Mobilise-D Statistical analysis plan for the Technical Validation Study

Title	Mobilise-D Statistical analysis plan for the Technical Validation Study

Statistical Analysis Plan Author

Encarna Micó Amigo, Newcastle University (UNEW)

Anne-Elie Carsin (ISGlobal)

Sarah Koch (ISGlobal)

Tecla Bonci, The University of Sheffield (USFD)

Andrea Cereatti (UNISS)

Francesca Salis (UNISS)

Rana Rehman, Newcastle University (UNEW)

Cameron Kirk (UNEW)

Alma Cantu (UNEW)

Brian Caulfield (UCD)

Claudia Mazzà, The University of Sheffield (USFD)

Judith Garcia-Aymerich, ISGLobal

Aida Aymemir, Merck

Silvia Del Din, Newcastle University (UNEW)

Date and Version

10/12/2020 - Version 1

27/01/2021 - Version 2

15/05/2021 - Version 3

06/07/2021 - Version 4

24/08/2021 - Version 5

08/10/2021 - Version 6

16/12/2021 - Version 7

SIGNATURE PAGE

	Name	Role	Signature	Date
First draft written by	Silvia Del Din Encarna Micó Amigo	Newcastle University Academic Track (NUAcT) Fellow, UNEW Research Associate, UNEW		16/12/2021
Modified by				
Additionally approved by				

AMENDMENT HISTORY

Unique Identifier for Version	Date of the Version	Author	Brief description of change from the Previous Version
1.0	10/12/2020	Silvia Del Din Rana Rehman Encarna Micó Amigo	First draft, taking into account previous input from WP6 on pre-technical SAP (Sarah Koch, Anne-Elie Carsin, Judith Garcia-Aymerich and Aida Aymemir, Merck) and from Claudia Mazzà.
2.0	17/01/2021	Encarna Micó Amigo Silvia Del Din	First draft, taking into account previous input from WP6 on pre-technical SAP (Sarah Koch, Anne-Elie Carsin, Judith Garcia-Aymerich and Aida Aydemir, Merck) and from Claudia Mazzà: adaptation of comments for common sections.
3.0	15/05/2021	Silvia Del Din Encarna Micó Amigo	New version taking into account input from WP6 (Sarah Koch, Anne-Elie Carsin, Judith Garcia-Aymerich and Aida Aydemir, Merck), WP3 (Brian Caulfied for the contextual factor analysis section) and from Claudia Mazzà, Tecla Bonci and Francesca Salis.
4.0	06/07/2021	Silvia Del Din Encarna Micó Amigo	New version based on the feedback from Anne-Elie Carsin, Aida Aydemir, Tecla Bonci, Claudia Mazzà, Sarah Koch and Judith Garcia-Aymerich, from the last discussions set on the 6th and 11th of May. Changes to the experimental protocol section (in line and consistent with TVS protocol paper). General re-structure, more details and figures added about granularity/ aggregation level of descriptive statistics used for performance metrics and statistical analyses (both in section 3.2.2 and 3.2.1). Addition of sub-analysis section “Impact of different single sensor device electronics (same capabilities) and sensor wearing modality on results”. Addition of Appendix C.
5.0	24/08/2021	Silvia Del Din Encarna Micó Amigo	Last changes after final round of feedback.

Unique Identifier for Version	Date of the Version	Author	Brief description of change from the Previous Version
6.0	08/10/2021	Silvia Del Din Encarna Micó Amigo	Changes and additional comments after latest feedback.
7.0	16/12/2021	Silvia Del Din Encarna Micó Amigo Cam Kirk	Changes in text, figures and tables, addressed additional comments after latest feedback from Aida and colleagues during the meeting held on 25/10/2021. Removed what was Appendix C (further example) and kept Appendix B as agreed. Appendix C now is the Appendix D of version 6.0 (data used in the visualisation tool).

LIST OF ABBREVIATIONS OF TERMS

Term	Definition
CABG	Coronary Artery Bypass Graft
CAT	COPD Assessment Test
CAU	University of Kiel, Germany
CE	Cadence Estimation
CI	Confidence Interval
CHF	Congestive Heart Failure
COPD	Chronic obstructive pulmonary disease
CRTD	Cardiac Resynchronization Therapy Device
CVS	Clinical Validation Study
DMOs	Digital Mobility Outcomes
EDSS	Expanded Disability Status Scale

FAU	University of Erlangen-Nuremberg
FEV-1	Forced Expiratory Volume in the first second
FN	False Negative
FP	False Positive
FVC	Forced Vital Capacity
GPS	Global Positioning System
GUI	Graphical User Interface
HA	Healthy older Adult
HTA	Health Technology Agencies
ICC	Intra-class Correlation Coefficient
INDIP	INertial module with DIstance Sensors and Pressure insoles
KCCQ	Kansas City Cardiomyopathy Questionnaire
LLFDI	Late-Life Function and Disability Instrument
MDS	Movement Disorder Society
MoCA	Montreal Cognitive Assessment
MS	Multiple Sclerosis
PCI	Percutaneous Coronary Intervention
PD	Parkinson’s Disease
PFF	Proximal Femoral Fracture
PPV	Positive Predictive Value
RBMF	Robert Bosch Foundation for Medical Research, Germany
RS	Reference System
SAP	Statistical Analysis Plan

SD	Standard Deviation
SL	Stride Length
SO	Secondary Outcomes
SP	Stereophotogrammetric System
SPPB	Short Physical Performance Battery
SRTM	NASA's Shuttle Radar Topography Mission
SS	Single Device System
SSL	Secure Socket Layer
TASMC	Tel Aviv Sourasky Medical Center, Israel
TLS	Transport Layer Security
TN	True Negative
TP	True Positive
TUG	Timed Up and Go
TVS	Technical Validation Study
TW	Tolerance Window
UNEW	Newcastle University, UK
UPDRS	Unified Parkinson's Disease Rating Scale
USDF	The University of Sheffield, UK
WB	Walking Bout
WP2	Work Package 2 (Mobilise-D)
WS	Walking Speed

List of Contents

1. Introduction

1.1 Context of Mobilise-D

Mobilise-D ("Connecting digital mobility assessment to clinical outcomes for regulatory and clinical endorsement") is an EU-funded IMI2-JU consortium that aims to develop and implement a digital mobility assessment solution to demonstrate that real-world digital mobility outcomes (DMOs) can successfully predict relevant clinical outcomes and provide a better, safer and quicker way to arrive at the development of innovative medicines.

The overarching objectives of MOBILISE-D are to:

i. deliver a robust, validated, technology independent solution, within a standards-based framework, for digital mobility assessment;

ii. provide evidence that the outcomes accurately measure and monitor disability and predict clinical outcomes in four of the most relevant mobility-limiting conditions: chronic obstructive pulmonary disease (COPD), Parkinson's disease (PD), multiple sclerosis (MS) and proximal femoral fracture (PFF);

iii. obtain regulatory, payer and health technology agencies (HTA) approval of DMOs. In addition, EFPIA partners will conduct exploratory studies in congestive heart failure (CHF) using digital technology to measure DMOs and will employ the validated algorithm to further the assessment of this approach.

The objectives of Mobilise-D will be achieved through two phases: a technical and a clinical validation phase.

During the first 2 years, Mobilise-D will carry out the technical validation (WP2) of a wearable device to quantify real-world walking speed and other DMOs. The wearable device includes sensors as well as a robust and accurate algorithm to derive DMOs. 120 participants will be recruited from six different groups across five clinical sites. These cohorts include: Chronic Obstructive Pulmonary Disease (COPD), Parkinson's disease (PD), Multiple Sclerosis (MS), Proximal femoral fracture (PFF), Congestive Heart Failure (CHF) and healthy older adults (HA).

From Year 3 to Year 5, Mobilise-D will carry out the clinical validation of DMOs in four patient cohorts of slow walkers (COPD, PD, MS, PFF).

1.1.1 Technical Validation Study (TVS)

The Technical Validation Study (TVS) is a multi-centre study with five clinical sites in three countries. Participants will be recruited at clinical sites that all have access to the populations of interest and have the capability to conduct technical studies. The five sites are: Tel Aviv Sourasky Medical Center (TASMC), Israel, Robert Bosch Foundation for Medical Research (RBMF), Germany, University of Kiel (CAU), Germany, Newcastle University (UNEW), UK, and The University of Sheffield (USFD), UK.

The primary objective of the TVS is to conduct a technical validation of a sensor and algorithm combination capable of measuring walking speed and other DMOs (Figure 1). Criterion validity will be evaluated by comparing these DMOs to those evaluated by reference systems (stereo-photogrammetric or INertial module with DIstance sensors and Pressure insoles (INDIP) system), in five clinical cohorts (COPD, PD, MS, PFF, and CHF) and in healthy adults.

A secondary objective is to evaluate sensor usability and acceptability. This will be achieved by obtaining perspectives from both the participants and the researchers conducting the assessments.

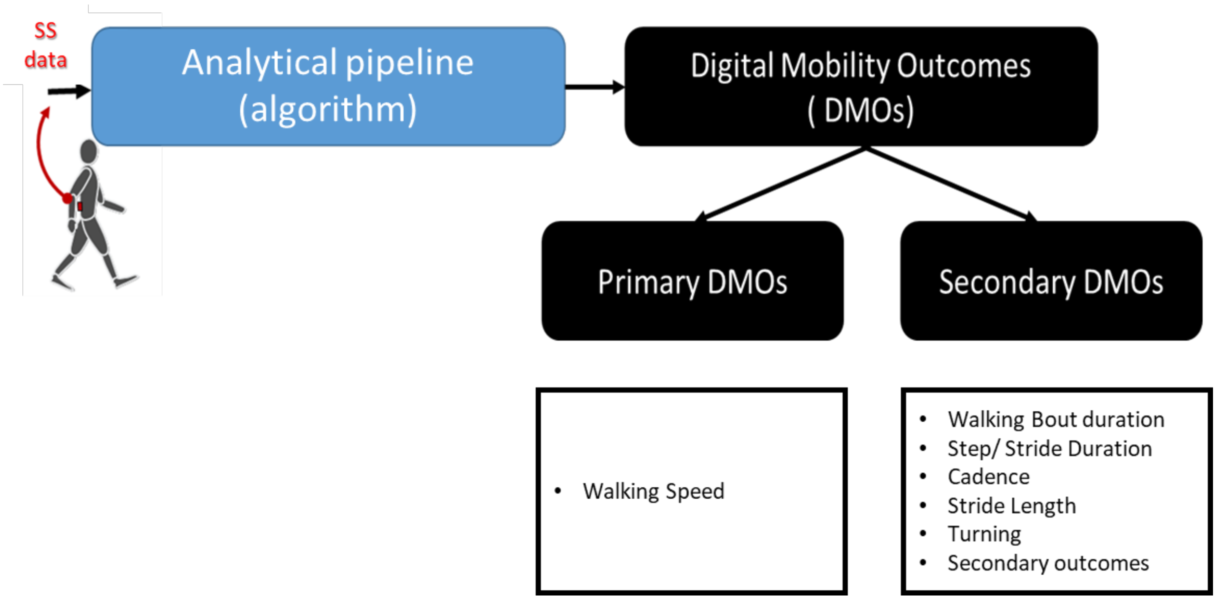


Figure 1. Overall schematics for estimation of primary DMOs and secondary DMOs from a device, single sensor (SS) / algorithm (pipeline) pair.

To assess criterion validity of the single sensor (SS)/algorithm pair both for the in-lab and real-world assessments, the testing protocol will include three different experimental assessments (see Table 1). Further detail can be found in the TVS protocol document (Deliverable 2.3).

1.1.2 Experimental Protocol

The following protocol reflects changes that have occurred since the submission of the first version of the TVS study protocol (Version 1.0 15 August 2019) when finalised list of DMOs and other details had not been finalised and agreed upon. Amendments and changes to the protocol have been since implemented (the current version of the TVS protocol is v1.5 24 September 2021) and are reflected in this section.

The performance of the analytical pipeline (algorithm) to determine walking-related DMOs is mostly affected by three factors:

- the type of motor task (e.g. slow as opposed to fast, straight as opposed to curvilinear or inclined walking, etc.)
- the population of interest (e.g. healthy vs pathological gait)
- the context of observation (e.g. home vs outdoors)

To accommodate these factors, we have developed a comprehensive, multi-stage protocol that includes a variety of tests conducted both in a laboratory context and in the real world (Table 1).

Table 1. Summary of the experimental protocol and reference systems used for the validation of the algorithms in the laboratory and in the real world.

Context of assessment	Reference Systems	Tested device	Mobility Tasks
Laboratory	Stereophotogrammetry (SP)	DynaPort MM+ INDIP	Structured mobility tasks and daily living activities
Real world (2.5 hours)	INDIP[1] Mobile Phone with Aeqora App[2]* Beacon**	DynaPort MM+	Unsupervised real-world activities (including predetermined tasks)
Real world (7 days)	Mobile Phone with Aeqora App* Beacon**	DynaPort MM+	Unsupervised daily living

[1] INDIP= INertial module with DIstance Sensors and Pressure insoles.

[2] https://play.google.com/store/apps/details?id=uk.ac.shef.oak.aeqora&hl=en&gl=US

*For indoor/outdoor the Mobile Phone with Aeqora App is the reference system

**For walking aid use the Beacon is the reference system

1. Laboratory based assessment

Laboratory-based observations will be used to quantify validity and consistency within and between groups and different types of walking tasks under controlled ideal conditions. Structured and task-based mobility activities and a simulated daily activity session, mimicking habitual movements performed at home or at work will be included. The outcome of this comparison will provide the level of highest expected accuracy and minimum detectable changes for a given DMOs.

Measurement tools

Reference system

A stereophotogrammetric (SP) system (100Hz) will be used as the gold standard in structured and simulated tests of daily activities to validate the DMOs calculated from the single DynaPort, single sensor (SS) raw data. A bespoke marker set will include four markers on each foot for detecting the gait events and four markers on the lower back device to track the displacement of the SS (Figure 2).

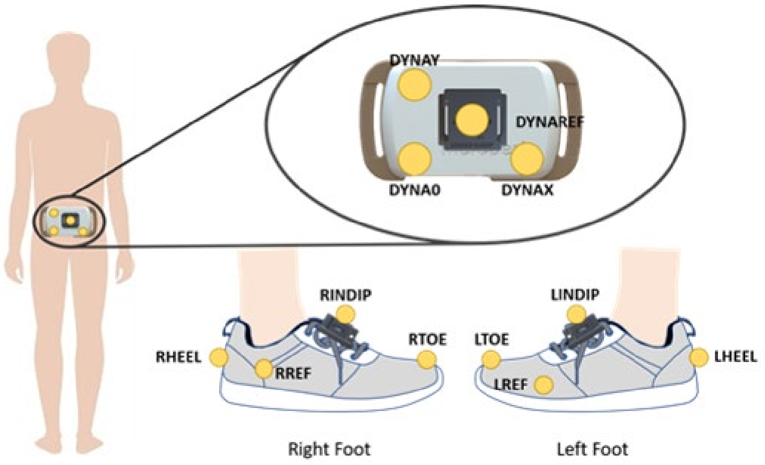


Figure 2. Illustration of the adopted marker set configuration

To ensure quality and consistency in the SP data collection, accommodating different SP systems across sites, a spot check will be used, which will establish each system's specific level of accuracy. A graphical user interface (GUI) for automated pre-processing of the SP data will ensure consistency in associated procedures (labelling, gap filling (1), etc.).

Tested devices

During this observation each participant will also be equipped with an additional multi-sensor system (INDIP) and with the Dynaport device. The INDIP system (Figure 3) includes four inertial modules (one on the lower back, one on the non-dominant wrist and two on the feet), two distance sensors and two force-sensitive resistor pressure insoles including 16 force-resistive sensing elements (manufacturer 221e S.r.l., Italy). The INDIP has been designed to be used as a reference for real-world experiments, and in this phase of the protocol its performance will be validated against the SP system for the populations of interest.

Notes on validity of the INDIP system:
To note is that the individual components of the INDIP system and the associated algorithms for the estimates of the DMOs have already been extensively validated in previous studies on various healthy and pathological cohorts (2-4), while the final assembled system in its fully synchronized configuration was being developed to address the specific project' s needs.

It should be considered that gait events (initial and final contacts) are directly detected based on the use of multi-sensor instrumented insoles which provide a direct measure of the forces resulting from the foot-ground interaction, thus representing a gold standard for real-world GEs detection. Spatial parameters (i.e. stride length) are estimated from the inertial modules attached to the feet based on state of the art algorithms which exploit zero-velocity update technique and optimally filtered and direct and reverse integrated technique (OFDRI) for cumulative error reduction. The performance of the adopted method was firstly validated on four different healthy and pathological populations with and without walking aids (10 healthy elderly, 10 hemiparetic, 10 Parkinson and choreic gait) and different walking speeds. The stride length was estimated for all subjects with less than 3% error. The same method was further validated on a total of more than 20,000 strides collected on 236 older adults including healthy, Parkinsonian and Mild Cognitive impaired participants collected in a multicentric study. In this second study, stride length and gait velocity mean absolute errors were on average 2% (≈ 25 mm).

In a very recent study in which the same inertial modules integrated in the INDIP system were used, percentage errors for stride length were 1.9%, 2.5% and 2.6% for comfortable, slow, and fast straight walking conditions, respectively. Finally, regarding the estimation of the spatial parameters, it has been shown that the OFDRI technique is the best performing among the double integration methods for mobile gait analysis.

The new 'assembled' configuration is expected to perform equivalently and as such we can anticipate mean absolute percentage errors of 1% on the stride duration, 2-3% in the estimate of the walking speed, and 2-3% in the estimate of the stride length. Importantly, though, it needs to be noted that in the above presented studies statistical analysis, performance metrics (e.g. errors) and experimental protocols in which the INDIP system has been tested are different from the ones presented in this SAP. While we expect these values to be confirmed in the present study, we will still use the information on the accuracy of the INDIP that will be quantified in the lab-based scenario as a reference to establish the system performance and will use those as our minimal detectable differences also for the real-world scenario. Preliminary results on 20 healthy adults, 16 PD (Parkinson disease), and 12 MS (multiple sclerosis) showed median absolute percentage errors for stride duration, walking speed, and stride length of ≤0.8%, ≤3.7%, and ≤2.3%, respectively.

To formally test this, spatio-temporal parameters will be estimated exploiting the sensors redundancy and implementing previously validated sensor fusion algorithms. The lower back INDIP unit and the DynaPort will be rigidly attached to each other. The data from the SP (100Hz), INDIP (inertial modules and insoles, 100Hz, Distance sensor, 50Hz) and DynaPort (100Hz) systems will be synchronised using a hardware-based approach for the SP and the INDIP system, and timestamps to align recordings from the INDIP and the DynaPort.

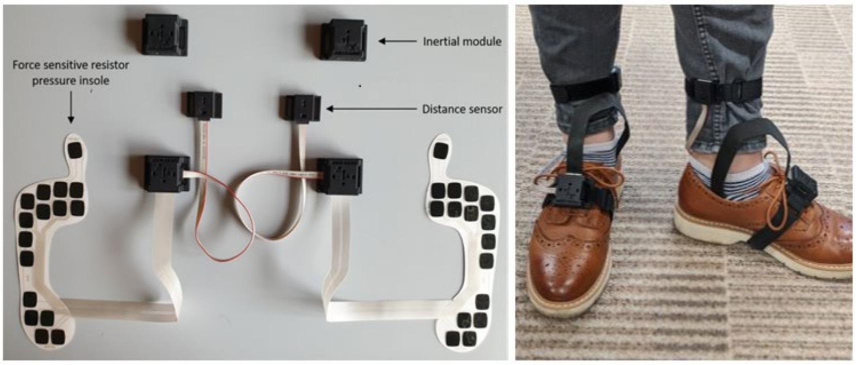


Figure 3. The INDIP system

Mobility tasks

Structured Mobility Tests

The following tests will be performed:

- ***Timed Up and Go (TUG):*** The TUG is a widely used clinical assessment of a person's mobility. The participant is asked to sit in a chair, stand up, walk 3 m in a straight line, make a 180° turn, walk back to the chair, turn and sit down (Figure 4 A).
- ***Straight Walking:*** Straight walking is the most common test of walking. The participant walks for a distance of 5m from a standing start and will be repeated at three different walking speeds: preferred, fast and slow (Figure 4 B).
- ***L Test:*** The participant is asked to sit in a chair, stand up, walk straight, turn 90° to the left around a cone, walk straight to the second cone, make a 180° turn to the left, walk straight before making a final 90° turn to the right and return to the chair to sit down (Figure 4 C). Besides being a clinically validated test, the main purpose of including this test is the variation in curvilinear walking and the inclusion of different types of turns.
- ***Surface Test:*** The participant walks around a defined circuit, including over a carpeted mat (2m in length), by turning around the cones (Figure 4 D). The circuit is completed twice, creating the longest walking bout out of all the tasks (approx. 20m).
- ***Hallway Test:*** The participant walks along a 6m walkway stepping up and down a step positioned in the walkway. At the end of the walkway, the participant will complete a sharp 180° turn and walk back along the walkway (again stepping up and down off the step) until reaching the end point of the test (Figure 4 E).

