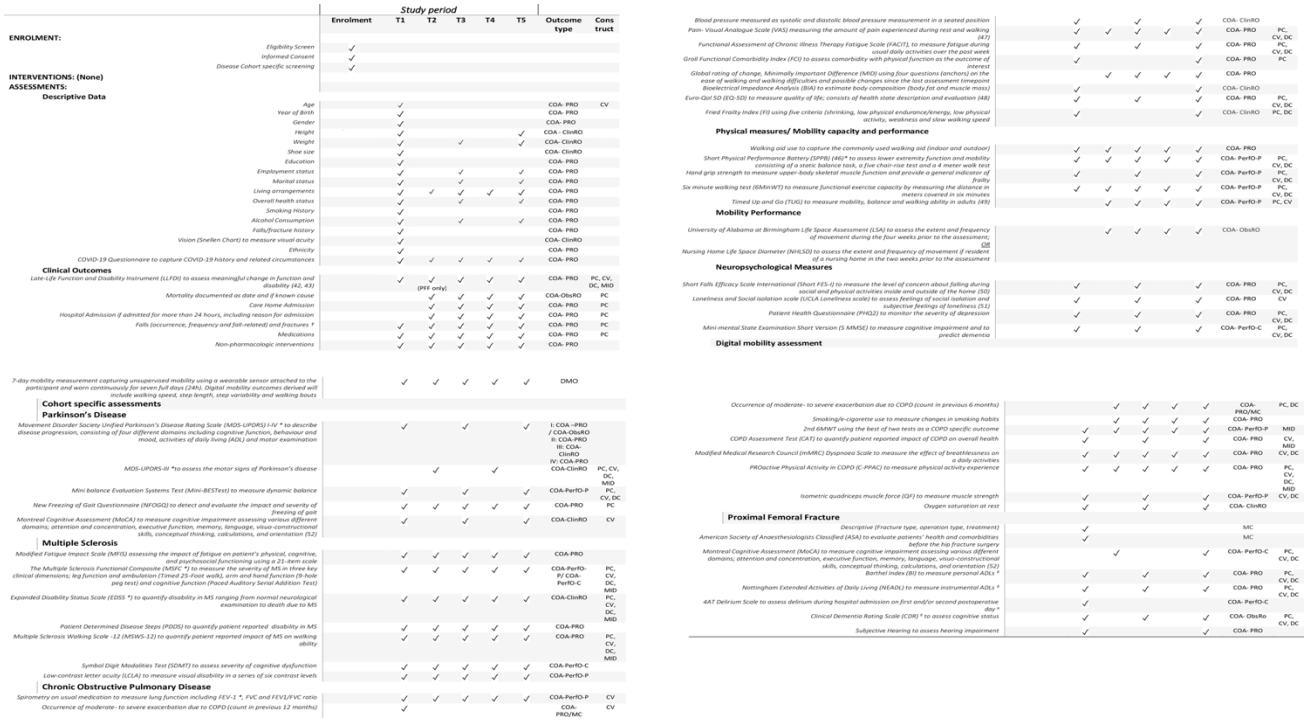


Figure 3. Assessment flow for Clinical Validation Study

Demographic and Clinical Data:

The clinical and demographic data that is captured from all cohorts, and the cohort specific data, is described in Table 1 below.