Use of arm rests for the TUG and L-test and hand rails for the hallway test is permitted when needed.

Daily living activities

These lab-based tasks will be used to simulate daily activities expected in the real life, similar to previous studies. The participant starts by sitting in chair one and then executes a series of daily living tasks while sitting and moving around the room (see Figure 4 F and G).

Tasks include activities such as replacing objects, setting the table, drinking a glass of water and eating a cookie.

To ensure patient safety and comfort, two variations of the simulated daily activity session have been created.

Level Two (Figure 4 F) consists of some complex movements that may not be suitable for all participants but will be crucial for the development and validation of the single sensor system/algorithm pair.

Level One (Figure 4 G) is a simplified version accommodating the participants who may find some of the tasks in Level Two too physically demanding or uncomfortable.

Level Two daily-life activities mimic movements performed at home, and are split into seven steps and instructions as follows:

1. "Stand up from chair one and walk to table one."

2. "Set table one for dinner using the supplies on table two (cup, plate, napkin and cutlery)."

3. "Carry/move chair two to table one and sit down at table one for a break (minimum 3-5 minutes)." A refreshment of their choice and a cookie will be offered to the participant during this break.

4. "Stand up from the chair and walk to the corner diagonally opposite while walking around the cones in the middle of the floor."

5. "Walk to the laptop/TV while stepping over the taped lines on the floor."

6. "Stand and watch a 3-minute video on the TV/laptop." The participant will be advised that if they wish to sit before the 3 minutes is over a chair will be brought to them by a researcher.

7. "Pick up all the cones on the floor and sit down back in chair one."

Level One daily-life activities mimic movements performed at home, and are split into seven steps and instructions as follows:

1. "Stand up from chair one and walk to table one."

2. "Set table one for dinner using the supplies on table two (cup, plate, napkin and cutlery)."

3. "Sit down (chair two will carried to correct position by researcher) at table one for a break (minimum 3-5 minutes)." A refreshment of their choice and a cookie will be offered to the participant during this break.

4. "Stand up from the chair and walk to the corner diagonally opposite while walking around the cone in the middle of the floor".

5. “Walk to the laptop/TV while stepping over the taped line on the floor.”

6. “Stand and watch a 3-minute video on the TV/laptop. The participant will be advised that if they wish to sit before the 3 minutes is over a chair will be brought to them by a researcher.”

7. “Sit back in chair one.”

The task set up for the participant can be a mix of the two levels. By tailoring the task to the ability of the participant the data necessary for the validation can be collected while still accounting for the participants safety and wellbeing. The set up for each participant is being recorded and noted.

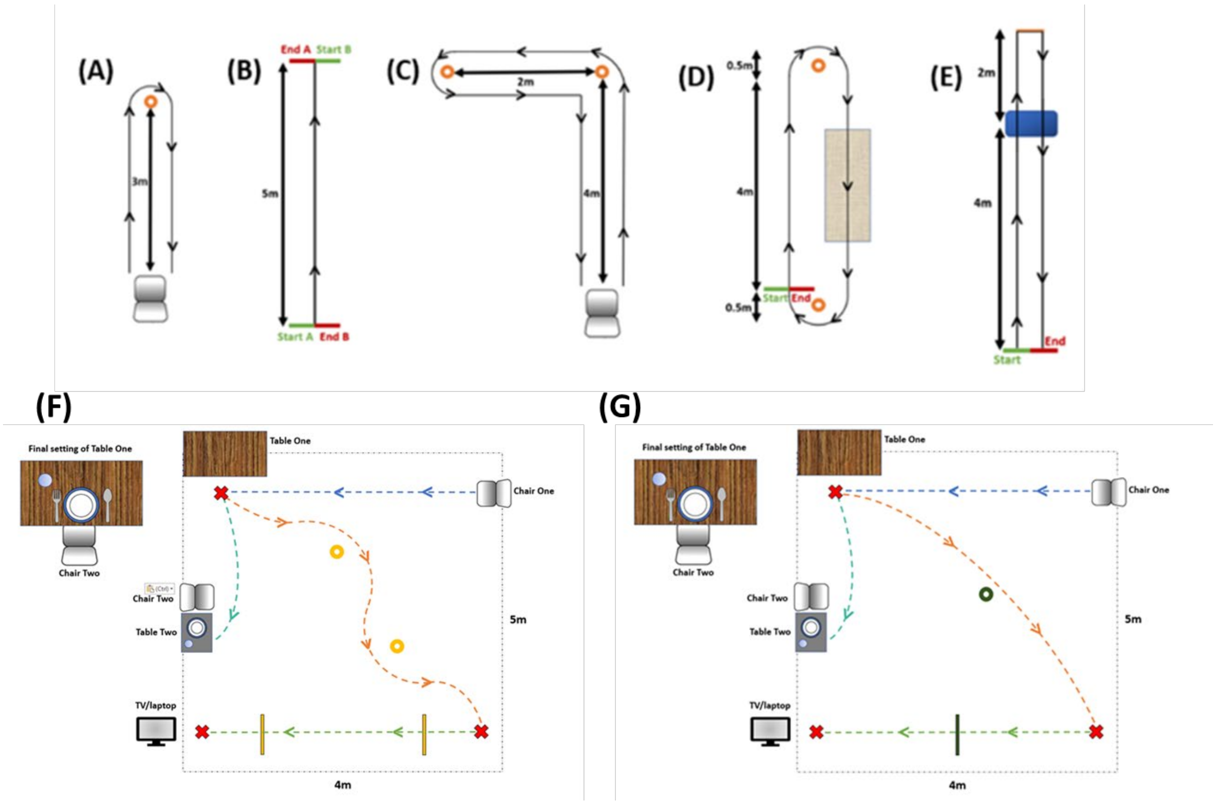


Figure 4. Diagrams of the selected tasks: (A) Timed Up and Go, (B) Straight Walking Test, (C) L-Test, (D) Surface Test, (E) Hallway Test, (F) Daily living activities level 2, (G) Daily living activities level 1.

2. Real - world (2.5 hours observation) assessment

This phase of the protocol will quantify validity and consistency across individuals and different types of walking tasks in the real world. It will be performed in a habitual environment (home/work/community) chosen by the participants. The duration of the observation has been established as trade-off between experimental, clinical and technical requirements.

Measurement tools

Reference systems

Participants will be asked to wear the INDIP, which in this phase of the protocol will be used as a reference system for the quantification of the DMOs provided by the single sensor algorithms, as applied to the DynaPort data.

Contextual Factors

In order to quantify the effects of contextual confounding factors, the participants will also be provided with a system detecting outdoors walking, gradient of descent/ascent (walking uphill/downhill). The system is developed as a mobile Android application (Aeqora app) and the device selected was a Samsung S9 with Android 10. The app is composed of three parts: (i) the core tracker, (ii) the interface and (iii) the server infrastructure collecting data across users. The core tracker, adapted from a library developed by the University of Sheffield, utilises the mobile phone's internal sensors to compute the type of activity (e.g. walking) and intensity (e.g. cadence) to identify geo-located bouts of movement. It operates in the background and senses mobility features through a range of sensors (e.g. step counters, activity recognition, accelerometer, gyroscope, etc.) as well as from location services (GPS, network, Bluetooth, etc.). It collects the data and stores the raw sensor data into a local database in real time.

Data collected during the experiments will be sent to a cluster of servers that uses algorithms to integrate the phone's data with contextual information about the locations where the participant is walking: where possible walks will be matched to OpenStreetMap roads and paths, to remove GPS noise, the slope variation of each walk is computed on tiles, with a resolution of 5m within the UK (using Ordnance Survey Terrain 5) and 30m in the other locations (using NASA's Shuttle Radar Topography Mission (SRTM) data), indoors and outdoors walking is recognised. Moreover, weather is associated with participant location based on the most proximate weather station.

The use of walking aids will also be monitored in this phase. For this purpose, a Bluetooth beacon (BlueBeacon Tag, BlueUp) will be attached to the walking aid and its activity will be detected by the phone's mobile tracker and saved by the app. The distance between the phone and the Beacon and data from the accelerometer contained in the Beacon will be integrated to determine when the aid is in use.

Tested devices

During this observation each participant will also be equipped with the DynaPort device.

Mobility Tasks (daily living activities)

To capture the largest possible range of activities during this assessment, participants will be guided by the following list of activities to be included: rise from a chair and walk to another room; walk to the kitchen and make a drink; walk up and down a set of stairs (if possible); walk outdoors (if possible, for a minimum of two minutes); if walking outside, walk up and down an inclined path.
No supervision or structure to how these tasks should be completed will be given to the participants.

3. Real - world (7 days monitoring) assessment

This observation will quantify the effects of device wearing time, hourly and daily fluctuations of DMOs, and contextual confounding factors (such as location of the walk, weather, type of housing, etc.).

Measurement tools

Reference systems

The participants will be asked to carry a mobile phone equipped with the Aeqora App. Bluetooth beacons will also be used to track the use of walking aids.

Tested devices

The participants will be asked to wear the DynaPort.

Mobility Tasks

Participants will be monitored continuously for seven days, without any specific instruction being provided, except for that of wearing the provided measurement tools.

1.2 Objectives

The specific objectives of the TVS SAP are to:

- *Primary Objective:* To characterise and evaluate criterion validity performance metrics (including selected criterion (concurrent) validity) of the primary DMOs estimated by the device/ algorithms pair, against the reference systems (RS): **Walking Speed** (WS).
- *Secondary Objective:* To characterise and evaluate criterion validity performance metrics (including selected criterion (concurrent) validity) of the secondary DMOs estimated by the device/ algorithm pair, against the reference systems: **Walking Bout** (WB) number, duration, start and end events**, Step/ Stride** number and duration, **Cadence**, **Stride Length**, **Turning** and **other secondary outcomes** (SO: number of final contact events, swing duration, stance duration).

Statistical analyses

Several statistical analyses and performance metrics will be used to characterise the quantified DMOs and evaluate selected criterion (concurrent) validity metrics. These have been agreed by the Mobilise-D consortium and are described in the objective methodology for assessing concurrent validity domain of algorithms/device locations pairs developed and published by USFD in collaboration with UNEW and UCD (5).

Each of the statistical analyses and performance metrics are tailored to the nature of the specific DMOs and their level of aggregation/ minimum granularity (e.g. step or stride, walking bout, test) and are detailed in section 3.

1.3 Study Population

1.3.1 Study Population and Eligibility Criteria

This multi-centre study is sponsored and coordinated by The Newcastle upon Tyne Hospitals NHS Foundation Trust, UK. Participants will be recruited in five sites across Europe: Tel Aviv Sourasky Medical Center (TASMC), Israel (ethics approval granted by the Helsinki Committee, Tel Aviv Sourasky Medical Center, Tel Aviv, Israel, 0551-19TLV), Robert Bosch Foundation for Medical Research (RBMF), Germany (ethics approval granted by the ethical committee of the medical faculty of The University of Tübingen, 647/2019BO2), University of Kiel (CAU), Germany (ethics approval granted by the ethical committee of the medical faculty of Kiel University, D438/18), The Newcastle upon Tyne Hospitals NHS Foundation Trust (UNEW – NuTH), UK and Sheffield Teaching Hospitals NHS Foundation Trust (USFD), UK (ethics approval granted by London – Bloomsbury Research Ethics committee, 19/LO/1507).

Inclusion and exclusion criteria, grouped by Global cohort (All cohorts) and disease-specific cohort, are summarized in Table 2.

All participants will give written informed consent prior to undergoing a clinic/laboratory-based session to record generic and disease-specific characterisations. This will include participant reported outcomes, assessments and medical notes review.

Table 2 Inclusion and exclusion criteria adopted for the different disease cohorts.

Cohort	Inclusion criteria	Exclusion criteria
All cohorts	-able to walk 4 meters independently with or without walking aids -able to give informed consent -willingness to wear the sensor set-ups during the study -shoe size 36 (3 UK) or above -able to read and write in first language of the respective country -Montreal Cognitive Assessment (MoCA) >15 -available for home /office visit during study period	-occurrence of any of the following 3 months prior to inclusion: myocardial infarction, hospitalization for unstable angina, stroke, coronary artery bypass graft (CABG), percutaneous coronary intervention (PCI), implantation of a cardiac resynchronization therapy device (CRTD) -current medical condition that could interfere with the patient's compliance
COPD	-≥45 years of age -Diagnosis of COPD (post-bronchodilator forced expiratory volume in the first second (FEV_1) to forced vital capacity (FVC) ratio <0.70) -clinical stability, defined as at least 4 weeks without antibiotics and/or oral corticosteroids to treat either a moderate or severe exacerbation -non-smokers, current or ex-smokers with a smoking history equivalent to at least 10 pack years (1 pack year = 20 cigarettes smoked per day for 1 year)	-having undergone major lung surgery (e.g. lung volume reduction, lung transplant) -having a lung tumor -primary respiratory diseases other than COPD (e.g. asthma) - impaired mobility related to non-COPD causes, as judged by the investigator
PD	-aged 18+ years -Diagnosis of PD according to the Movement Disorders Society criteria	-impaired mobility related to non-PD causes, as judged by the investigator
MS	-aged 18+ years -Diagnosis of MS based on the revised McDonald's criteria	-impaired mobility related to non-MS causes, as judged by the investigator
PFF	-65+ years of age -surgical treatment (fixation or arthroplasty) for a low-energy fracture of the proximal femur (ICD-10 diagnosis S72.0, S72.1, S72.2) as diagnosed on X-rays of the hip and pelvis within last 12 months	-impaired mobility related to non-PFF causes, as judged by the investigator
CHF	-≥45 years of age -Diagnosis of chronic heart failure NYHA class II-IV	- history of COPD ≥GOLD III - impaired mobility related to non-CHF causes, as judged by the investigator
HA	65+ years of age	

CHF: Congestive heart failure, COPD: Chronic obstructive pulmonary disease, HA: Healthy older adult, MS: Multiple Sclerosis, PD: Parkinson's disease, PFF: Proximal femoral fracture.

1.3.2 Study Size

A convenience sample of 112 participants will be recruited via their clinical care team or research registers to represent the five disease cohorts (COPD, PD, MS, PFF, and CHF), as well as healthy older adults (HA). 20 participants will be recruited from the PD, MS, PFF and HA cohorts, 17 participants will be recruited from the COPD cohort, and 15 participants will be recruited from the CHF cohort. Each cohort will be recruited across multiple sites to ensure generalizability (e.g. differing cultures and contexts). The original sample size of 120 had been defined according to Consensus-based Standards for the selection of health Measurement Instruments guidelines for measurement properties (COSMIN, https://www.cosmin.nl/). This sample size allowed for 'excellent' methodological quality of non-inferiority studies, and is the one endorsed by the COSMIN checklist, a standardized tool for assessing the methodological quality of studies on measurement properties.

After 50% enrolment, this sample size was reviewed and adjusted to 112 due to the reduced numbers of COPD (17) and CHF (15) participants required (please see section Sample size evaluation).

Sample size evaluation

The original sample size (120, 20 participants per cohort) has been reviewed and calculation has been performed by WP6 using new data and is presented in **Appendix A**. The sample size evaluation confirmed that 20 participants for PD, MS, PFF, HA, 17 participants for COPD and 15 for CHF is a sufficient sample size.

2. Outcomes and Variables

2.1 Digital Mobility Outcomes: primary and secondary DMOs

Several statistical analyses and performance metrics will be used to characterise the quantified DMOs and evaluate selected criterion (concurrent) validity metrics. These have been agreed by the Mobilise-D consortium and are described in the objective methodology for assessing concurrent validity domain of algorithms/device locations pairs developed and published by USFD in collaboration with UNEW and UCD (5).

Each of the statistical analyses and performance metrics are tailored to the nature of the specific DMOs, and their level of aggregation/ minimum granularity (e.g. step or stride, walking bout, test) are detailed in section 3.

To characterise the primary and secondary DMOs evaluated using the SS/algorithm pair, we will evaluate criterion validity performance metrics. Particularly, to enable this comparison, a reference system (RS) is needed for validation of the quantified DMOs. Since RS is not available for the third condition (7-day real world), only data collected during the Laboratory and Real world 2.5 hour assessments will be utilised in this analysis.

For the Laboratory assessment, the stereo-photogrammetric system (SP) will be used as the RS (in two cases, comparing the SS versus the SP, and comparing the INDIP system versus the SP), and for the Real world 2.5 hour assessment the INDIP system will be used as the RS (Table 1, Figure 2). Note that the INDIP system will be used as a reference system for the Real world 2.5 hours assessment after its validity will be confirmed in the Laboratory assessment (Table 1).

Table 3 lists the primary and secondary DMOs that will be evaluated for the Laboratory and Real world 2.5 hour assessments, for each participant and cohort.

As agreed by WP2 (6), the Walking Bout (WB) definition (see section 2.2) will be applied to the tests collected during the Laboratory assessment. In this case, given the short duration of the Laboratory tests, the results from the WBs detected during repetitions of a given test (i.e. trials of a test) will be combined and aggregated. For example, for the Straight Walk test at comfortable gait speed, the WBs detected for each repetition (trial) will be combined into a single WB to evaluate: (i) WB-level parameters (e.g., Walking speed) as the mean value of the two walking speeds evaluated for each WB of the test; (ii) Stride/step-level parameters as the mean, standard deviation (SD) of the DMO, together with the statistical analyses obtained over all accumulated steps or strides of the test.

Table 3 List of primary and secondary DMOs

Outcome block	DMOs	Definition	Units	Minimum granularity (level of aggregation/analysis)
Primary DMOs				
Walking Speed	Walking Speed	Walking Speed: stride speed (ratio between the displacement covered within a WB and the time to cover it), calculated as the mean of stride speed values as follows: $\boldsymbol{WalkingSpeed_i} = \frac{\sum_{k=1}^{n\_STRIDE_i} Stride\_Speed_k}{n\_STRIDE_i}$ where $i = 1, \ldots, n$ are the different WBs, $Stride\_speed_k$ is the stride speed of the k – stride in the relevant i –WB, n_STRIDE_i is the number of qualified strides identified in the relevant i –WB	Meters/second	Walking Bout
Secondary DMOs				
Walking Bout	Number of Walking Bouts	Based on the identification of gait as an activity (yes/no) to a sample level of 0.1 seconds, n	Count	Lab Test/Real-world 2.5 hours assessment
	Walking Bout Start	Start of a walking bout, SWB_i i=1, ...n	Seconds	Walking Bout
	Walking Bout End	End of a walking bout, EWB_i i=1, ...n	Seconds	Walking Bout
	Walking Bout Duration	The time between the start and the end of a walking bout, $DWB_i = EWB_i – SWB_i$ i=1, ...n	Seconds	Walking Bout

Outcome block	**DMOs**	**Definition**	**Units**	**Minimum granularity** **(level of aggregation/analysis)**
Step/ Stride Duration	Initial Contact Events - Number of Events (steps and strides)	Based on the qualified identification of Initial contact events (j), n_IC_{ij} i=1, ...n, j=1..m	Counts	Walking Bout
	Step Duration	Duration between two consecutive initial contact events, $D\_IC_{ij}\_IC_{ij+1}= IC_{ij+1} – IC_{ij}$ i=1, ...n; j=1,..m-1	Seconds	Step, Walking Bout
	Stride Duration	Duration between two non-consecutive (alternate) initial contact events, so between two initial contact events of the same foot $D\_IC_{ij}\_IC_{ij+2}= ICi_{j+2} - IC_{ij}$ i=1, ...n, j=1,..m-2	Seconds	Stride, Walking Bout
Cadence	Cadence	Defined as step frequency within a minute, evaluated as double the stride frequency: $\boldsymbol{Cadence_i} = StrideFrequency_i * 2$ where $i = 1, \dots, n$ are the different WBs Stride Frequency is defined as: $\boldsymbol{Stride\ Frequency_i} = \frac{\sum_{k=1}^{n\_STRIDE_i} \left({}^{60}/_{STRIDE\ d_k} \right)}{n\_STRIDE_i}$ where $i = 1, \dots, n$ are the different WBs, n_STRIDE_i is the number of right and left strides in the relevant i –WB $STRIDE\ d_k$ is the duration [seconds] of the k – stride in the relevant i –WB	Steps/minute	Walking Bout
Stride Length	Stride Length	Average stride length within a walking bout; i.e. the average value of all stride lengths within the walking bout:	Meters	Walking Bout

Outcome block	DMOs	Definition	Units	Minimum granularity (level of aggregation/analysis)
		$\boldsymbol{Stride\ Length_i} = \frac{\sum_{k=1}^{n\_STRIDE_i}(STRIDE\ l_k)}{n\_STRIDE_i}$ where $i = 1, \ldots, n$ are the different WBs, n_STRIDE_i is the number of right <u>and</u> left strides in the relevant i –WB $STRIDE\ l_k$ is the length [meters] of the k – stride in the relevant i –WB		
Turning*	Number of Turns	Based on the identification of turns (yes/no) to a sample level of 0.1 seconds, p	Counts	Lab Test/Real-world 2.5 hour assessment
	Turn Start	Start of each turn within the walking bout "I", ST_{iw} i=1, …n; w=1, …p	Seconds	Turn, Walking Bout
	Turn End	End of each turn within the walking bout, ET_{iw} i=1, …n; w=1, …p	Seconds	Turn, Walking Bout
	Turn Duration	Time between the start and the end of the turns within the walking bout, $DT_I = ET_{iw} - ST_{iw}$ i=1, …n; w=1, …p	Seconds	Turn, Walking Bout
	Maximal Turn Angle	Maximal angle achieved in the turn Max_Angle T_w = Max ($AngleT_w$) w=1, …p	Degrees	Turn, Walking Bout
Other Secondary Outcomes	Final Contact Events – Number of Events/Steps	Based on the qualified identification of Final Contact events within each WB_i i=1, …n N_FC_{ji} I=1..n; j=1, …m	Counts	Step, Walking Bout
	Swing Phase Duration	The time between the Final Contact of a foot and the following Initial Contact of the same foot $Swing\_duration_{ij} = IC_{ij+1} - FC_{ij}$ i=1, …n; j=1,..m-1	Seconds	Stride, Walking Bout

Outcome block	DMOs	Definition	Units	Minimum granularity (level of aggregation/analysis)
	Stance Phase Duration	The time between Initial Contact and the Final Contact of the same foot $\text{Stance\_duration}_{ij} = FC_{ij+1} - IC_{ij}$ i=1, ...n; j=1,..m-1	Seconds	Stride, Walking Bout

*Turning block will only be assessed for the Real world 2.5 hours assessment and validated against the INDIP system.

2.2 Digital Mobility Outcomes: definitions of primary and secondary DMOs

In this section, each of the primary and secondary DMOs that will be quantified on a walking bout level are described.

Primary DMO:

Walking Speed

- Walking Speed is calculated in meters/second as the mean of stride speed values as follows:

$$\boldsymbol{WalkingSpeed_i} = \frac{\sum_{k=1}^{n\_STRIDE_i} Stride\_Speed_k}{n\_STRIDE_i}$$

where $i = 1, \dots, n$ are the different WBs,
$Stride\_speed_k$ is the of the k – stride in the relevant i –WB,
n_STRIDE_i is the number of eligible strides identified in the relevant i –WB

Secondary DMOs:

Walking Bout (WB)

The definition of a WB** was established and agreed by the Mobilise-D consortium (6) and is as follows: *"Walking bout: A walking bout (WB) is a walking sequence containing at least two consecutive strides of both feet (e.g. R-L-R-L-R-L or L-R-L-R-L-R, being R/L the right/left foot contact with the ground). Start and end of a walking bout are determined by a resting period or any other activity (non-walking period). The initial step of a WB follows a non-walking period and the final step precedes the next non-walking period".*

The following criteria are used to identify WB:

Walking Bout Parameter	Value (Mobilise-D consensus)
Minimum number of strides	2 Left + 2 Right
Maximum break duration	3 seconds
Max. Absolute height difference (elevation change)*	a. 0.20 ± 0.05 meter b. 4 consecutive strides (2R+2L) exceed the vertical thresholds

* Absolute height difference = height difference between two successive strides.

For a stride to be included in a walking bout, it must abide by the following criteria**:**

Stride Parameter	Value (Mobilise-D consensus)
Duration	0.2 – 3 seconds
Minimum Length	0.15 meter

- The number of Walking Bouts identified within a Lab test (from the Laboratory assessment)/ Real world 2.5 hour assessment is based on the classification of “gait” as an activity that fulfils the WB definition.
- The Walking Bout Start will be identified as the start time of each walking bout
- The Walking Bout End will be identified as the end time of each walking bout
- The duration of Walking Bout (in seconds) will be calculated as the time difference between Walking Bout Start and End time

Step and Stride Duration

- Number of Initial Contact Events identified within a walking bout.
- Step duration, which refers to the duration (in seconds) calculated between two consecutive initial contacts. The average and SD of the step durations will be evaluated.
- Stride duration, which refers to the duration (in seconds) between two non-consecutive (alternate) initial contact events, so between two initial contact events of the same foot. The average and SD of the stride durations will be evaluated.

Information on laterality of initial contact needs to be validated as it will be utilised for evaluation of WB (where information about Right and Left strides is part of the definition) and further derived outcomes like asymmetry.

Although not an “outcome” per se, for each initial contact we will assess and validate its Laterality (IC identified as Left or Right) as follows:

- Laterality of each Initial Contact identified within a Walking Bout (either Left or Right). Here we will assess the number of qualified labels
- Sequencing Order, refers to the number of qualified sequenced labels (after Left: Right, after Right: Left).

Cadence

Cadence refers to the frequency of steps within a minute, it is evaluated in steps/ minute as double the stride frequency:

$$\boldsymbol{Cadence_i} = StrideFrequency_i * 2$$

where $i = 1, \ldots, n$ are the different WB and Stride Frequency is evaluated as:

$$\boldsymbol{Stride\ Frequency_i} = \frac{\sum_{k=1}^{n\_STRIDE_i} \left({}^{60}/_{STRIDE\ d_k} \right)}{n\_STRIDE_i}$$

where $i = 1, \ldots, n$ are the different WBs

n_STRIDE_i is the number of right <u>and</u> left strides in the relevant i –WB

$STRIDE\ d_k$ is the duration [seconds] of the k – stride in the relevant i –WB

Stride Length

- Average stride length within a walking bout will be evaluated as the average value of all stride lengths (stride length corresponds to the displacement, in meters, of each stride) within the walking bout as follows:

$$\boldsymbol{Stride\ Length_i} = \frac{\sum_{k=1}^{n\_STRIDE_i}(STRIDE\ l_k)}{n\_STRIDE_i}$$

where $i = 1, \dots, n$ are the different WBs, n_STRIDE_i is the number of right and left strides in the relevant i –WB,
$STRIDE\ l_k$ is the length [meters] of the k – stride in the relevant i –WB

In addition, also the SD of all stride lengths will be evaluated.

Turning

- Number of turns within a test (from lab assessments) or from the real world 2.5 hour assessment. Note that the turns will be provided within a Walking Bout level. Thus, only the turns identified in the detected WB will be assessed within a test/real world 2.5 hour assessment, potential turns detected outside WBs won't be analysed.
- Start of turn, which refers to the start events/instants of each turn identified within a Walking Bout.
- End of turn, which refers to the end events/instants of each turn identified within a Walking Bout.
- Duration of turn (in seconds), which refers to the duration of each turn identified within a Walking Bout.
- Angle of turn (degrees), which refers to the magnitude (i.e. angle) of each turn identified within a Walking Bout.

Other Secondary Outcomes (SO)

A range of spatiotemporal, signal-based and complex features will be derived from the signals included in a Walking Bout. Only secondary outcomes that are also evaluated by a RS (and so that can be compared) are included in this SAP.

- Number of Final Contact Events identified within a Walking Bout.
- Swing phase duration, which refers to the duration (time intervals) of the swing phase within a gait cycle, calculated as the time between the Final Contact of a foot and the following Initial Contact of the same foot. The average and SD of the swing phase durations will be evaluated.
- Stance phase duration, which refers to the duration (time intervals) of the stance phase, calculated as the time between Initial Contact and the Final Contact of the same foot. The average and SD of the swing phase durations will be evaluated.

2.3 Other Variables

Demographic and disease characteristics data such as sex, age, type of disease, use/non-use of walking aid and gait speed measured by the RS will be considered for the analysis as potential confounding factors or sub-group analyses, see section 3.3 and **Appendix B**. For the Real world (2.5 hours) assessment, contextual information derived from the mobile phone will be evaluated and matched for each detected Walking Bout, and will include the following information (Table 4). Please see more details in Deliverable 2.5, as part of WP2 tasks.

Table 4 contextual factors to be evaluated from the mobile phone for the Real world 2.5 hours assessment.

Contextual Information	Output	Unit	Level of Aggregation
Weather *	date	Date	Local Date
	temperature	Celsius	Average for day
	wind Speed	m/s	Average for day
	wind direction	degrees	Average for day
	precipitation	mm	Total for day
	snow	mm	Depth of snow on the day
Stay points	Stay point ID	categorical	Unique identifier
	Mean duration	seconds	The mean amount of time spent at the staypoint over all visits
	Total duration	seconds	The total amount of time spent at the staypoint
	Number of days	count	The number of days when this staypoint was visited.
	Types	categorical	Location class and type derived from Open Street Map (OSM)**
	Start and end times	Date (timestamp)	The start; end times for each visit
Indoor/ Outdoor	Date	Timestamp	Date/time
	Probability	0-1	Probability of being indoor (One value for each second. Where, 0 == outdoor, 1 == indoor)
Common path	Path ID	categorical	Unique identifier
	Frequency	count	The number of times path is travelled
	Frequency any directions	count	The number of times path is travelled in either direction
	Elevation	m	Change in elevation***
	Distance	m	Distance covered in the common path identified
	Start End Times	Date (timestamp)	Start and End time of the common path identified
Walking aid	Walking Aid ID	categorical	Walking aid category/type
	Start End Times	Date (timestamp)	Start and End time when a Walking aid was utilised

* WeatherBit (https://www.weatherbit.io) data aggregation service is used to extract weather information.

** Open Street Map (https://www.openstreetmap.org/#map=6/54.910/-3.432)

*** Worldwide elevations are provided by NASA's Shuttle Radar Topography Mission (SRTM), which has a granularity of 30m. Other, finer grained, elevation National specific data sources may be available, for example Ordnance Survey Terrain5 for the UK, which has a granularity of 5m.

3. Data Analysis

This section provides a detailed description of how criterion (concurrent) validity of the primary and secondary DMOs derived from the tested system will be compared to their respective reference system.

Only data collected during the Laboratory and Real world 2.5 hours assessments will be used for the purpose of this SAP (Table 1 and Figure 5), given the fact that during these assessments, at least a single reference system is available. DMOs derived from the SS / algorithm pair (in the Laboratory and Real world 2.5 hour assessments), and the same DMOs (when available) evaluated by the INDIP system (in the Laboratory assessment), will be compared to corresponding DMOs derived from their respective reference system. In particular, for the Laboratory assessment, the SP system will be used as RS for the following comparisons: SS vs. SP and INDIP vs. SP. For the Real world 2.5 hour assessments, the INDIP system will be used as a RS: SS vs. INDIP, please see Table 1 for more details.

The analyses presented in this SAP will be performed separately by disease cohort (HA, CHF, COPD, PD, MS, PFF). The results will be further stratified based on the biomechanical properties/categories of the participant's walking bout (WB): average stride speed, duration of the walking bout, walking speed (see section 3.2.1 for more details).

3. 1 Overview of planned analyses

Table 5 provides an overview of the planned analyses to characterise primary and secondary DMOs by quantifying criterion validity performance metrics, including concurrent validity metrics. Similar types of DMOs for which the same set of statistical analyses and performance metrics are evaluated, are identified by colour (blue, orange, green, grey and yellow, e.g. the DMOs which present the same structure and granularity), and therefore these require the same evaluation criteria.

In the table, for each of the DMOs the following metrics are summarized: the performance metrics that will be used to compare SS based DMOs vs. DMOs derived from the RS, the criterion validity metrics that will be used to compare SS based DMOs vs. DMOs derived from the RS, and the type of plots that will be used to visualise performance metrics and to support interpretation of the results.

We indicated entries in Table 5 with a (✓) when the analysis is performed for the comparison of focus, whereas we indicated with a (×) when the analysis cannot be performed for the comparison of focus. Table 5 also reports the target ranges (7-9) for the key performance metrics and criterion validity metrics (absolute and relative errors and Intra-Class Correlation ICC(2,1), see section 3.2.1 for more details). Target ranges will be used for description and interpretation of the validation results, and will help identifying SS DMOs performance metrics that have met all the desirable target values (and so being the closest to the RS DMOs, as an ideal situation) to the ones who may have met only a partial number of target values or technical requirements.

Table 5. List of Primary and Secondary DMO, corresponding performance/validity metrics with acceptable ranges and minimum aggregation level.

					PERFORMANCE METRICS					CRITERION VALIDITY METRICS						PLOT TYPES			MINIMUM AGGREGATION LEVELS			
Laboratory Assessment (SS vs. SP)	Laboratory Assessment (INDIP vs. SP)	Real world: 2.5 hours (SS vs. INDIP)	DMOs block	DMOs	Sensitivity (95% CI)	Positive Predictive Value (95% CI)	Accuracy (95% CI)	Specificity (95% CI)	F1-score (95% CI)	Absolute Errors (of mean & SD)	Relative Errors (of mean & SD)	Maximum and RMS of Absolute Errors; Maximum Relative Error *based on True Positives	Cohen's Kappa	Concurrent validity, ICC (95% CI)	Statistical Significance (p- value)	Bland Altman Plots (limits of agreement: ±2 SD)	Scatter Plots for Correlation (R-value and regression equation)	Histogram Plots	Steps/Stride level	WB level	Participant level (test/ assessment)	Cohort/subgroup level
				Acceptable values	>0.7	>0.7	>0.7	>0.7	>0.7		<20%		>0.7	>0.7	p>0.05							
Primary DMO																						
✓	✓	✓	Walking Speed	Walking Speed						🟩	🟩			🟩	🟩	🟩	🟩	🟩		✓	✓	✓
Secondary DMOs																						
✓	✓	✓	Walking Bout	Number (Identification) of Walking Bouts	🟦	🟦	🟦	🟦	🟦												✓	✓
				Start of Walking Bouts						🟧	🟧							🟧		✓	✓	✓
				End of Walking Bouts						🟧	🟧							🟧		✓	✓	✓
				Duration of Walking Bouts						🟩	🟩			🟩	🟩	🟩	🟩	🟩		✓	✓	✓
✓	✓	✓	Step/ Stride Duration	Initial Contact Events (Time)	▨	▨			▨	▨	▨	▨						▨	✓	✓	✓	✓
				Step and Stride Duration						🟩	🟩			🟩	🟩	🟩	🟩	🟩		✓	✓	✓
				Laterality						🟨	🟨		🟨					🟨		✓	✓	✓
				Sequence Order						🟨	🟨							🟨		✓	✓	✓
✓	✓	✓	Cadence	Cadence (steps/min)						🟩	🟩			🟩	🟩	🟩	🟩	🟩		✓	✓	✓
✓	✓	✓	Stride Length	Stride Length						🟩	🟩			🟩	🟩	🟩	🟩	🟩		✓	✓	✓
×	×	✓	Turning	Number (Identification) of Turns	🟦	🟦	🟦	🟦	🟦											✓	✓	✓
				Start of Turns						🟧	🟧							🟧		✓	✓	✓
				End of Turns						🟧	🟧							🟧		✓	✓	✓
				Duration of Turns						🟩	🟩			🟩	🟩	🟩	🟩	🟩		✓	✓	✓
				Maximal Angle of Turns						🟩	🟩			🟩	🟩	🟩	🟩	🟩		✓	✓	✓
✓	✓	✓	Secondary Outcomes	Final Contact Events (Time)	▨	▨			▨	▨	▨	▨						▨	✓	✓	✓	✓
				Stance Phase Duration						🟩	🟩			🟩	🟩	🟩	🟩	🟩		✓	✓	✓
				Swing Phase Duration						🟩	🟩			🟩	🟩	🟩	🟩	🟩		✓	✓	✓

Abbreviations: CI: Confidence Intervals, ICC(2,1): Intra-Class Correlation, SS: Single Sensor (Dynaport), INDIP: INertial module with DIstance Sensors and Pressure insoles, SD: Standard Deviation, SP: Stereophotogrammetric system.

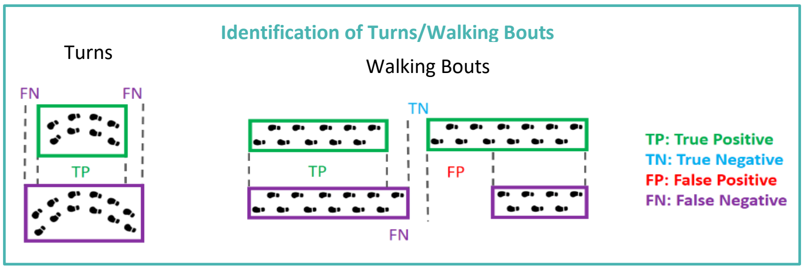
Figure 5 shows an example of an accelerometry signal recorded with a SS located on the lower back during a laboratory gait test. The schema presents an overview of the DMOs and the corresponding metrics/analyses proposed for each of them, with the same colour code as for Table 5. The left side of the figure illustrates a typical gait signal of a walking bout recorded with a single sensor. This diagram reflects the movement recorded with a triaxial accelerometer and describes how several DMOs are derived. The five panels represent the 5 different sets of the estimated DMOs from the single sensor unit (SS - green colour) and the corresponding DMOs from the reference system (RS - purple colour). These are indicated (colour used in the frame and in diagram title) with the same colours used in Table 5 and corresponding to the colours used for the DMOs marked in the signals diagram (figure on the left). The panels (blue, orange, green, grey and yellow) correspond to the five different sets of DMOs calculated, for which the same metrics/analyses are implemented, and the granularity is indicated.