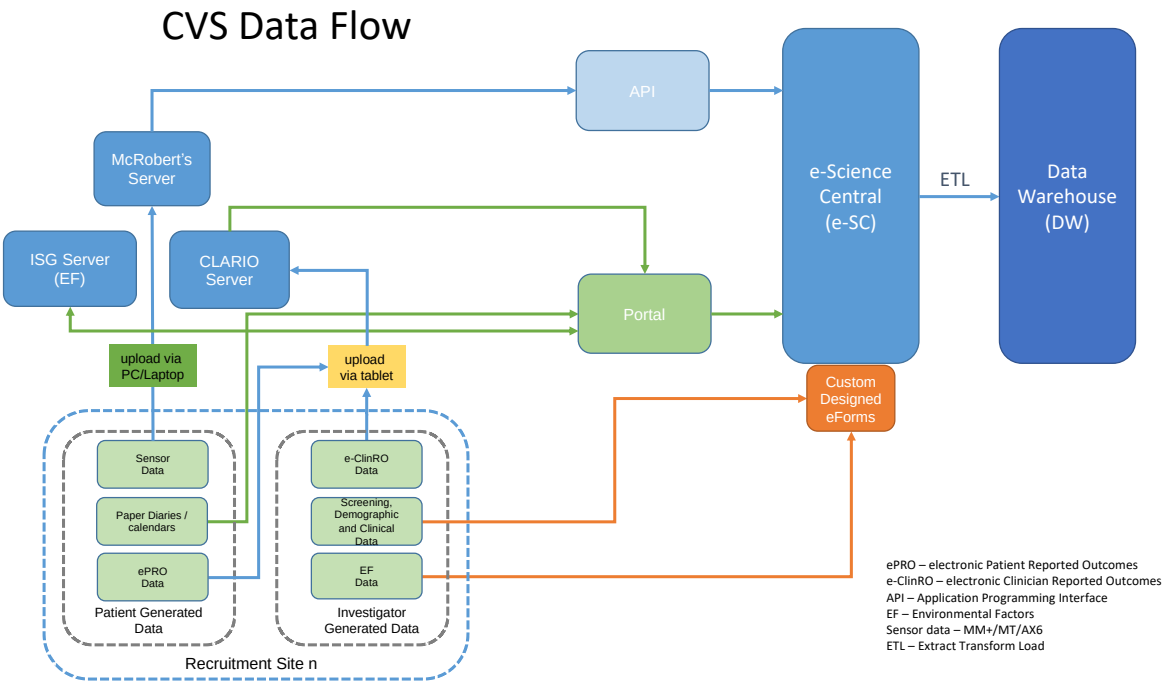


T1, Screening/Baseline; T2, 6 month assessment; T3, 12 month assessment; T4, 18 month assessment; T5, 24 month assessment; *, indicates key (primary) cohort specific outcome measure; SPPB, short physical performance battery – PFF key primary cohort specific outcome measure; † falls and fracture data are collected retrospectively, 12 month retrospective at T1 and 6 month retrospective at T2-T5; β, pre-fracture status is measured at T1, current status is measured at T3 and T5; α, only applicable to acute patients;

O, type of outcome measure in accordance with FDA terminology; COA, clinical outcome measure – describes or reflects how a patient feels, functions, or survives; PROM, Patient-reported outcome; ObsRO – Observer-reported outcome; ClinRO, Clinician-reported outcome; PerfO, Performance-based outcome; PerfO-P, Performance-based outcome physical measure; PerfO-C, Performance-based outcome cognitive/mental measure; C, validation construct assessed; PC, predictive capacity; CV, construct validity; DC, detect change over 24 months; MID, Minimum Important Difference; MC, medical chart;

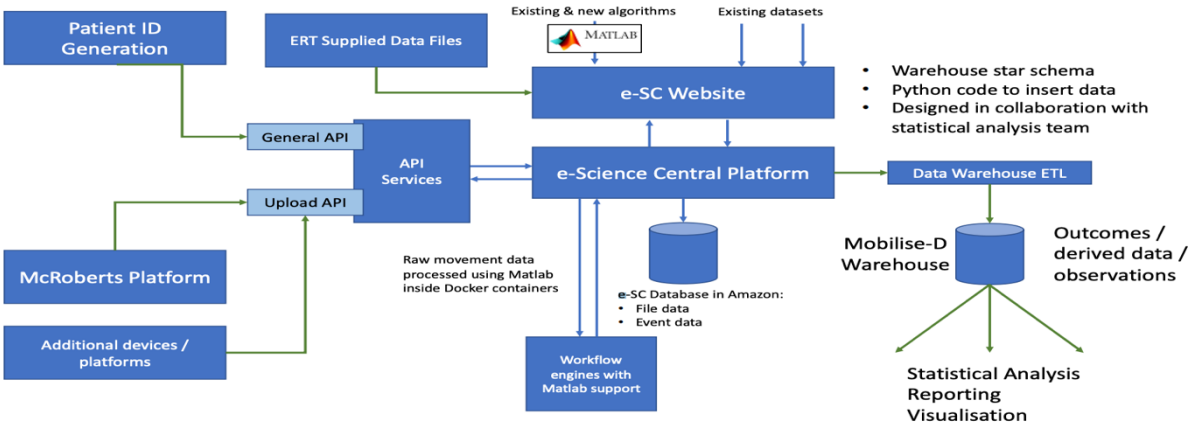
Mobility Data:
At each measurement period (T1-5) mobility data are captured from each participant over 7 days as they go about their habitual daily routines using one of the McRoberts Dynaport MoveMonitor or Axivity AX6 wearable sensors.