Minimum granularity

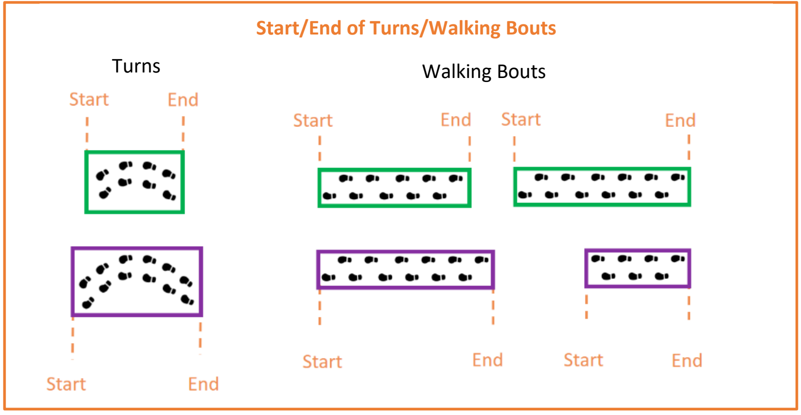
Single Sensor (SS)

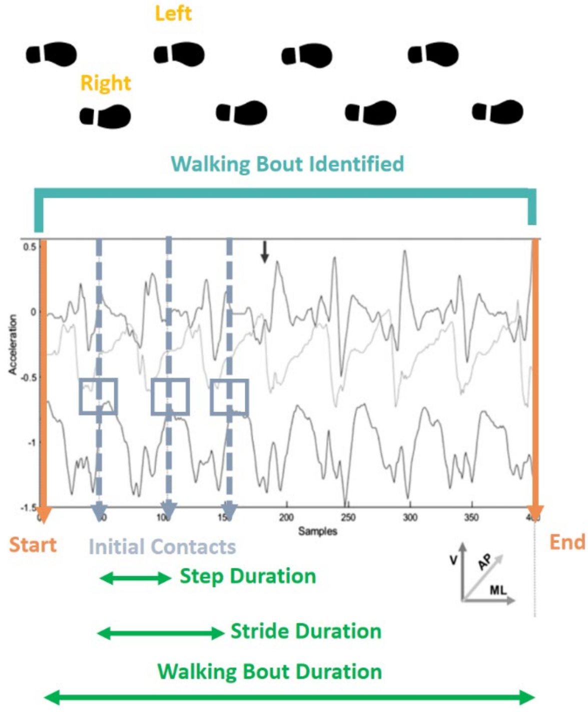
Reference System (RS)

Minimum granularity: **Test (lab assessment) or Real world 2.5 hour assessment**
Performance Metrics: **Sensitivity, Positive Predictive Value, Accuracy, Specificity, F1-score**
Plots: **Histograms of Errors**

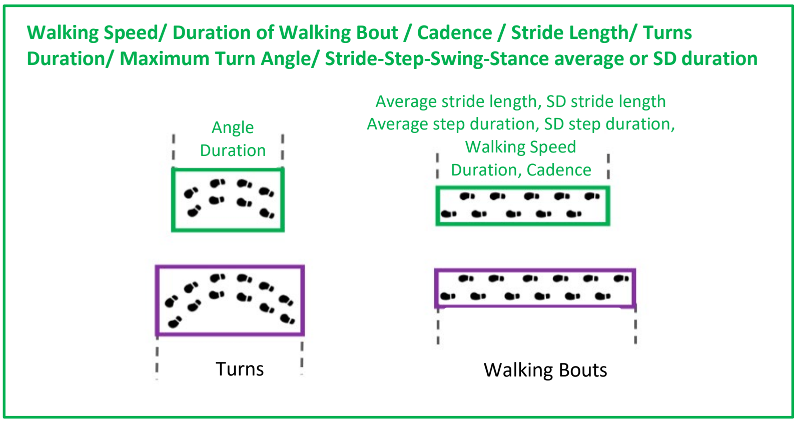


Minimum granularity: **Test (lab assessment) or Real world 2.5 hour assessment**
Criterion Validity: **Absolute and Relative Errors, Delta precision normalised by the mean**
Plots: **Histograms of Errors**

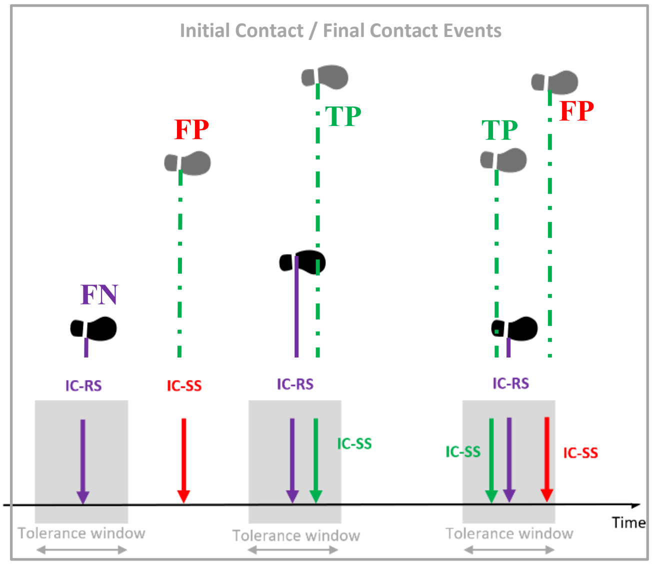




Minimum granularity: **Walking Bout**
Criterion Validity: **Absolute and Relative Errors, Delta precision normalized by the mean, Concurrent Validity with Confidence Interval, Significant Difference**
Plots: **Bland Altman plots, Correlation plots, Histograms of DMOs and Relative errors**


Walking Speed/ Duration of Walking Bout / Cadence / Stride Length/ Turns Duration/ Maximum Turn Angle/ Stride-Step-Swing-Stance average or SD duration
Angle
Duration
Average stride length, SD stride length
Average step duration, SD step duration,
Walking Speed
Duration, Cadence
Turns
Walking Bouts

Minimum granularity: **Initial Contact Event**
Performance Measures: **Sensitivity, Positive Predictive Value, F1-Score**
Criterion Validity: **Absolute and Relative Errors**
Plots: **Histograms of Relative Errors**



Minimum granularity: **Initial Contact Event**
Criterion Validity: **Absolute and Relative Errors, Cohen's Kappa**
Plots: **Histograms of Relative Errors**

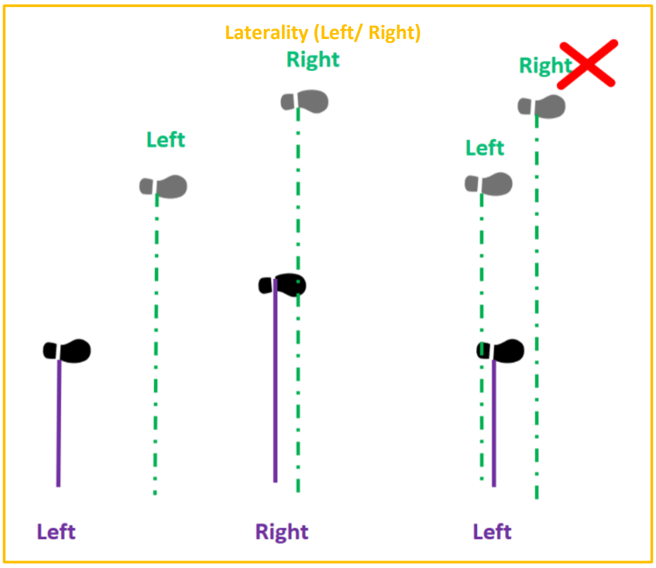


Figure 5. Overview of the planned metrics/analyses, organized by type of DMOs.

The template provided for the analysis of primary and secondary DMOs is included in **Appendix B**. In the template, each tab (sheet) corresponds to the metric/analyses proposed for each type of DMOs (e.g. Walking Bout, Step Duration, etc.).

3. 2. Statistical Analysis

3.2.1 Statistical analysis: definitions and glossary

All statistical analyses will be performed using the statistical analysis toolbox from Matlab R2021a.

In **section 3.2.2**, worked examples of evaluation of the following aggregation methods (i.e. the levels of data at which the analyses are performed) and statistical analyses for specific DMOs are presented, including examples of granularity of derived descriptive statistics for DMOs that present multiple outcomes per each WB and per participant.

<u>**Analyses based on all identified WB.**</u>

- **For the analyses based on** all detected WB (regardless of if they are identified as true positive, etc.)**, data will be analysed as follows:**

Continuous DMOs and **statistical analyses/metrics** will be reported using the following aggregation methods, depending on their minimum granularity (see Tables 3 and 5):

a. for the DMOs that present multiple values per WB (e.g. with a minimum granularity level set to a step/stride level, Tables 3 and 5), Figure 6-a:

   - <u>1st level of aggregation</u> – at *WB level*: the **mean and standard deviation (SD)** of the DMO, together to the **absolute and relative errors of the mean and the SD**. In the case of DMOs whose analyses are based on true positives (e.g. initial and final contact events), we will also obtain the ***maximum absolute and relative errors, and the root-mean-square of the absolute errors*** within each WB.

   - <u>2nd level of aggregation</u> – at *patient level* (across all WBs): **descriptive statistics (mean, SD, median, inter-quartile range, minimal, maximal and root-mean-square)** of each of the values obtained at the 1st level of aggregation, for each participant, across all WBs of each Lab test or Real world 2.5 hours assessment.

   - <u>3rd level of aggregation</u> – at *cohort level*: across all participants of the same cohort **(mean, 95% confidence interval (CI), SD, median, inter-quartile range, intra-class-correlation coefficient (when feasible) and p-value of normal distribution)** of all metrics obtained at the 2nd level of aggregation**.**

Note that for all DMOs the 95% CI will be evaluated only at cohort level (3rd level of aggregation). And in the case of ICC, we will perform the analysis at cohort level, working on DMO that are true positive events – so SS events that correspond to RS events (e.g. step durations based on true positives per WB, for all WBs, for all participants of the cohort), see more details on ICC in the Criterion validity section.

Annotated example (please note that for all annotated examples the indexes are different and independent from the one used in Table 3).

For each WB we have a set of values for the specific DMO (multiple values for each WB):

For participant 1 (PT1), we have identified:

$WB_1 = DMO_{1,1}, DMO_{1,2}, …, DMO_{1,m}$

$WB_2 = DMO_{2,1}, DMO_{2,2}, …, DMO_{2,\, p}$

…

$WB_{n\,=}\,DMO_{n,1}, DMO_{\,n,2}, …, DMO_{\,n,q}$

1) First descriptive statistics and performance metrics are evaluated within each WB – 1st level of aggregation for both RS and SS. See example of evaluation of the mean and applicable to the other descriptive statistics.

Mean DMO_{WB1} = mean ($DMO_{1,1}, DMO_{1,2}, …, DMO_{1,m}$) = $\frac{1}{m}\sum_{k=1}^{m}\mathrm{DMO}_{1,\mathrm{k}}$

Mean DMO_{WB2} = mean ($DMO_{2,1}, DMO_{2,2}, …, DMO_{2,p}$) = $\frac{1}{p}\sum_{k=1}^{p}\mathrm{DMO}_{2,\mathrm{k}}$

…

Mean DMO_{WBn} = mean ($DMO_{n,1}, DMO_{n,2}, …, DMO_{n,q}$) = $\frac{1}{q}\sum_{k=1}^{q}\mathrm{DMO}_{\mathrm{n,k}}$

2) Next, descriptive statistics are evaluated at participant level for all outcomes obtained at the 1st aggregation level, across all WBs – 2nd level of aggregation for both RS and SS (across all WBs belonging to a Lab test or Real world 2.5 hours assessment), for each participant (PT). See an example of evaluation of the mean (and weighted mean) which is applicable to the other descriptive statistics.

PT_1 Mean DMO = mean (Mean DMO_{WB1}, Mean DMO_{WB2}, …, Mean DMO_{WBn}) = $\frac{1}{n}\sum_{k=1}^{n}\mathrm{Mean}\ DMO_{WB_k}$

For another participant, participant 2 (PT_2), we will have, for example, “x” WBs: WB_1, WB_2, …, WB_x

PT_2 Mean DMO = mean (Mean DMO_{WB1}, Mean DMO_{WB2}, …, Mean DMO_{WBx}) = $\frac{1}{x}\sum_{k=1}^{x}\mathrm{Mean}\ DMO_{WB_k}$

Weighted Mean (weighted by the relative duration of the WB with respect to the total duration across all WBs):

PT_1Weighted Mean DMO = (Mean DMO_{WB1} * w_1, Mean DMO_{WB2} * w_2, …, Mean DMO_{WBn} * w_n) = = $\sum_{k=1}^{n}\mathrm{Mean\ DMO_{WBk}}$*$w_k$

w_k = duration of the walking bout WBk/ total duration across all n WBs (weights are normalized so that they sum up to a maximum of 1 for a given participant.

3) Then, general statistics of the outcomes obtained at 2nd level of aggregation are evaluated at cohort level, across all participants – 3rd level of aggregation for both RS and SS (belonging to a Lab test or Real world 2.5 hours assessment). Example of evaluation of the mean and applicable to other descriptive statistics.
For each DMO, the values will be averaged across all participants (N participants) belonging to a cohort (COPD, CHF, MS, PD, PFF). Corresponding 95% CI will be provided.

Annotated example:

If the HA cohorts consists of: PT_1, PT_2, …, PT_N

HA Mean DMO $_{2.5hours}$ = mean (PT_1 Mean DMO $_{2.5hours,}$ PT_2 Mean DMO $_{2.5hours}$, …, PT_N Mean DMO $_{2.5hours}$) = $= \frac{1}{N}\sum_{k=1}^{N} PT_k \; Mean \; DMO_{2.5hours}$

b. for the DMOs that present only one value per WB (e.g. WB duration, average stride length, cadence, etc.; i.e., those with a minimum granularity level set to a WB, Tables 3 and 5), the analyses will be performed as follows (Figure 6-b):

- 1st level of aggregation – at *patient level* (across all WBs): **descriptive statistics (mean, SD, median, inter-quartile range, minimal, maximal and root-mean-square)** of each of the values/DMOs/metrics obtained at the walking bout level, for each participant, across all WBs of each Lab test or Real world 2.5 hours assessment.
- 2nd level of aggregation – at *cohort level*: across all participants of the same cohort **(mean, 95% mean confidence interval, SD, median, inter-quartile range, intra-class-correlation coefficient (when feasible), and p-value of normal distribution)** of all metrics obtained at the 1st level of aggregation**.**

Note that the 95% CI will be evaluated only at cohort level (2nd level of aggregation). In the case of ICC, we will perform the analysis at cohort level, working on DMOs evaluated on WB that are true positive events only (so WB identified by the SS that correspond to WB identified by RS), see more details on ICC in the Criterion validity section.

Annotated example:

For each WB we have one DMO (one value for each WB).

For participant 1 (PT_1), we have identified:

WB_1 = DMO_1

WB_2 = DMO_2

…

$WB_{n\,=}$ DMO_n

For participant 2 (PT_2) we have identified:

WB_1 = DMO_1

WB_2 = DMO_2

…

$WB_{x\,=}$ DMO_x

1) First, descriptive statistics and DMOs are evaluated a at participant level, across all WBs – 1st level of aggregation (belonging to a Lab test or Real world 2.5 hours assessment) for both RS and SS, for each participant (PT). An example of the evaluation of the mean (and weighted mean), applicable to the other descriptive statistics, is presented as follows:

PT_1 Mean DMO = mean (DMO_1, DMO_2, …, DMO_n) = $\frac{1}{n}\sum_{k=1}^{n} DMO_{WB_k}$

For another participant (PT_2), we will have, for example, x WBs: WB_1, WB_2, …, WB_x

PT_2 Mean DMO = mean (DMO_{WB1}, DMO_{WB2}, …, DMO_{WBx}) = $\frac{1}{x}\sum_{k=1}^{x} DMO_{WB_k}$

Weighted Mean (weighted by the relative duration of the WB with respect to the total duration across all WBs):

PT_1 Weighted Mean DMO = (Mean DMO_{WB1} * w_1, Mean DMO_{WB2} * w_2, …, Mean DMO_{WBn} * w_n) = = $\frac{1}{n}\sum_{k=1}^{n}$ Mean DMO_{WBk}*w_k

w_k = duration of the walking bout WBk/ total duration across all n WBs (weights are normalized such that they sum up to 1 for a given participant).

2) Then, general statistics are evaluated at cohort level, across all participants – 2nd level of aggregation (across all WBs belonging to a Lab test or Real world 2.5 hours assessment) for both RS and SS. See example of evaluation of the mean, applicable to the other descriptive statistics. For each DMO, the values will be averaged across all participants (N participants) belonging to a cohort (COPD, CHF, MS, PD, PFF). Corresponding 95% CI will be provided.

Annotated example:
If the HA cohorts consists of: PT_1, PT_2, …, PT_N

HA Mean DMO = mean (PT_1 Mean DMO, PT_2 Mean DMO, …, PT_N Mean DMO) = = $\frac{1}{N}\sum_{k=1}^{N} PT_k\ Mean\ DMO$

Categorical data analysis, e.g. laterality labels of the foot performing the initial contact events: left or right, will be reported using the absolute number of correctly identified labels, and the relative number (i.e. normalized to the total number of initial contact events). Differently to numerical analysis, instead of intra-class-correlation coefficient (ICC) we will use the Cohen's kappa at the highest level of aggregation (cohort level), by including all initial contact events labels, based on the true positives.

Sub-group Analysis:
For both continuous DMOs and categorical data, despite obtaining the results per cohort, the results will also be obtained and presented in "subgroups" of a specific cohort, based on the gait speed of the participants as quantified by the RS. Thus, we will stratify and average all the results by the stride gait speed (fast speed defined as >1 m/s, medium speed: 0.5 to 1 m/sec and slow speed <0.5 m/sec gait speed (10), more details in section 3.3), Figure 6.

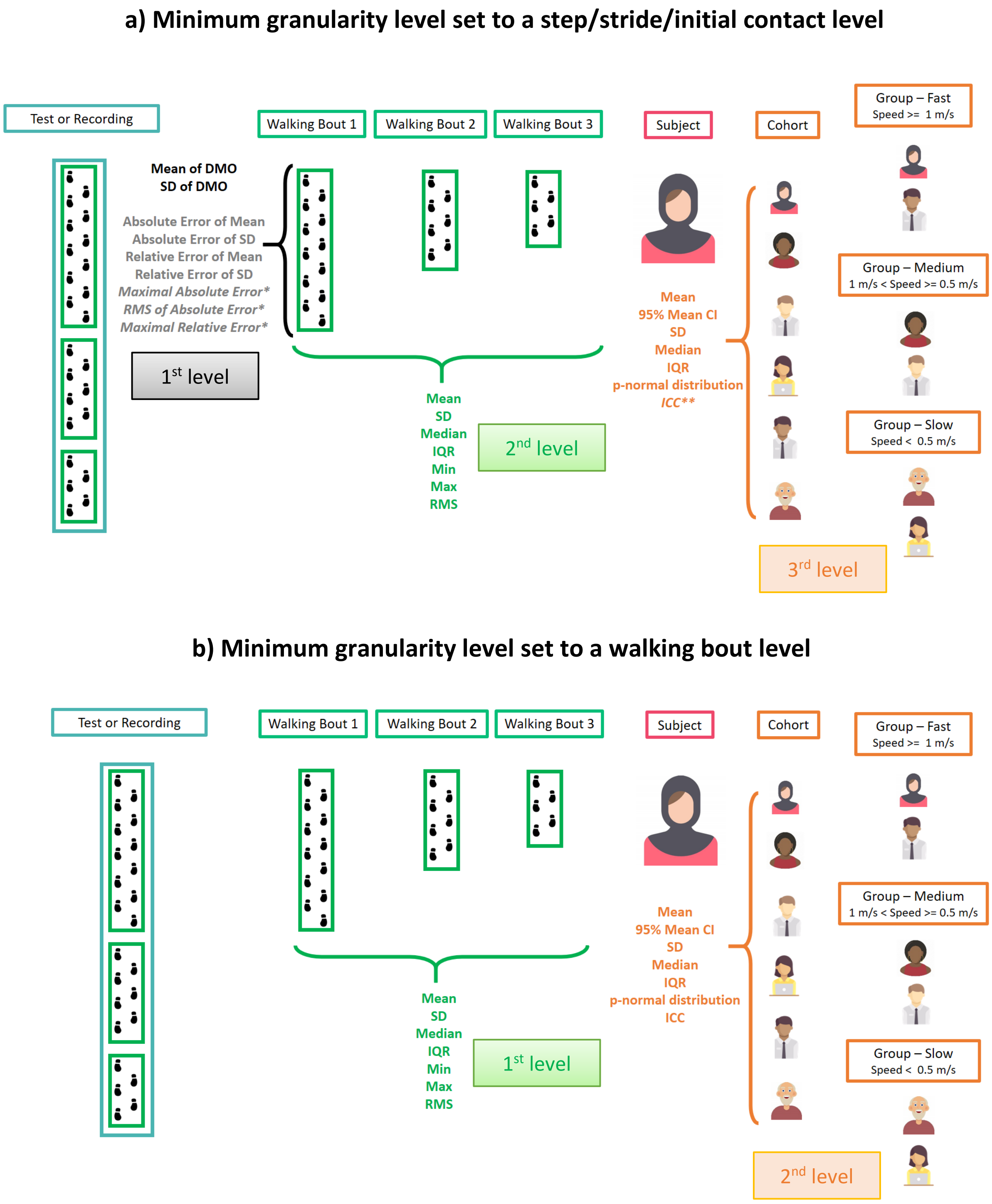


Figure 6. Aggregation method for: a) DMOs that present multiple outcomes per each WB; b) DMOs that presents only one single outcome per each WB. Note that in section a) we marked with "*" to refer to the metrics that only apply to the DMOs whose analyses are based on true positives (e.g. initial and final contact events). In section b) we marked with "**" to refer to the ICC that only will be calculated when feasible (e.g. excluding initial and final contact events). CI: Confidence Interval; SD: Standard Deviation; IQR: inter-quartile range; RMS: root-mean square; ICC: intra-class-correlation coefficient.

The definitions for each of the metrics presented in Table 5 and Figure 5 are described below:

Performance metrics

We first categorize event variables identified/not identified by SS relative to RS:

- **True Positive (TP)**: event correctly identified by the SS; the event is identified by the SS and by the RS (Figure 10).
- **True Negative (TN)**: event correctly not identified by the SS; the event is not identified by the SS and neither by the RS (Figure 10).

*Note that TN will not be estimated when identifying initial and final contact events.

- **False Positive (FP)**: event incorrectly identified by the SS; the event is identified by the SS, but not by the RS (Figure 10).
- **False Negative (FN)**: event incorrectly not identified by the SS; the event is not identified by the SS (missing event), but it is identified by the RS (Figure 10).

Next, using above definitions for TP, TN, FP, and FN, we derive the performance metrics.

- **Sensitivity:** calculated as:

$$Sensitivity = \frac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FN}}$$

- **Positive Predictive Value:** calculated as:

$$Positive\ Predictive\ Value = \frac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FP}}$$

- **Accuracy:** calculated as:

$$Accuracy = \frac{\mathbf{TP} + \mathbf{TN}}{\mathbf{TP} + \mathbf{TN} + \mathbf{FP} + \mathbf{FN}}$$

- **Specificity:** calculated as:

$$Specificity = \frac{\mathbf{TN}}{\mathbf{TN} + \mathbf{FP}}$$

*Note that accuracy and specificity can only be evaluated only when True negative (TN) events can be identified, which is the case of the identification of walking bouts and turns, but not for the identification of initial and final contact events.

- **F1-score:** the combination of positive predictive value and sensitivity measures, calculated as:

$$\boldsymbol{F1-score} = 2 * \frac{Positive\ Predictive\ Value * Sensitivity}{Positive\ Predictive\ Value + Sensitivity}$$

Granularity of derived descriptive statistics: The above performance metrics, related to the identification of turns/walking bouts, will be derived for each Lab test and Real world 2.5 hours assessment, for each participant.

In the case of initial and final contact events, since their analyses are based on true positives, the **descriptive statistics** will **apply for each aggregation level**:

- 1st level of aggregation – at ***WB level***: including the maximum of all absolute and relative errors and the root-mean-square of all absolute errors (i.e. for each event).
- 2nd level of aggregation – at ***patient level*** (across all WBs): descriptive statistics are evaluated for each participant, across all WBs of each Lab test or Real world 2.5 hours assessment, from all the outcomes obtained at the 1st level of aggregation.
- 3rd level of aggregation – at ***cohort level***: across all participants of the same cohort, for all outcomes derived from the 2nd level of aggregation. Note that for the initial and final contact events, the ICC will not be performed.

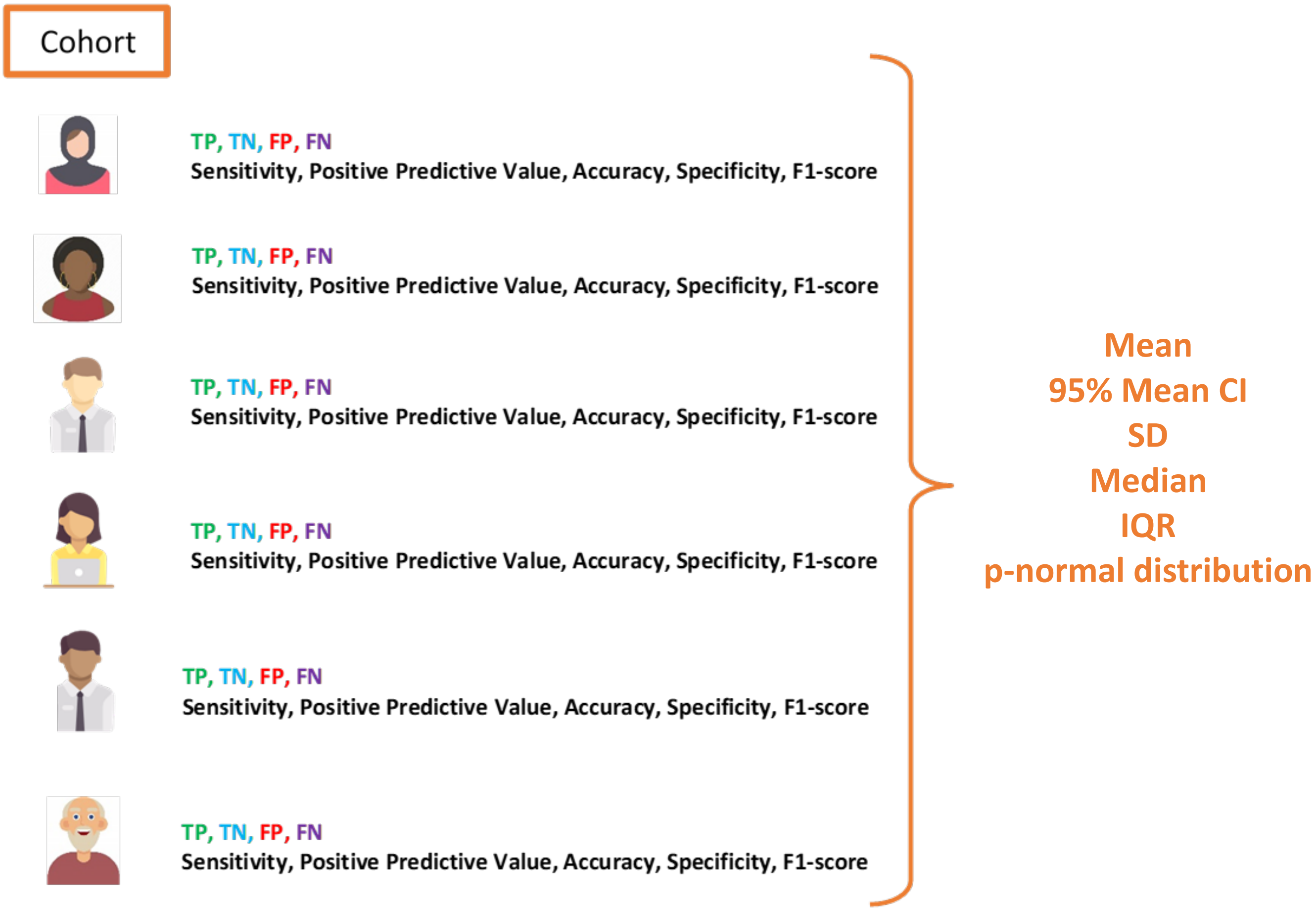

Cohort
TP, TN, FP, FN
Sensitivity, Positive Predictive Value, Accuracy, Specificity, F1-score
TP, TN, FP, FN
Sensitivity, Positive Predictive Value, Accuracy, Specificity, F1-score
TP, TN, FP, FN
Sensitivity, Positive Predictive Value, Accuracy, Specificity, F1-score
TP, TN, FP, FN
Sensitivity, Positive Predictive Value, Accuracy, Specificity, F1-score
TP, TN, FP, FN
Sensitivity, Positive Predictive Value, Accuracy, Specificity, F1-score
TP, TN, FP, FN
Sensitivity, Positive Predictive Value, Accuracy, Specificity, F1-score
Mean
95% Mean CI
SD
Median
IQR
p-normal distribution

Figure 7. The figure represents the approach proposed for the calculation of the performance metrics of a WB identification at cohort level (3rd level of aggregation). The mean, 95% confidence interval, standard deviation (or median and inter-quartile range, IQR), values across participants are derived at cohort level.

Criterion Validity

- **Absolute error:** is calculated as the absolute difference between the DMOs estimated by the RS and the DMOs estimated by the SS.

$$\boldsymbol{Absolute\ Error\ (DMO)} = |DMO\ estimated\ by\ SS - DMO\ estimated\ by\ RS|$$

Absolute errors of derived DMOs descriptive statistics, e.g. mean, standard deviation (SD), will be evaluated as follows:

$$\boldsymbol{Absolute\ Error\ (Mean\ of\ DMO)} = |Mean\ (DMO\ SS) - Mean\ (DMO\ RS)|$$

$$\boldsymbol{Absolute\ Error\ \ (SD\ of\ DMO)} = |SD\ (DMO\ SS) - SD\ (DMO\ RS)|$$

- **Relative error:** is calculated as the percentage of the ratio between the absolute difference of the DMOs estimated by the SS and the DMOs estimated by the RS, and the DMOs estimated by the RS.

$$\boldsymbol{Relative\ error} = \left(\left|\frac{DMO\ estimated\ by\ SS - DMO\ estimated\ by\ RS}{DMO\ estimated\ by\ RS}\right|\right) \times 100$$

Relative errors of derived DMOs descriptive statistics, e.g. mean, standard deviation (SD), will be evaluated as follows:

$$\boldsymbol{Relative\ Error\ (Mean\ of\ DMO)}\ = \frac{|Mean\ (DMO\ SS) - Mean\ (DMO\ RS)|}{Mean\ (DMO\ RS)}$$

$$\boldsymbol{Relative\ Error\ (SD\ of\ DMO)}\ = \frac{|SD\ (DMO\ SS) - SD\ (DMO\ RS)|}{Mean\ (DMO\ RS)}$$

Note that the Relative Error of the SD is also referred to as the precision normalized by the mean.

<u>**Granularity of derived descriptive statistics**</u>**:** The above statistical metrics will be evaluated at **each aggregation level**:

- 1st level of aggregation – at WB level: descriptive statistics are evaluated **within each WB** for all the DMOs that present multiple values per WB (e.g. DMOs with a minimum granularity level set to a step/stride event level) (11, 12), **statistical metrics at this level will be evaluated considering only the WBs that are identified as True Positives** (i.e. a WB detected by the RS that is also correctly identified by the SS, to have a correspondence of values for both systems).
- 2nd level of aggregation – at patient level (across all WBs): descriptive statistics are evaluated **for each participant, across all WBs** of each Lab test or Real world 2.5 hours assessment, **statistical metrics at this level will be evaluated considering all WBs** detected by the RS (including FN) and all WB detected by the SS (TP, FP).
- 3rd level of aggregation – at cohort level: across all participants of the same cohort (Figure 7).

Note that the 95% CI will be evaluated only at cohort level (3rd level of aggregation).

- ❖ **Concurrent Validity:** The Intra-class Correlation Coefficient, ICC(2,1) (13) will be calculated for continuous DMOs (e.g. WB duration, step duration, cadence) **only considering WBs that are True Positives** (i.e. a WB detected by the RS that is also correctly identified by the SS, to have a correspondence of values for both systems). Thus, the relationship between the DMOs from the tested system, and the corresponding DMOs from the reference system will be evaluated. The correlation coefficient (r), corresponding p value and 95% confidence intervals will be reported.

❖ $$\text{ICC }(2,1) = \frac{BMS - EMS}{BMS+(k-1)EMS+k(JMS-EMS)/n}$$

BMS: between observations mean square
EMS: residual mean square
JMS: between systems mean square
k: number of tested devices (k=2). n=number of observations

Example for 1 DMO, 2 participants with 3 WBs each (k=2 devices, n=6 observations).

ID	PT	WB	DMO$_{RS}$	DMO$_{ss}$
PT1$_{WB1}$	1	1	1.00	1.00
PT1 $_{WB2}$	1	2	0.99	0.97
PT1 $_{WB3}$	1	3	0.97	0.95
PT2 $_{WB1}$	2	1	2.30	2.00
PT2 $_{WB2}$	2	2	2.40	2.10
PT2 $_{WB3}$	2	3	2.90	2.30

Analysis of variance:
BMS= 1.1439
EMS=0.02845
JMS=0.1281

ICC(2,1)= (BMS-EMS)/(BMS+(k-1)EMS+k(JMS-EMS)/n)= 0.925

In the case of categorical DMOs, as the laterality labels assigned to each initial contact event (left or right), a Cohen's kappa analysis will be performed. The Cohen's kappa index will be calculated using the following definition:

$$\kappa = \frac{2 \times (TP \times TN - FN \times FP)}{(TP + FP) \times (FP + TN) + (TP + FN) \times (FN + TN)}$$

TP, TN, FN, FP as previously defined.

<u>Granularity of derived descriptive statistics</u>: In the case of DMOs that present multiple values within a WB (e.g. step duration), the ICC(2,1) will be evaluated in two different ways (see annotated examples):

- 1st level of aggregation – at cohort level (across all WBs): descriptive statistics are evaluated considering all single DMOs values (e.g. all step durations) **accumulated over all WBs** within a Lab test/ real world (2.5 hour) assessment.

- 2nd level of aggregation – at cohort level: across all participants of the same cohort: considering the mean DMOs (e.g. mean step duration) **for each WB**, leading to several mean DMOs (as many as WBs available) for a Lab test/ Real world (2.5 hour) assessment. The ICC(2,1) between SS and RS will be evaluated considering all mean DMOs outputs (one for each WB), of all participants belonging to a certain cohort/group, as follows:

1st level: **Annotated example**:

For each WB we have a set of DMOs (multiple values for each WB):

For participant 1 (PT1), for the 2.5 hour assessment, we have identified, for the SS:

WB_{1SS} = DMO_{1SS}, DMO_{2SS}, …, DMO_{mSS} = $DMO_{\mathbf{WB1}1SS,}$ $DMO_{\mathbf{WB1}2SS}$, …, $DMO_{\mathbf{WB1}mSS}$

WB_{2SS} = DMO_{1SS}, DMO_{2SS}, …, DMO_{pSS} = $DMO_{\mathbf{WB2}1SS,}$ $DMO_{\mathbf{WB2}2SSs}$, …, $DMO_{\mathbf{WB2}pSS}$

…

$WB_{nSS =}$ DMO_{1SS}, DMO_{2SS}, …, DMO_{qSS} = $DMO_{\mathbf{WBn}1SS,}$ $DMO_{\mathbf{WBn}2SS}$, …, $DMO_{\mathbf{WBn}qSS}$

And we will have values for the RS:

For participant 1 (PT1), for the 2.5 hour assessment, we have identified, for the RS:

WB_{1RS} = DMO_{1RS}, DMO_{2RS}, …, DMO_{mRS} = $DMO_{\mathbf{WB1}1RS,}$ $DMO_{\mathbf{WB1}2RS}$, …, $DMO_{\mathbf{WB1}mRS}$

WB_{2RS} = DMO_{1RS}, DMO_{2RS}, …, DMO_{pRS} = $DMO_{\mathbf{WB2}1RS,}$ $DMO_{\mathbf{WB2}2RS}$, …, $DMO_{\mathbf{WB2}pRS}$

…

$WB_{nRS =}$ DMO_{1RS}, DMO_{2RS}, …, DMO_{qRS} = $DMO_{\mathbf{WBn}1RS,}$ $DMO_{\mathbf{WBn}2RS}$, …, $DMO_{\mathbf{WBn}qRS}$

1) First ICC(2,1) is evaluated at cohort level, across all WBs using only TP DMO events within each WB – 1st level of aggregation (belonging to a Lab test or Real world 2.5 hours assessment), for each participant (PT) of a cohort. For example, we will have data for each participants of a specific cohort (e.g. HA). PT2 will have a similar structure for “x” WB, and so on, with PT_{20} having “z” WBs.

If the HA cohorts consists of: PT_1, PT_2, …, PT_{20}

ICC(2,1) (DMO) = ICC(2,1) [($PT_1DMO_{\mathbf{WB1}1SS}$,$PT_1DMO_{\mathbf{WB1}1RS}$), ($PT_1DMO_{\mathbf{WB1}2SS}$,$PT_1DMO_{\mathbf{WB1}2RS}$),…, (PT1$DMO_{\mathbf{WB2}1RS}$,$PT_1DMO_{\mathbf{WB2}1SS}$),…, (PT1$DMO_{\mathbf{WBn}1SS}$,$PT_1DMO_{\mathbf{WBn}1RS}$),…, (PT1$DMO_{\mathbf{WBn}qSS}$,$PT_1DMO_{\mathbf{WBn}qRS}$), ($PT_2DMO_{\mathbf{WB1}1SS}$,$PT_2DMO_{\mathbf{WB1}1RS}$), ($PT_2DMO_{\mathbf{WB1}2SS}$,$PT_2DMO_{\mathbf{WB1}2RS}$),…, ($PT_2DMO_{\mathbf{WB1m}SS}$,$PT_2DMO_{\mathbf{WB1}mRS}$),…, ($PT_2DMO_{\mathbf{WB2}1SS}$,$PT_2DMO_{\mathbf{WB2}1RS}$),…, ($PT_2DMO_{\mathbf{WBx}1SS}$,$PT_2DMO_{\mathbf{WBx}1RS}$),…, ($PT_{20}DMO_{\mathbf{WB1}1SS}$, $PT_{20}DMO_{\mathbf{WB1}1RS}$), …,($PT_{20}DMO_{\mathbf{WB2}1SS}$,$PT_{20}DMO_{\mathbf{WB2}1RS}$),…, ($PT_{20}DMO_{\mathbf{WBz}qSS}$,$PT_{20}DMO_{\mathbf{WBz}qRS}$)]

2^{nd} level: Annotated example:

For each WB we have a set of DMOs (multiple values for each WB):

For participant 1 (PT1), for the 2.5 hour assessment, for the SS, we will evaluate the mean of the DMOs within the WB:

Mean $\text{DMO}_{\text{WB1SS}}$ = Mean(DMO_{1ss}, DMO_{2ss}, …, DMO_{mss})

Mean $\text{DMO}_{\text{WB2SS}}$ = Mean(DMO_{1ss}, DMO_{2ss}, …, DMO_{pss})

…

Mean $\text{DMO}_{\text{WBnSS =}}$ Mean(DMO_{1ss}, DMO_{2ss}, …, DMO_{qss})

And we will have values for the RS:

For participant 1 (PT1), for the 2.5 hour assessment, we have identified, for the RS:

Mean $\text{DMO}_{\text{WB1RS}}$ = Mean(DMO_{1RS}, DMO_{2RS}, …, DMO_{mRS})

Mean $\text{DMO}_{\text{WB2RS}}$ = Mean(DMO_{1RS}, DMO_{2RS}, …, DMO_{pRS})

…

Mean $\text{DMO}_{\text{WBnRS =}}$ Mean(DMO_{1RS}, DMO_{2RS}, …, DMO_{qRS})

2) Then, the ICC(2,1) is evaluated at cohort level, across all participants – 2^{nd} level of aggregation (across all WBs belonging to a Lab test or Real world 2.5 hours assessment). For each DMO, the values will be evaluated across all participants belonging to a cohort (COPD, CHF, MS, PD, PFF).

We will have data for each participants of a specific cohort (e.g. HA).
PT2 will have a similar structure for "x" WB, and so on, with PT_{20} having "z" WBs.

If the HA cohorts consists of: PT_1, PT_2, …, PT_{20}

ICC(2,1) (DMO) = ICC(2,1) [($\text{PT}_1\text{MeanDMO}_{\text{WB1SS}}$, $\text{PT}_1\text{MeanDMO}_{\text{WB1RS}}$), ($\text{PT}_1\text{MeanDMO}_{\text{WB2SS}}$,$\text{PT}_1\text{MeanDMO}_{\text{WB2RS}}$),…, ($\text{PT}_1\text{MeanDMO}_{\mathbf{WBn}\text{SS}}$, $\text{PT}_1\text{MeanDMO}_{\mathbf{WBn}\text{RS}}$), ($\text{PT}_2\text{MeanDMO}_{\text{WB1SS}}$, $\text{PT}_2\text{MeanDMO}_{\text{WB1RS}}$),…, ($\text{PT}_2$Mean $\text{DMO}_{\text{WB2SS}}$, $\text{PT}_2\text{MeanDMO}_{\text{WB2RS}}$),…, ($\text{PT}_2\text{MeanDMO}_{\mathbf{WBx}\text{SS}}$, $\text{PT}_2\text{MeanDMO}_{\mathbf{WBx}\text{RS}}$), …, ($\text{PT}_{20}\text{MeanDMO}_{\text{WB1SS}}$, PT_{20}MeanDMO $_{\text{WB1RS}}$), …, ($\text{PT}_{20}\text{MeanDMO}_{\text{WBzRS}}$, $\text{PT}_{20}\text{MeanDMO}_{\text{WBzSS}}$)]

Note that in the case of categorical DMOs, as for the laterality labels assigned to initial contact events, only method "a" will apply when performing the Cohen's kappa analysis. In this case, all labels assigned to each initial contact event (of all WBs of a test/ real world (2.5 hour) assessment, of a specific participant) will be accumulated (for SS and RS). Then, the Cohen's kappa index will be evaluated considering all outputs (all labels of all WBs) of all belonging to a certain cohort/group following the formula provided in the concurrent validity section.

- ❖ **Statistically significant difference:** To evaluate statistically significant differences between the DMOs measured using the SS and RS, parametric (paired t-test) and non-parametric (Wilcoxon signed-rank test) tests will be performed and will consider the distribution of the data for continuous DMOs. This will be done **only considering WBs that are True Positives** (i.e. a WB detected by the RS that is also correctly identified by the SS, to have a correspondence of values for both systems). Data distribution will be visually inspected with histograms, and tested with the Shapiro Wilk test.