Data Flow:
The data flow for the various data sources in the CVS is outlined below.



3 e-Science Central Platform

The data requirements for the Mobilise-D programme (including the TVS and CVS) have been addressed using the open source e-Science Central (e-SC) platform, an open source platform that was developed to enable cloud based scientific data ingestion, storage, analysis and collaboration (2). An implementation of e-SC is being used in Mobilise-D to collect, process and analyse all data associated with the TVS and CVS (see below).



The platform consists of two main components. Firstly a secure, scalable, cloud-based platform to ingest and process data uploaded from a variety of sources including the McRoberts and Axivity sensors, the clinical evaluation forms, and environmental/contextual data sources. Features extracted from this data are then loaded into a Data Warehouse with a novel schema designed specifically for study data. This allows appropriate consortium investigators scientists to explore, analyse and visualise this data in a variety of different ways. A key aspect of the warehouse design is that it also stores metadata describing the types and format of the data. This enables automatic report generation, exploratory data analysis and error checking. The overall result is a flexible, general purpose system that is open-source and uses the cloud for scalability (3).

4 Data Heterogeneity & Scale

The platform must deal with heterogenous data types that are ingested from different sources. Furthermore, it must do this at scale – **by the end of the Mobilise-D programme over 100TB of data will have been processed by the platform.**

Data was collected from multiple sources during both the TVS and CVS studies and ultimately persisted in the e-SC based Data Management Platform. These sources included:

- Accelerometer data uploaded automatically via an Application Programming Interface (API) endpoint
- eCRF collected via tablets provided by Clario that was uploaded manually as .csv files as the studies progressed
- JSON form data collected directly by the e-SC platform for cases where there were no commercial eCRF forms available

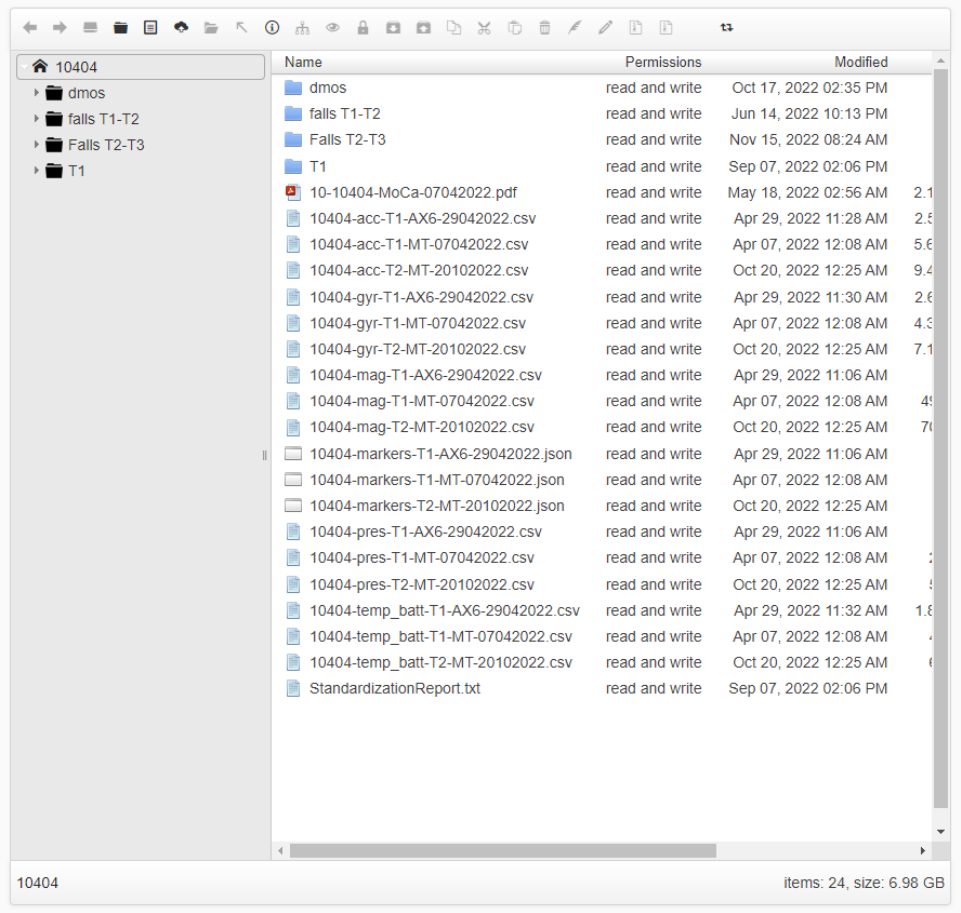
2. Hiden H, Woodman S, Watson P, Cala J. Developing cloud applications using the e-Science Central platform. *Phil. Trans. R. Soc. 2013 A.***371**20120085http://doi.org/10.1098/rsta.2012.0085