  <u>Granularity of derived descriptive statistics</u>: Statistically significant differences will be evaluated on DMOs both at 1st and 2nd level of aggregation as follows.

1) **First, statistically significant differences are evaluated at cohort level, across all WBs – 1st level of aggregation** (belonging to a Lab test or Real world 2.5 hours assessment), for each participant (PT) of a cohort.

Student paired *t*-test will test the hypothesis that mean DMOs μ_{rs} measured by RS are not different than mean DMOs μ_{ss} measured by SS. The Student statistic T will be calculated as follows:

$$T= \bar{d}\sqrt{n} / Sd$$

$\bar{d}$, the mean difference between the DMO measured from RS and from SS at 1st level, as follows:
$\bar{d}$ = mean (PT_iDMO_{jkRS}-$PT_iDMO_{jk}SS$), for all participant *i,* WBs *j* and steps *k*.
n= total number of observations (e.g. sum all evaluated steps for all WB and all participants)
Sd= Standard deviation of $\bar{d}$

Similar approach will be followed to derive Wilcoxon signed-rank test according to definition (14).

2) Then, statistically significant differences are evaluated at cohort level, across all participants – 2nd level of aggregation (across all WBs belonging to a Lab test or Real world 2.5 hours assessment). For each DMO, the values will be evaluated across all participants belonging to a cohort (COPD, CHF, MS, PD, PFF).

We will derive student T as:

$$T= \bar{d}\sqrt{n} / Sd$$

$\bar{d}$ is the mean difference between the mean DMO from RS and from SS at WB level, as follows:
$\bar{d}$ = mean (PT_i $meanDMO_{WBjRS}$-$PT_imeanDMO_{WBjSS}$)), participant *i*, WB *j.*
n= total number of observations (e.g. sum all WBs for all participants)
Sd= Standard deviation of $\bar{d}$

Similar approach will be followed to derive Wilcoxon signed-rank test according to definition (14).

<u>Analyses based on True Positive WB.</u>

- **For the analyses based on** WB that are true positive events only (so WB identified by the SS that correspond to WB identified by RS), **the above performance metrics and data analyses will be carried out across all Walking Bouts of all participants of a specific cohort:**

1st level of aggregation – at cohort level (across all WBs): descriptive statistics are evaluated considering all single DMOs values (e.g. all step durations) **accumulated over all WBs** within a Lab test/ real world (2.5 hour) assessment

Annotated example for evaluation of mean DMO across all True positive WB identified in a cohort (HA) (please note that for all annotated examples the indexes are different and independent from the one used in **Table 2**). Example and annotations are transferrable to other performance metrics (e.g. evaluation of relative error, etc.)

For each WB we have a set of values for the specific DMO (multiple values for each WB):

For participant 1 (PT1), we have identified:

WB_1 = $DMO_{1,1}$, $DMO_{1,2}$, …, $DMO_{1,m}$

WB_2 = $DMO_{2,1}$, $DMO_{2,2}$, …, $DMO_{2,\ p}$

… $WB_{n\ =}$ $DMO_{n,1}$, $DMO_{\ n,2}$, …, $DMO_{\ n,q}$

Mean $DMO_{PT1,WB1}$ = Mean ($DMO_{1,1}$, $DMO_{1,2}$, …, $DMO_{1,m}$)

Mean $DMO_{PT1,WB2}$ = Mean ($DMO_{2,1}$, $DMO_{2,2}$, …, $DMO_{2,p}$)

…

Mean $DMO_{PT1,WBn}$ = Mean ($DMO_{n,1}$, $DMO_{\ n,2}$, …, $DMO_{\ n,q)}$

With a total number of true positive Walking bouts WBi= n

If the HA cohorts consists of: PT_1, PT_2, …, PT_{20} , let's assume we have identified a total of n WB with n = $\sum_{i=1}^{20} WBi$,

HA Mean DMO = mean (Mean $DMO_{PT1,WB1}$, Mean $DMO_{PT1,WB2}$, …, Mean $DMO_{\ PT1,WBn}$, …., Mean $DMO_{PT20,WBn}$) = $\frac{1}{n}\sum_{k=1}^{n} \mathrm{Mean}\ DMO_{WB_k}$,

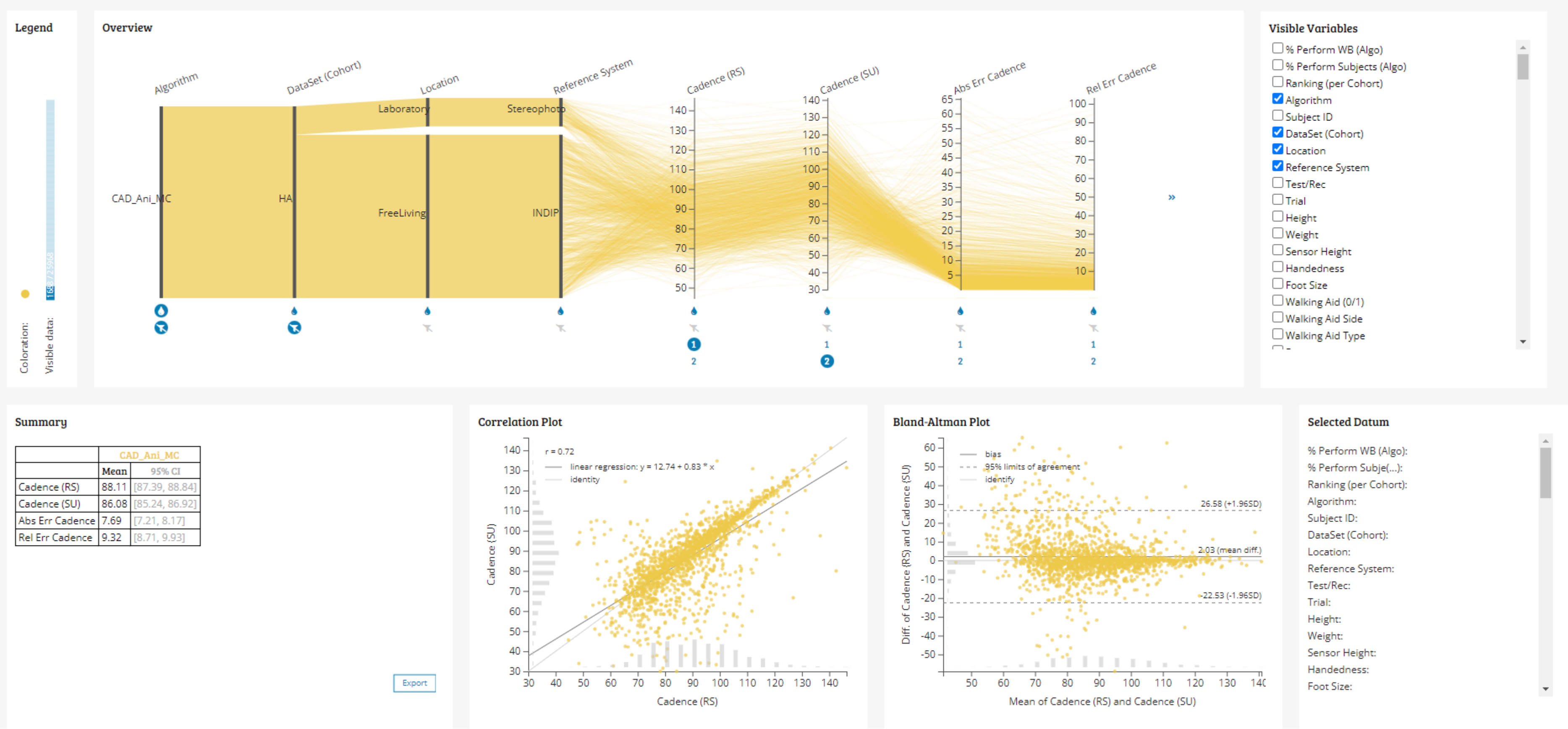
With n = total number of True positive WB identified across the "i" participants (e.g. i =20) of the HA cohort.

Plot types

An interactive visualization toolbox developed by Dr Alma Cantu (as part of collaboration with WP3) will be used for data exploration, to visualise different granularities and aggregation levels, and plot generation. The toolbox uses as an input the DMOs results dataset (**Appendix C** shows an example for the DMO cadence loaded on the visualisation toolbox platform. This toolbox enables to access, explore, visualise all available information (or variables - e.g. technical, statistical, demographic data), and to interact with the data by selecting available variables. Each datum in the dataset corresponds to a single WB, for a specific participant, allowing to visualise data on a WB level. Moreover, summary statistics (mean, min, max, SD, median, RMS, IQR) over all WBs of a participant can be presented, addressing in this way the different aggregation levels of analysis, as described in this SAP.

The toolbox (Figure 8) presents on the right side the complete list of available variables ("Visible Variables"). A specific selected subset of this information (i.e. a list of preferred variables) can be selected, which will be displayed in the top of the toolbox ("Overview"), with one axes per variable. The toolbox allow variables (e.g. Cadence Mean RS, Relative Error Cadence Mean) to be filtered or selected. This specific filtered or selected data will be used for: a) plots (as described in the next paragraph), b) evaluate statistical summaries of the selected data (mean, 95% confidence interval, SD, min, max, median, RMS, IQR, normal distribution) ("Information", bottom left of Figure 8).

For data plotting, the interactive selection of two systems (RS and SS) for the DMO of interest will be used for the Scatter and Bland Altman plots. Moreover, the histograms of each of the DMOs selected for the plots will be represented in a light grey colour, in the same direction of the plot axes. Finally, any of the point presented in these plots can be identified and selected, and a small report will be presented on the right side of the plots ("Selected Datum"), containing all available information for this particular data point. In this way, the outliers presented in these plots can be identified, as well as all their complete clinical, demographic, technical and statistical information, which potentially might help identifying the reasons of its "outlier" nature; and thus, potentially identifying sources of errors.



Summary

	CAD_Ani_MC	
	Mean	95% CI
Cadence (RS)	88.11	[87.39, 88.84]
Cadence (SU)	86.08	[85.24, 86.92]
Abs Err Cadence	7.69	[7.21, 8.17]
Rel Err Cadence	9.32	[8.71, 9.93]

Visible Variables

- □ % Perform WB (Algo)
- □ % Perform Subjects (Algo)
- □ Ranking (per Cohort)
- ☑ Algorithm
- □ Subject ID
- ☑ DataSet (Cohort)
- ☑ Location
- ☑ Reference System
- □ Test/Rec
- □ Trial
- □ Height
- □ Weight
- □ Sensor Height
- □ Handedness
- □ Foot Size
- □ Walking Aid (0/1)
- □ Walking Aid Side
- □ Walking Aid Type

Selected Datum

% Perform WB (Algo):
% Perform Subje(...):
Ranking (per Cohort):
Algorithm:
Subject ID:
DataSet (Cohort):
Location:
Reference System:
Test/Rec:
Trial:
Height:
Weight:
Sensor Height:
Handedness:
Foot Size:

Figure 8. Screenshot of the interactive visualisation toolbox using cadence results. Example of mean cadence results evaluated across each identified WB for a full cohort (HA, n=20).

The following plots will be used to explore DMOs data:

- **Scatter plots** will be used to visualise the correlation of DMOs evaluated by the two systems (SS and RS), scatter plots of the DMOs from SS, versus the DMOs from RS, will be presented at WB level and participant level. We will report the correlation value (r) as well as the regression equation.
- **Bland Altman plots** (15) will be used to visualise and quantify the agreement between the systems (SS and RS) for each DMOs, at WB level and participant level. The Bland Altman plot shows the difference between the two measurements (DMOs (SS)- DMOs (RS), Y axis) versus the average of these measures ((DMOs(SS)+DMOs(RS))/2, X axis). The limits of agreement will also be provided and calculated as follows:

  $$\boldsymbol{Limits\ of\ Agreement}\ (\boldsymbol{LoA}) = 1.96\ \times SD\ of\ the\ difference\ (\mathrm{SS} - \ \mathrm{RS})$$

- **Histograms:** Histograms displaying the distribution of DMOs from the systems (SS and RS) and statistical metrics (e.g. relative errors) will be used to visualize the frequency and identify potential outliers, as indicated in Table 5 and Figure 13 and 14. The outliers presented will be assessed by exanimating all available information regarding this outlier (e.g. demographic and technical information).

Examples of the above mentioned plots are shown in Figure 13 and 14. Bland Altman, scatter and histogram plots will be used to represent continuous DMOs.

3.2.2 Examples of DMOs Analysis

Based on Table 5 and Figure 5, in Section 3.1 the DMOs have been categorised into 5 groups: 5 different set of analyses (corresponding to the colours used in Table 5 and Figure 5) according to their minimum aggregation level, structure and type of DMOs, and associated statistical analyses.

For each group of DMOs that uses similar set of analyses, **we present here one example** of the derivation and the statistical analyses that will be implemented (using simulated data). The examples refer to the evaluation of performance metrics and statistical analyses for one participant and, depending on the DMO, they will show:

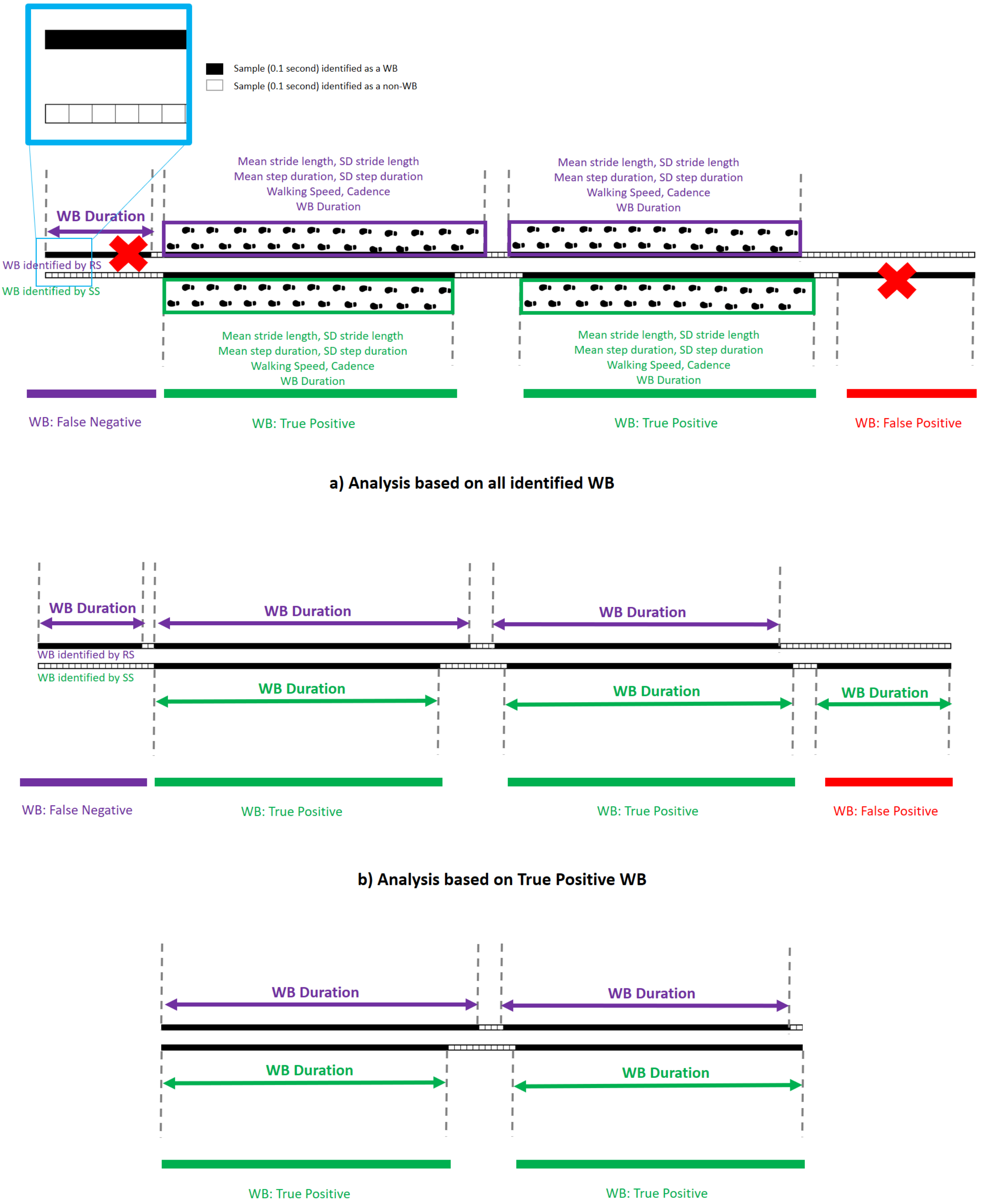
- **1st level of aggregation** – at WB level: descriptive statistics evaluated for the participant, **within each WB** presented in the examples.
- **2nd level of aggregation** – at patient level (across all WBs): descriptive statistics evaluated **for the participant, across all WBs** of the presented examples.

1. ***Primary DMO and Secondary DMOs: (i) Walking Speed, (ii) Duration of Walking bout, (iii) Step duration, (iv) Stride duration, (v) Cadence, (vi) Stride length, (vii) Duration of Turn, (viii) Maximal Angle of Turn, (ix) Stance phase duration, (x) Swing phase duration.***

Statistical analyses. Criterion validity: absolute and relative errors, concurrent validity (r, p and 95% confidence interval of ICC(2,1)), statistically significant difference tests. Plots: scatter plots for correlations, Bland Altman plots for agreement and histograms of raw DMOs data and relative errors.

Example data. The following example of data is based on the DMO "Walking Bout Duration". This example is applicable to any DMO which is obtained as a single value within a WB, i.e. a DMO whose minimal aggregation granularity is set to a WB level, as indicated in Figure 6 b. (e.g. average step duration, etc.) and coloured with green in Figure 5. See in Figure 9 the example on the DMO "WB Duration", for one participant and for a lab-test, presenting the identified WBs by the SS (WBs n = 3) and the WBs identified by the SS (WBs n = 3).

Note that a True positive WB will be considered when at least an 80% of the total duration of the WBs defined by the two systems of overlap in time.



Sample (0.1 second) identified as a WB
Sample (0.1 second) identified as a non-WB
WB Duration
WB identified by RS
WB identified by SS
Mean stride length, SD stride length
Mean step duration, SD step duration
Walking Speed, Cadence
WB Duration
Mean stride length, SD stride length
Mean step duration, SD step duration
Walking Speed, Cadence
WB Duration
Mean stride length, SD stride length
Mean step duration, SD step duration
Walking Speed, Cadence
WB Duration
Mean stride length, SD stride length
Mean step duration, SD step duration
Walking Speed, Cadence
WB Duration
WB: False Negative
WB: True Positive
WB: True Positive
WB: False Positive

a) Analysis based on all identified WB

b) Analysis based on True Positive WB

Figure 9. Example of the analysis of the DMO: "WB Duration", defined between the start and the end of a WB and evaluated by both systems (SS and RS). The RS and the SS identified 3 WBs, but the first WB identified by the RS is not identified by the SS (false negative); and the last WB identified by the SS has no corresponding WB from the RS (false positive); the two other WBs are considered true positives, as they are detected in both systems and are relatively near in time. Two type of analyses will be performed: a) based on all identified WBs, for which the absolute and relative errors will be computed for the "WB Duration", i.e. the WBs estimated independently in each system; b) based on true positive, i.e., based on only the WBs for which both systems identified a

corresponding WB. In this case, despite the absolute and relative errors of the mean and SD of the "WB Duration", we will also estimate the maximal absolute and relative error, and the RMS of the absolute error.

Based on the WB duration evaluated by the SS and RS, an example set of data is presented in Table 6, based on the data in Figure 9.

Table 6. Analysis performed on the WB Duration for both type of analyses: a) based on all identified WBs, and b) based on true positive WBs, and corresponding to Figure 9. The sign (*) indicates that the statistical measure (e.g., mean) is performed across all WBs available, in this case, for all WBs identified independently with each measurement system. The sign (**) indicates that the statistical measure (e.g., mean) is performed across all WBs available, i.e. for all true positive WBs.

	STATISTICAL METRICS	Walking Bout 1	Walking Bout 2	Walking Bout 3	Mean WB duration (s)
Based on all WBs	**WB Duration (s) - RS**	1.70	5.20	4.80	**3.90**
	WB Duration (s) - SS	4.70	4.70	2.10	**3.83**
	Absolute Error	0.07			
Based on True Positive WBs	**WB Duration (s) - RS**	5.20	4.80		5
	WB Duration (s) - SS	4.70	4.70		4.70
	Absolute Error (s)	0.50	0.10		

Some of the proposed statistical analyses across WBs for a single test or recording are shown in Table 7 and will be evaluated for each participant, based on the data presented in Table 6.