3. P. Watson and H. Hiden, "The e-Science Central Study Data Platform," *2022 IEEE 18th International Conference on e-Science (e-Science)*, Salt Lake City, UT, USA, 2022, pp. 55-64, doi: 10.1109/eScience55777.2022.00020.

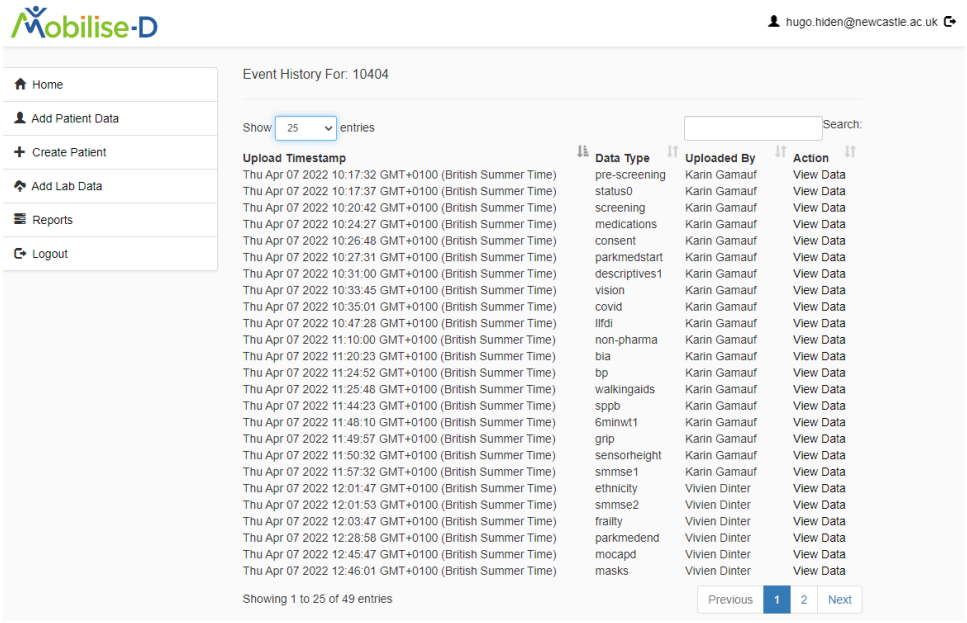
- Additional files such as falls diaries that were stored directly as files
- Derived DMO data

Because of the heterogeneous nature of the data, the e-SC platform performed the role of a “Data Lake” that stored and organised the various data sets.

For file based data, each participant had a “Folder” for file storage:



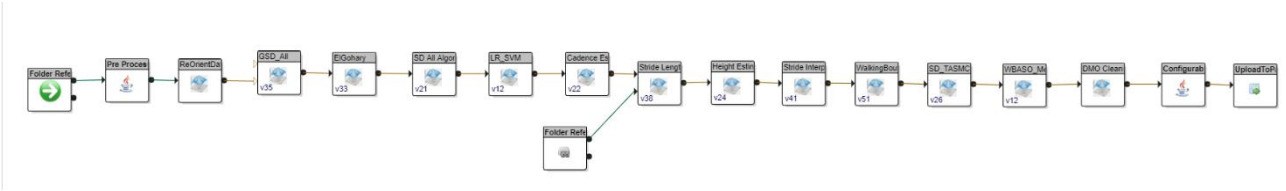
For JSON form data, participant forms were stored directly as JSON within a PostgreSQL database:



This data was combined into the data warehouse after the processing of each visit to produce a unified data set that was used to supply data to the statistical analysis WP members.

5 Data Processing

DMOs were calculated using a multi-step pipeline that combined the algorithms identified during the TVS with selected demographic data derived from the eCRF forms that were implemented within the e-SC platform. A typical processing pipeline is shown below:



This pipeline calculates DMO data for a single day of CSV data collected for one visit of a single participant within a cohort. The early stages in the pipeline deal with various pre-processing tasks such as gathering any meta-data required by the algorithms and ensuring that the data is corrected for cases where the device orientation was incorrect. Subsequent steps perform the processing steps identified by the TVS analysis and the final stages store the derived DMO data in the patient's record within the platform.

The processing scale for this work was significant as it required the organisation of multiple visit files and the co-ordination of multiple instances of the processing pipeline. **Each invocation of the pipeline required approximately 20 minutes of processing time and 16GB of RAM.** Given the number of participants and volumes of data, calculating DMO data for all the cohorts within the CVS could not be performed on single workstations in a reasonable time frame.

To process a single visit (T1) for the CVS cohort required 10 virtual servers, each with 8 CPU cores and 32GB of RAM. Processing all study participants' data took approximately 7 days for a single visit. This could have been accelerated by deploying more virtual servers, but 7 days was deemed to be a reasonable time given the additional administrative and monitoring tasks that were required in order to keep the platform operational during the period of intense load.

In total, **the processing of the T1 visit required approximately 6TB of storage space** to store the derived DMO data, standardized accelerometer data sets and temporary working files.

6 Data Sharing

Initially clinical data was made available upon request from the data management team and provided in the form of .csv extracts from the data warehouse. However, with the increase in demand this type of support model became unsustainable, and a tiered level of data access was implemented as below;

- L1. Site representatives with access to local site data only.
- L2. Cohort leads with access to cohort-specific data at multiple sites.
- L3. Analysts with access to entire pseudonymised dataset.

A basic flutter app was developed which uses e-SC credentials for authentication and secrets manager for authorisation (specified L1, L2, or L3). Users can access reports based on their privilege level at any time. However, data is accessible with the caveat that it is not fully cleaned. This is an ongoing process, managed by a Data Cleaning Task Force to provide a 'cleaned' dataset for analysis as per the Statistical Analysis Plan (SAP). A Data Dictionary (DD) was also developed and shared within the consortium which describes every variable in the data warehouse. Sites can use the DD together with the report extracts to identify any data errors or data integrity issues which are addressed by the study monitoring team. Access to on-demand reports also facilitates sub-studies and publications. Policies and procedures have been developed to ensure any dissemination of data and/or code from the platform follows strict guidelines for copyright and licensing. Data is shared on Zenodo using Creative Common 4.0 (CC BY 4.0) license and code is shared on GitHub using Apache 2.0 license.