Table 7. Example of statistical analyses on the WB Duration for a single participant, for both type of analyses: a)* based on all identified WBs, and b)** based on true positive WBs and based on Table 6.

STATISTICAL METRICS	Summary Statistics across all WBs for a single participant and a single Lab Test or Real world 2.5 hour assessment
Absolute Error (s)*	0.07
Relative Error (%)*	1.80
Absolute Error – Mean (s)**	0.30
Absolute Error – SD (s)**	0.28
Relative Error – Mean (%)**	6.00
Relative Error – SD (%)**	5.32
Absolute Error (s) - Maximum**	0.50
Absolute Error (s) - RMS**	0.36
Relative Error (%) - Maximum**	9.61

2. *Secondary DMOs: (i) Number of Walking Bouts, (ii) Number of Turns*

Statistical analyses. Performance metrics: accuracy, sensitivity, specificity, positive predictive value, F1-Score.

Example Data. The following example is based on the DMO “Number of Walking bouts”. For calculating each of the afore mentioned performance metrics, there is a need to identify the true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) at a sample level of 0.1 second (corresponding to 10 samples of the SS raw data considering the sampling frequency of 100 Hz). Thus, for each sample (0.1 seconds) of data, the SS and the RS output will be analysed and compared. If a single sample (depicted as a rectangle in Figure 10) obtained from the RS is identified as part of a WB according to WP2 definition (coloured in black), and this is also identified in the SS as part of a WB, then this sample will be coded as a TP (green). If a sample is identified as part of a WB by the RS only (coloured black for the RS, but white for the SS), then this will be considered as a FN (purple). When a sample is identified as part of a WB from the SS only (coloured black for the SS, but white for the RS), this will be considered a FP (red). For all the samples which are not part of a WB for either of the systems (white in both), these will be considered a TN (blue). These samples correspond to activities other than walking, for example standing, shuffling or sitting. See example in Figure 10 below one participant for a lab-test comprising 24 samples of data (potential WB periods) with identified WB from the SS and the RS.

In the case of the DMO “Number of turns”, the identification/detection of turns will be only performed within all the samples corresponding to an identified WB.

An example of these performance metrics for three identified WBs is provided in Table 8.

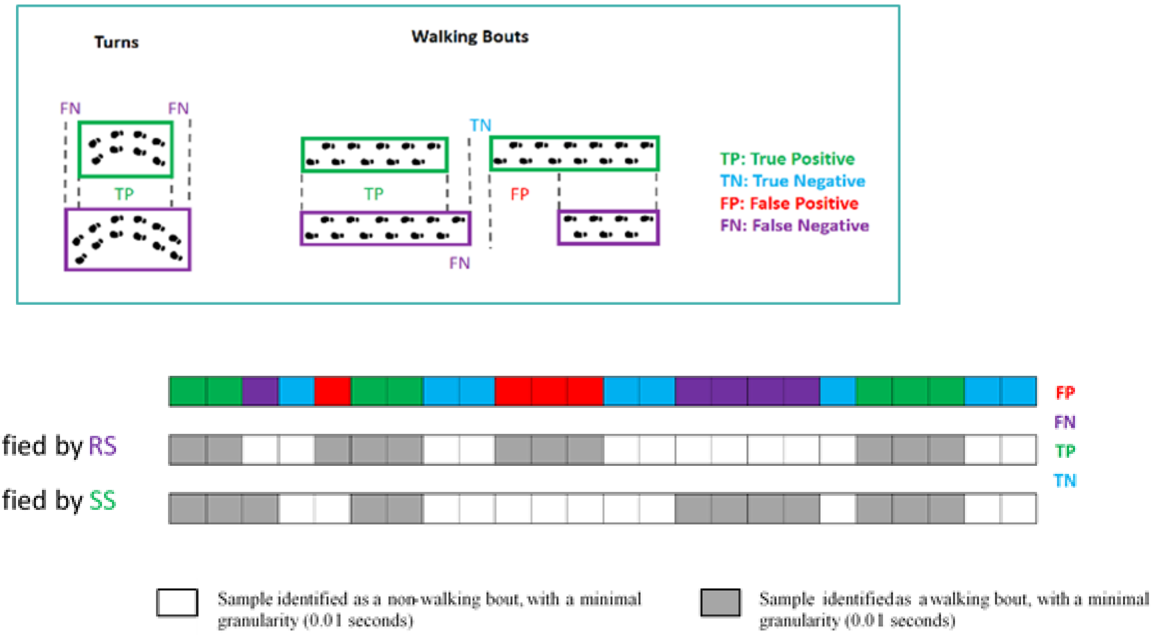

Turns
Walking Bouts
FN
FN
TP
TN
TP
FP
FN
TP: True Positive
TN: True Negative
FP: False Positive
FN: False Negative
FP
FN
TP
TN
WBs identified by RS
WBs identified by SS
Sample identified as a non-walking bout, with a minimal granularity (0.01 seconds)
Sample identifiedas a walking bout, with a minimal granularity (0.01 seconds)

Figure 10. Example of how each sample (0.1 seconds) is identified as TP, TN, FP and FN and compared between the SS and RS for the identification/detection of WBs for one participant. Each sample of the SS and RS outputs are depicted as a rectangle, where white rectangles represent samples of non-WB periods, and black rectangles denote samples of a detected WB period.

Table 8 below shows how the example in the figure above is further processed for the identified WB by the RS (number of WB n = 4), for one participant. Table 8 illustrates how the corresponding performance metrics are derived for each Lab test or for the whole Real world 2.5 hour assessment (and not for each WB), for each participant.

Table 8. Example of the samples of three walking bouts for one participant. “Non-WB” indicate samples which are not considered as part of a WB, whereas “WB” indicate samples which are considered as part of a WB.

Sample		SS	RS	WB
Sample 1	TP	WB	WB	WB1
Sample 2	TP	WB	WB	
Sample 3	FN	Non-WB	WB	
Sample 4	TN	Non-WB	Non-WB	
Sample 5	FP	WB	Non-WB	
Sample 6	TP	WB	WB	WB2
Sample 7	TP	WB	WB	
Sample 8	TN	Non-WB	Non-WB	
Sample 9	TN	Non-WB	Non-WB	
Sample 10	FP	WB	Non-WB	
Sample 11	FP	WB	Non-WB	
Sample 12	FP	WB	Non-WB	
Sample 13	TN	Non-WB	Non-WB	
Sample 14	TN	Non-WB	Non-WB	
Sample 15	FN	Non-WB	WB	WB3
Sample 16	FN	Non-WB	WB	
Sample 17	FN	Non-WB	WB	
Sample 18	FN	Non-WB	WB	
Sample 19	TN	Non-WB	Non-WB	
Sample 20	TP	WB	WB	WB4
Sample 21	TP	WB	WB	
Sample 22	TP	WB	WB	
Sample 23	TN	Non-WB	Non-WB	
Sample 24	TN	Non-WB	Non-WB	

TP = 7; **TN** = 8; **FP** = 4; **FN** = 5.

Table 9 shows the performance metrics derived for a single participant based on the samples identified as TP, TN, FP, FN.

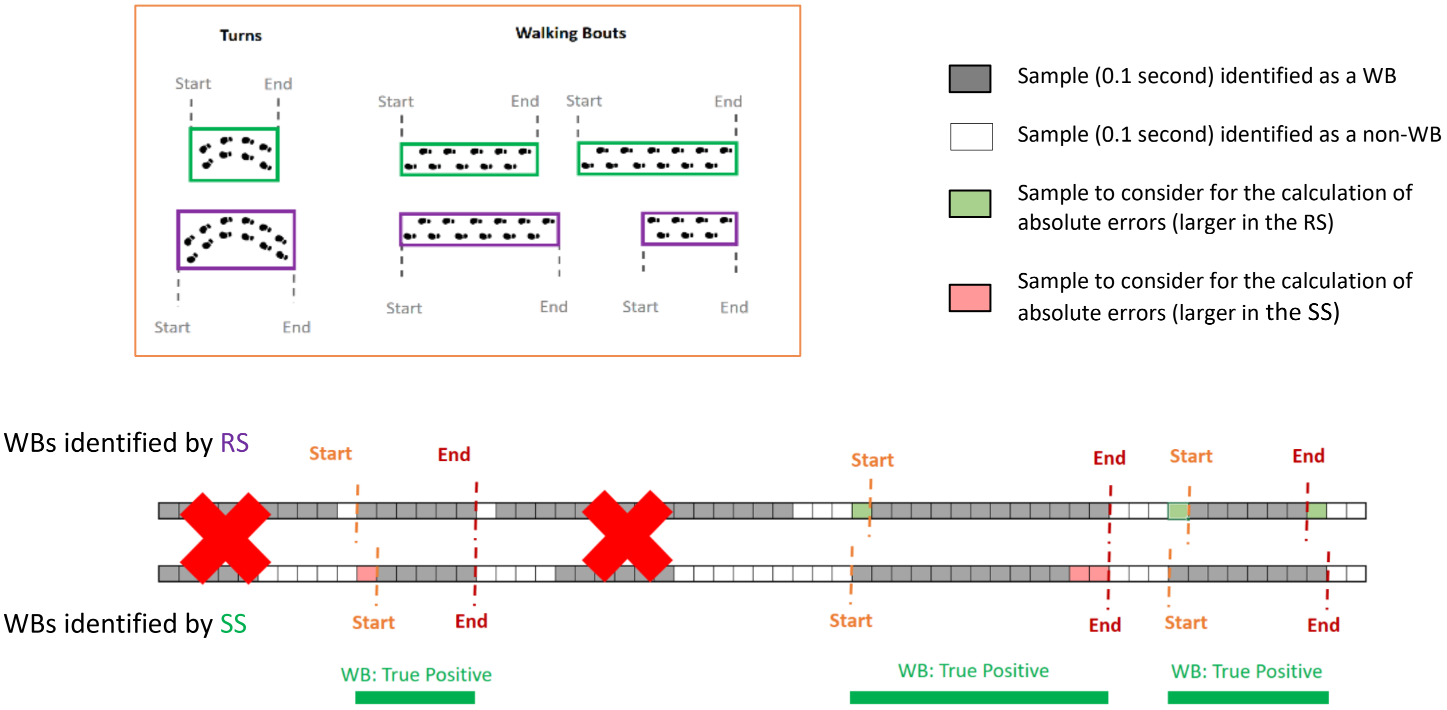
Table 9. Example of the performance metrics calculated for the example of Lab test for one participant as represented in Table 8.

Performance Metrics	**Overall Performance across all Walking bouts (Lab Test or Real world 2.5 hour assessment)**
Sensitivity	0.583
Specificity	0.667
Accuracy	0.625
Positive Predictive Value	0.636
F1 score	0.608

3. *Secondary DMOs: (i) Start and End of Walking Bouts, (ii) Start and End of Turns*

Statistical analyses. Criterion validity: absolute error and relative errors. Plots: histograms of relative errors.

Example data. The following example of data is based on the DMOs “Start of Walking Bout” and “End of Walking Bout” corresponding to true positive walking bouts, based on the timing events of the starts and ends identified by the two systems of comparison (tested: SS vs. reference: RS, Figure 11). These examples are applicable to the start/end of turs (which are identified within WBs). The schema (left side) also represents the start/end of turns, as in Figure 5. See in Figure 11, below, the example for one participant and a lab-test, presenting the identified WBs by the SS (WBs n = 5) and the WBs identified by the RS (WBs n = 5), from which only 3 of the WBs are considered true positive WBs in both systems, and based on which the DMOs “Start” and “End” will be assessed.



Turns
Walking Bouts
Start
End
Start
End
Start
End
Start
End
Start
End
Start
End
Sample (0.1 second) identified as a WB
Sample (0.1 second) identified as a non-WB
Sample to consider for the calculation of absolute errors (larger in the RS)
Sample to consider for the calculation of absolute errors (larger in the SS)
WBs identified by RS
Start
End
Start
End
Start
End
WBs identified by SS
Start
End
Start
End
Start
End
WB: True Positive
WB: True Positive
WB: True Positive

Figure 11. Example of the start and end time (events) identified for a walking bout, and how absolute errors (“differences” marked in light red and light green) between the true positive WBs from the SS and RS are estimated for one participant and a lab test or 2.5 hours assessment.

Based on the time difference between the start and end of the DMOs (e.g. walking bout), the statistical metrics will be calculated (see section 3.2.1). Examples are presented in Table 10, based on Figure 11.

Table 10. Examples of start and end time events in seconds (s), derived from the SS and RS over three WBs.

		Walking Bout 1			Walking Bout 2			Walking Bout 3		
		Start	**Duration**	**End**	**Start**	**Duration**	**End**	**Start**	**Duration**	**End**
Based on True Positive WBs	**RS - time (s)**	1.000	0.600	1.600	3.600	1.200	4.800	5.200	0.600	5.800
	SS - time (s)	1.100	0.500	1.600	3.500	1.100	4.600	5.100	0.800	5.900
	Absolute Error (s)	0.100	0.100	0.000	0.100	0.100	0.200	0.100	0.200	0.100
	Relative Error (%)	0.167	0.167	0.000	0.083	0.083	0.167	0.167	0.333	0.167

Table 11 shows an example (3 walking bouts for a participant) of some of the statistical analyses proposed for the DMOs “start” and “end of WBs”.

Table 11. Example of summary statistical analyses for the DMOs “start” and “end of WBs” for one participant, based on the data from Table 10.

	DMO	**STATISTICAL METRICS**	**Min across all WBs**	**Max across all WBs**	**Mean across all WBs**	**SD across all WBs**	**Median across all WBs**	**IQR across all WBs**	**RMS across all WBs**
Based on True Positive WBs	Start	**Absolute Error (s)**	0.100	0.100	0.100	0.000	0.100	0.000	0.100
		Relative Error (%)	0.083	0.167	0.139	0.049	0.167	0.063	0.145
	End	**Absolute Error (s)**	0.000	0.200	0.100	0.100	0.100	0.150	0.129
		Relative Error (%)	0.000	0.167	0.111	0.096	0.167	0.125	0.136
	Duration	**Absolute Error (s)**	0.100	0.200	0.133	0.058	0.100	0.075	0.141
		Relative Error (%)	0.083	0.333	0.194	0.127	0.167	0.188	0.220

4. *Secondary DMOs: (i) Initial Contact and Final Contact Events*

Statistical analyses. Performance measures: sensitivity, positive predictive value, F1-score. Criterion validity: absolute and relative errors. Plots: histogram of relative errors.

Example Data. The following example is for “Initial Contacts (IC) time events” detected as for each of the three WBs, for a single participant. The calculation of sensitivity, positive predictive value and F1-Score requires to determine the nature of each initial/final contact event within each WB (as TP, FP and FN events). This is achieved by comparing the events detected by the SS to those detected by the RS (Figure 12). In this analysis, TN will not be evaluated, since TN correspond to all non-IC events/samples, which do not exist due to the way this data is processed. Thus, specificity and accuracy cannot be evaluated.

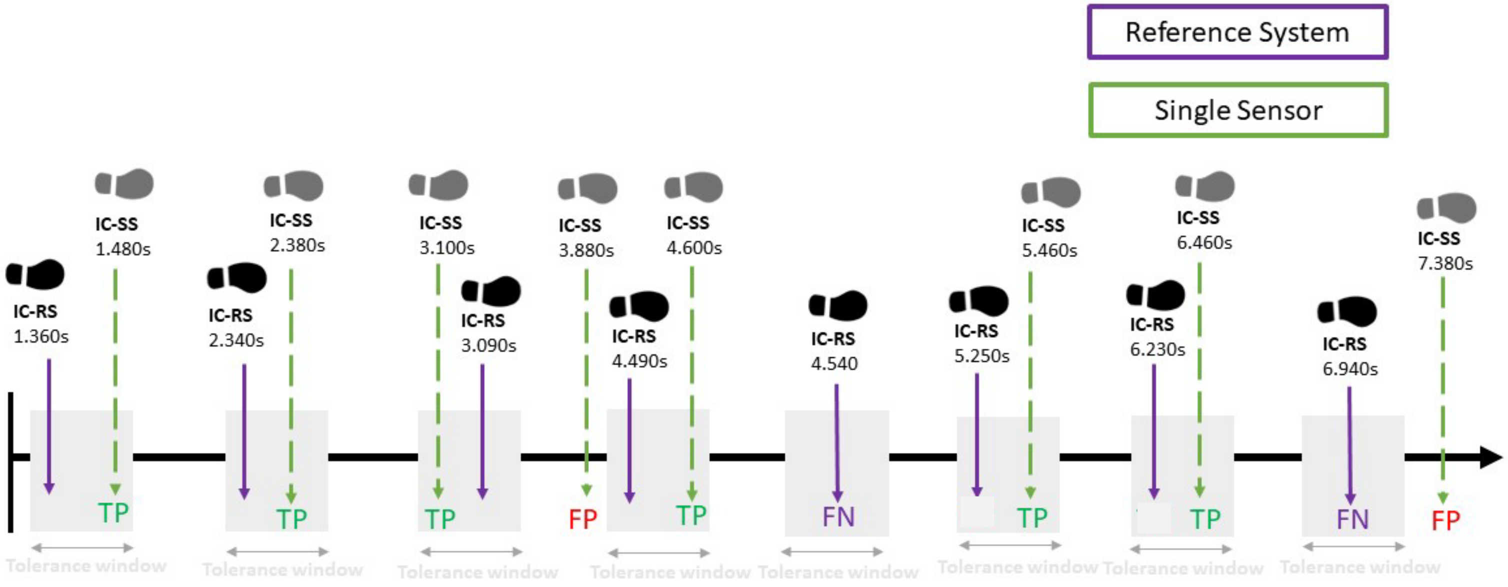


Figure 12. Example of Initial Contact identification based on ICs identified for Walking Bout 1 in Table 12. This figure presents how FN, FP and TP events are identified based on the comparison of IC events detected by the SS (IC-SS in red) versus the ones detected by the RS (IC-RS in green) for one participant. FN, FP and TP are defined with respect to the selected temporal tolerance window (TW), in grey, as defined within WP2 Task 2.2.2.

Figure 12 presents the TP, FP and FN with respect to the tolerance window (TW). The TW is defined as a fixed interval of 0.5 data (8, 16). The TW is centred on each initial contact detected with the RS (IC-RS) and determines the temporal interval in which an initial contact from the SS (IC-SS) may be identified as either a TP or FP.

The TW is centred on each IC-RS, thus, the TW starts 0.25 seconds before the IC-RS and ends 0.25 seconds after the IC-RS (17). A single IC-RS is set in the middle of each TW.

For each WB, there will be the same number of TW as the number of IC-RS.

The nearest IC-SS (in time) with respect to the IC-RS within the specified TW will be considered a TP. Each IC-RS will lead to a single TP, as long as there is an IC-SS within the TW. In case there are several IC-SS within a single TW, only the IC-SS that is nearest in time to the IC-RS would be considered a TP, whereas the others will be considered as FP.

All the IC-SS identified outside the TW will be considered as FP. FN will be identified when an IC-RS is not within the TW and does not correspond to any IC-SS.

The data presented in Figure 12 correspond to the ICs identified for Walking Bout 1 in Table 12.

Note that to estimate relative errors, the differences between SS and RS outputs (absolute errors) are divided by the average step duration estimated by the RS. An example of the summary statistical analyses at WB level is shown in Table 13, based on the data presented in Table 12. These correspond to the statistical metrics obtained within each available WB (including all the ICs of the respective WB). In Table 14 we also present some of the statistical measures obtained over all available WBs of a Lab test /Real world 2.5 hour assessment in a single participant, derived from the exemplary data in Table 12:

Table 12. IC events in seconds (s) detected by the SS and RS for one participant and three WBs. The info corresponding to Walking Bout 1 is presented in figure 12.