7 Study Management & Monitoring

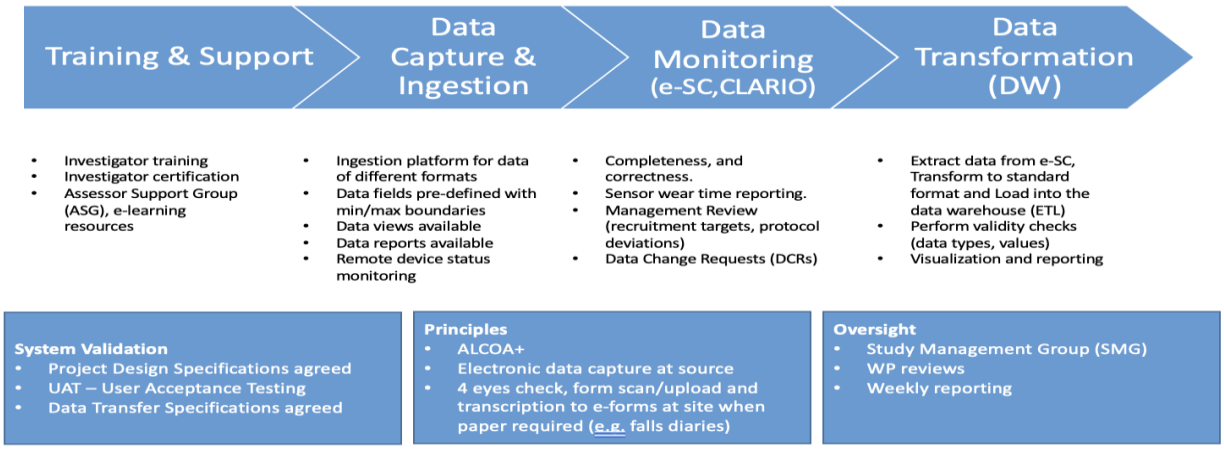
The DMP was implemented within the context of a comprehensive programme of study support and monitoring that was implemented across multiple sites in different countries.

This included:

1. **Creation of educational resources for site investigators.** This included participation in the webinar series that preceded the CVS and the production of several 'how to' guides that were distributed to sites and made available on SharePoint.
2. Provision of support to investigators through participation in the regular **Assessor Support Group** meetings.
3. Provision of ongoing support for issues that might arise via **telephone or email contact**.
4. **Dynamic management of hardware resources** across the sites. This task consumed significantly more resources and effort than was originally forecast due to 2 major issues that arose. Firstly, the sensor partner (McRoberts) was unable to supply the required number of MM+ sensors so we needed to redistribute the available sensors across the sites while securing a supply of additional sensors from Axivity. Secondly, the sheer volume of data that was being handled though the CLARIO platform put this enterprise level platform under severe pressure. The DMP team had to work closely with sites and CLARIO to diagnose the issue and implement appropriate solutions without interrupting the study.
5. Comprehensive study monitoring programme that involved real time monitoring of recruitment rates and data completeness throughout the programme, with appropriate follow up. This monitoring programme is described in detail in the Study Monitoring Plan.

8 Quality Control

A comprehensive Quality Control programme was put in place to ensure that the DMP met its requirements throughout the Mobilise-D programme. This QC framework incorporated measures that were implemented right across the programme from training and support measures, through data capture and ingestion, the monitoring programme, and during the implementation of the processing and ETL phases (see figure below).



As per the Study Monitoring Plan, the Data Manager issues monthly reports to sites with data queries. To correct any issues identified the sites must create a Data Change Request (DCR) and submit to the support team for the relevant platform (e-SC, McRoberts, or Clario). The Data Management Team maintains a log of all DCRs. The majority of DCRs are for form data entry errors on either e-SC or Clario. Initially each error required a single DCR and we have recorded in excess of 400 e-SC DCRs, and less than 100 McRoberts DCRs. However, due to minor changes in form field types (integer to real) and subsequent recalculations we uncovered in excess of 7000 changes required on the Clario platform. This was a significant number and would require significant resources to correct. We worked with Clario to identify each change using a unique ID in their database and complete all changes programmatically and include in a bulk DCR. Audit trail documentation consists of a Memo to File (MTF) that describes the issue and resolution plan for all affected patients and will be included in the study archive.

9 Concluding Remarks

The data management team has implemented a scalable platform that is capable of ingesting and processing multimodal data at scale. The team has had to address evolving data requirements throughout the CVS in particular, and has even needed to accommodate an additional sensor platform being introduced to the CVS. Despite this they have managed to meet all of the requirements and constraints necessary to deliver the Mobilise-D research programme and maintain the level of data integrity that is necessary to underpin the digital endpoint qualification process. The team will continue to refine and improve the DMP and associated processes as we move into the final stages of the Mobilise-D programme.