DMO: IC (s)	**Walking Bout 1**				**Walking Bout 2**				**Walking Bout 3**			
		SS		**RS**		**SS**		**RS**		**SS**		**RS**
IC 1	**TP**	1.360		1.480	**FP**	10.140	**FN**	10.810	**TP**	86.225		86.300
IC 2	**TP**	2.340		2.380	**TP**	11.290		11.440	**TP**	86.925		87.040
IC 3	**TP**	3.100		3.090	**TP**	12.150		11.990	**TP**	87.675		87.750
IC 4	**FP**	3.880		4.490	**TP**	12.650		12.560	**TP**	88.475		88.440
IC 5	**TP**	4.600	**FN**	4.540	**TP**	13.390		13.210	**TP**	89.125		89.230
IC 6	**TP**	5.460		5.250	**TP**	13.910		13.750	**TP**	90.100		90.310
IC 7	**TP**	6.460		6.230	**TP**	14.470		14.280	**TP**	91.200		91.150
IC 8	**FP**	7.380	**FN**	6.940	**TP**	15.080		14.920	**FP**	92.550	**FN**	91.820
IC 9					**TP**	15.950	**FN**	15.220	**FP**	93.575	**FN**	93.850
IC 10								15.840	**FP**	94.525	**FN**	94.220

Table 13. Example for derivation of metrics across all IC events detected within each available WB for participant, based on data presented in Table 12.

STATISTICAL METRICS	Walking Bout 1	Walking Bout 2	Walking Bout 3
Sensitivity	0.750	0.889	0.700
Positive Predictive Value	0.750	0.800	0.700
F1-score	0.750	0.842	0.700
Absolute Error – Mean (s)	0.120	0.150	0.095
Absolute Error – SD (s)	0.088	0.034	0.058
Relative Error – Mean (%)	15.190	26.200	11.198
Relative Error – SD (%)	11.151	5.905	6.831
Absolute Error (s) - Maximum	0.230	0.190	0.210
Absolute Error (s) - RMS	0.145	0.153	0.109
Relative Error (%) - Maximum	29.114	33.188	24.754

Table 14. Example of summary statistical analyses evaluated across all available WBs for one participant, based on data presented in Table 13.

STATISTICAL METRICS	Min across all WBs	Max across all WBs	Mean across all WBs	SD across all WBs	Median across all WBs	IQR across all WBs	RMS across all WBs
Sensitivity	0.700	0.889	0.780	0.098	0.750	0.142	0.784
Positive Predictive Value	0.700	0.800	0.750	0.050	0.750	0.075	0.751
F1-score	0.700	0.842	0.764	0.072	0.750	0.107	0.766
Absolute Error – Mean (s)	0.095	0.150	0.122	0.028	0.120	0.041	0.124
Absolute Error – SD (s)	0.034	0.088	0.060	0.027	0.058	0.041	0.064
Relative Error – Mean (%)	11.198	26.200	17.529	7.770	15.190	11.251	18.642
Relative Error – SD (%)	5.905	11.151	7.962	2.800	6.831	3.934	8.284
Absolute Error (s) - Maximum	0.190	0.230	0.210	0.020	0.210	0.030	0.211
Absolute Error (s) - RMS	0.109	0.153	0.136	0.023	0.145	0.033	0.137
Relative Error (%) - Maximum	24.754	33.188	29.019	4.218	29.114	6.325	29.222

5. *Secondary DMOs: Initial Contact Events: (i) Laterality of Left-Right Feet, (ii) Sequence Order or Left-Right Feet*

Statistical analyses. Criterion validity: absolute and relative errors of laterality and sequencing order, and Cohen’s kappa of categorical labels.

Example Data. Example data is presented to show the Initial Contact events identified within each of the 3 WBs (true positives), for a single participant, and the assigned categories/labels of the laterality “Left” or “Right” corresponding to each Initial Contact (Table 15). Note that the labels for assessment are based on the true positive initial contact events. Table 15 provides the exemplary data based on the true positive IC events determined for Walking Bout 1 in Table 12.

Table 15. Timing in seconds (s) of the IC events (based on the true positives) and the assigned laterality labels (Left/Right) identified by the SS and the RS for 3 walking bouts in one test of a single participant.

DMO: IC (s)	**Walking Bout 1**				**Walking Bout 2**				**Walking Bout 3**			
		SS		**RS**		**SS**		**RS**		**SS**		**RS**
IC 1	**Left**	1.360	**Left**	1.480	**Right**	11.290	**Right**	11.440	**Left**	86.225	**Left**	86.300
IC 2	**Right**	2.340	**Right**	2.380	**Left**	12.150	**Left**	11.990	**Right**	86.925	**Right**	87.040
IC 3	**Left**	3.100	**Left**	3.090	**Right**	12.650	**Right**	12.560	**Left**	87.675	**Left**	87.750
IC 4	**Left**	4.600	**Right**	4.490	**Right**	13.390	**Left**	13.210	**Right**	88.475	**Right**	88.440
IC 5	**Left**	5.460	**Left**	5.250	**Right**	13.910	**Right**	13.750	**Left**	89.125	**Left**	89.230
IC 6	**Right**	6.460	**Right**	6.230	**Right**	14.470	**Right**	14.280	**Right**	90.100	**Right**	90.310
IC 7					**Left**	15.080	**Left**	14.920	**Right**	91.200	**Left**	91.150
IC 8					**Right**	15.950	**Right**	15.840				
IC 9												
IC 10												

The data is presented for IC events with the assigned laterality categories (“Left” or “Right”). The timing of IC events is presented in seconds. Note that the criterion validity of the DMOs needs to be assessed by computing the number of different labels assigned to ICs (i.e. laterality, referring to the foot with which the initial contact event is performed) and the wrong sequencing, i.e. by assuming that after each right step there is a left step, and vice versa, and always in comparison to the reference system. These analyses will only be performed on the ICs (obtained from the SS) that are True Positives (so ICs identified by the RS and correctly identified also by the SS). Note that to assess the Sequence Order, only absolute and relative errors will be computed. Absolute and relative errors will be computed as detailed below:

- **Laterality - Absolute Error:** To validate and compare the performance of the algorithms we will count the number of labels wrongly assigned to the ICs from the SS, with respect to the total labels assigned by the RS. Note that the positive predictive value cannot be estimated, since all ICs will be assigned a label/category (either wrong or correct), precluding the estimation of FP.

- **Laterality - Relative Error:** Absolute laterality errors will be counted within each WB and divided by the total number of ICs within that particular WB, as identified by the RS. Thus, the number of errors will be divided by the total number of ICs within a WB.

- **Laterality - Concurrent Validity:** Cohen’s kappa (k) will be computed for all the labels (i.e. Left or Right) corresponding to the SS and the RS. The r value (correlation coefficient), 95% CI and p value will be reported. All p values will be reported regardless of statistical significance. We will obtain a single p value and a single Cohen’s kappa (k) value for each cohort (e.g. PD) or group (e.g. slow PD). Particularly, we will aggregate all the labels of all the ICs, from all the WBs of all the participants of a cohort/group.

- **Sequencing order – Absolute Error:** We will check whether the correct sequencing occurs seemingly for SS than for RS. For example, for every “Left” category we will ascertain whether it is followed by a “Right” category in both systems, and continues to alternate (after Left, Right; after Right, Left), also in both systems. The number of wrong sequences (which does not coincide between the SS and RS) will be counted for each WB.

- **Sequencing order – Relative Error:** The absolute errors reported for the sequencing order will be counted for a WB and divided by the total number of ICs within that particular WB detected by the RS.

Table 16 shows an example of some of the proposed statistical analyses obtained individually for each available WB, and based on the data from Table 15. In addition, in Table 17 (for a Lab test/ Real world 2.5 hours assessment) we present the summary statistical analyses for a participant.

Table 16. Example of statistical analyses for the assessment of "Laterality" and "Sequence order", over all available WBs (n=3) in one participant, based on data presented in Table 15.

		Walking Bout 1	Walking Bout 2	Walking Bout 3
Laterality	Absolute Number	1.000	2.000	1.000
	Relative Number (%)	0.167	0.250	0.143
Sequence Order	Absolute Number	2.000	4.000	1.000
	Relative Number (%)	0.333	0.500	0.143
Correct labels	Absolute Number	5.000	6.000	6.000
	Relative Number (%)	0.833	0.750	0.857

Table 17. Example of summary statistical analyses evaluated across all available WBs for one participant, based on data presented in Table 16.

		Min across all WBs	Max across all WBs	Mean across all WBs	SD across all WBs	Median across all WBs	IQR across all WBs
Laterality	Absolute number	1.000	2.000	1.333	0.577	1.000	0.750
	Relative number	0.143	0.250	0.187	0.056	0.167	0.080
Sequence order	Absolute number	1.000	4.000	2.333	1.528	2.000	2.250
	Relative number	0.143	0.500	0.325	0.179	0.333	0.268
Correct labels	Absolute number	5.000	6.000	5.667	0.577	6.000	0.750
	Relative number	0.750	0.857	0.813	0.056	0.833	0.080

Plots: Scatter Plot, Bland Altman Plot and Histogram

Utilising the visualisation toolbox on examples of cadence values estimated by the SS and the RS, scatter plots and Bland-Altman plots will be used to explore results on the following aggregation levels:

- **1st level of aggregation**: for all participants belonging to a cohort (e.g. HA), mean cadence values evaluated for each WB (one WB = one data point) detected within a lab test or 2.5 hour assessment, Figure 13. This plots will represent only **WBs that are True Positives** (i.e. a WB identified by the RS and that is correctly also identified by the SS, so we will have a correspondence of values for both systems)

- **2nd level of aggregation:** for each participant (one participant = one data point) belonging to a cohort (e.g. HA), mean values of cadence averaged across all WBs ("mean of means") identified within a lab test or 2.5 hour assessment, Figure 14.

The distribution of the variables used for axis of each graph (e.g. distribution of the mean cadence SS (y axis) and the mean cadence RS for the scatter plot (x axis)) are depicted in grey, on each axis, using histograms.

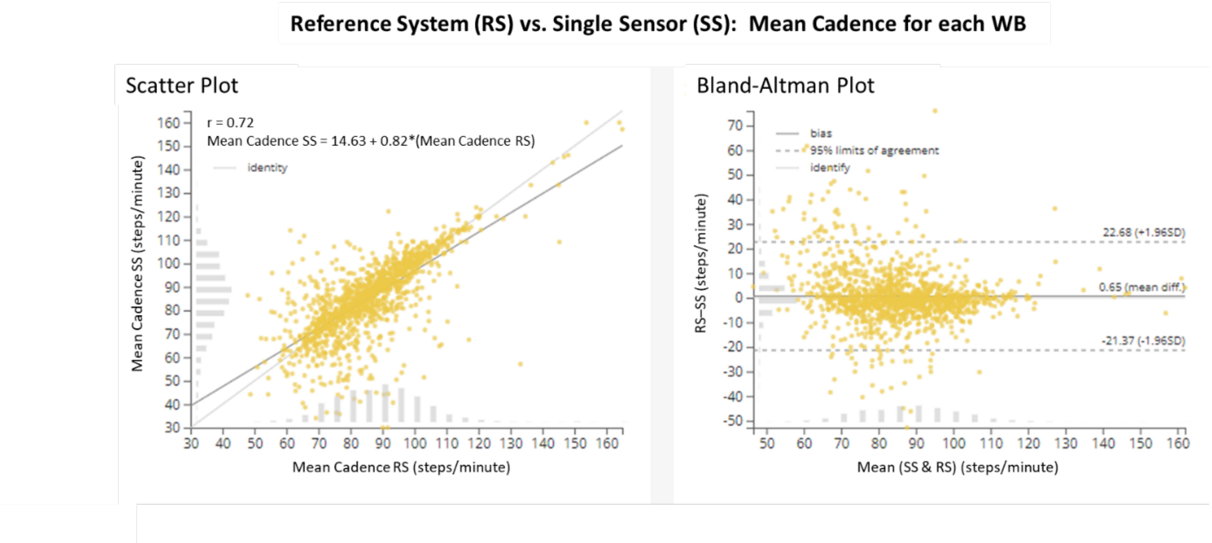


Figure 13. Scatter plot (on the left, with correlation results (r) and regression equation) and Bland Altman plot (on the right, with limits of agreement) of mean cadence values evaluated for each WB identified for all participants of a cohort (HA, one data point for each WB identified); histograms of the variables used for each graph (e.g. Mean cadence SS and Mean cadence RS for the scatter plot, etc.) are included on the x and y axis of each graph, in grey.

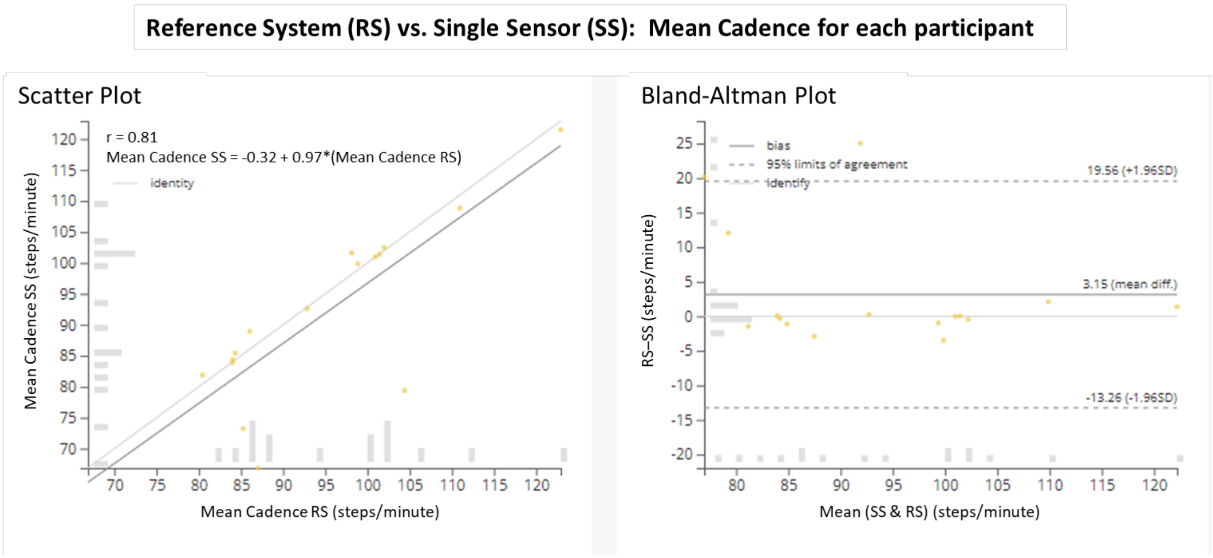


Figure 14. Scatter plot (on the left, with correlation results (r) and regression equation) and Bland Altman plot (on the right, with limits of agreement) of mean cadence values averaged across all WBs ("mean of means") identified for each participant of a cohort (HA, one data point for each participant); histograms of the variables used for each graph (e.g. Mean cadence SS and Mean cadence RS for the scatter plot, etc.) are included on the x and y axis of each graph, in grey.

3.3 Supplementary and subgroup analysis

Supplementary and subgroup analyses will be performed only if there are at least 5 participants (~25% of the full dataset for each cohort) in each sub-group.

3.3.1. Impact of Walking Bout length on results: To assess whether the criterion validity results are dependent on walking bout length, the presented analysis will be repeated based on the length of the WB (evaluated by the RS), considering only the Real world 2.5 hours assessment data.

All of the analyses presented in section 3.2.2 will be repeated on the following groups of WBs that are considered to be relevant for evaluation of DMOs (18-21).

- Group 1: ∀ Walking Bouts < 10 seconds
- Group 2: ∀ Walking Bouts > 10
- Group 3: 10 seconds < Walking Bouts < 30 seconds
- Group 4: 30 seconds < WB < 60 seconds
- Group 5: 60 seconds < WB < 120 seconds
- Group 6: WB > 120 seconds

3.3.2. Impact of contextual factors on results: The objective in this case is to determine the extent to which the performance of the different DMOs algorithms is affected by extrinsic confounding factors such as whether the WB was performed indoor or outdoor, with or without a walking aid, or walking on a flat or inclined path. To assess whether construct validity results are dependent on contextual factors the presented analysis of DMOs algorithm performance will be repeated based on the context (for the Real world 2.5 hours assessment).

The first part of this analysis will involve the assignment of contextual labels to the Real world 2.5 hour assessment walking bouts for each participant. The specific context labels, which will be considered in pairs of opposites, included in this analysis will include:

1. Indoor and outdoor WBs
2. WBs with and without a walking aid

Following assignment of context labels groups of WBs that align to each of the labels in the opposite pairs will be assembled. This will be done on the basis of matching pairs of WBs with similar durations for each participant. All of the analyses presented in section 3.2.2 will be repeated considering e.g.:

- all indoor WBs <10 seconds and all outdoor WBs <10 seconds;
- all WBs <10 seconds labelled with walking aid, all WBs <10 seconds labelled without walking aid;

and so on, including all the WB length described in section 3.3.1.

Comparative analysis of DMOs algorithm performance will then be performed on the basis of WB by contextual factor for each of the cohorts/groups.

3.3.3. Impact of covariates on results: We will analyse the results in subgroup analyses, not only based on the disease cohort, but as well considering other covariate:

- Biomechanical information: gait speed (for gait speed, this will be done by performing a subgroup analysis of the clinical cohorts based on their average stride gait speed, stratifying them on three levels: fast speed: > 1 m/s, medium speed: 0.5m/s - 1 m/s, slow speed < 0.5 m/s) (10).

3.4 Handling of missing data and identification of outliers

3.4.1 Missing values

Data values may be missing or unavailable (e.g. test within the Laboratory assessment (e.g. "Back and forth walk (preferred gait speed) along a straight walkway") due to failure of the recording device or technical issues, such as not detecting enough strides to identify a walking bout (as per WB definition criteria). All missing data are thus considered completely at random and will be reported in %. The technical reason for which data will be missing will be explored and reported for all involved systems, particularly in the case of a single system failing in listings.

For example:

1. The data for participant X for the real-world 2.5 hours assessment were missing for the RS because the recording was incomplete, and therefore excluded from the analysis.

2. The data for participant X during lab-based assessment, test 2 were missing due to synchronisation technical problems, and therefore excluded from the analysis.

Assuming that all missing data are missing completely at random, complete case approach (i.e. pairwise deletion) will be utilised to handle missing data (22). Since we are interested in the analysis at walking bout level and not at patient level, even if some participants missed some specific assessment (e.g., one of the tasks in the laboratory) or observations (e.g. 2.5hs), their remaining available data will still be included in the analyses.

3.4.2 Outliers

All variables will be examined for outliers. An outlier is defined as a data point exceeding 2 standard deviations from the mean in the cohort. Outliers will be reported for each DMOs and visually investigated (e.g. scatterplots and histogram) (23, 24). If the outlier occurred due to technical fault (e.g. faulty sensor, incorrect sensor set-up/ calibration etc.), then the outlier will be excluded and reported in the missing section (see 3.4.1). All other outliers will remain in the analysis.

4. References

1. Camargo J, Ramanathan A, Csomay-Shanklin N, Young A. Automated gap-filling for marker-based biomechanical motion capture data. Computer methods in biomechanics and biomedical engineering. 2020;23(15):1180-9.

2. Trojaniello D, Cereatti A, Pelosin E, Avanzino L, Mirelman A, Hausdorff JM, et al. Estimation of step-by-step spatio-temporal parameters of normal and impaired gait using shank-mounted magneto-inertial sensors: application to elderly, hemiparetic, parkinsonian and choreic gait. J Neuroeng Rehabil. 2014;11:152.

3. Bertoli M, Cereatti A, Trojaniello D, Avanzino L, Pelosin E, Del Din S, et al. Estimation of spatio-temporal parameters of gait from magneto-inertial measurement units: multicenter validation among Parkinson, mildly cognitively impaired and healthy older adults. Biomedical engineering online. 2018;17(1):58.

4. Rossanigo R, Caruso M, Salis F, Bertuletti S, Della Croce U, Cereatti A, editors. An Optimal Procedure for Stride Length Estimation Using Foot-Mounted Magneto-Inertial Measurement Units. 2021 IEEE International Symposium on Medical Measurements and Applications (MeMeA); 2021: IEEE.

5. Bonci T, Keogh A, Del Din S, Scott K, Mazzà C, On Behalf Of The Mobilise DC. An Objective Methodology for the Selection of a Device for Continuous Mobility Assessment. Sensors (Basel, Switzerland). 2020;20(22).

6. Kluge F, Del Din S, Cereatti A, Gaßner H, Hansen C, Helbostad JL, et al. Consensus based framework for digital mobility monitoring. PloS one. 2021;16(8):e0256541.

7. Trojaniello D, Ravaschio A, Hausdorff JM, Cereatti A. Comparative assessment of different methods for the estimation of gait temporal parameters using a single inertial sensor: application to elderly, post-stroke, Parkinson's disease and Huntington's disease subjects. Gait Posture. 2015;42(3):310-6.

8. Del Din S, Godfrey A, Rochester L. Validation of an Accelerometer to Quantify a Comprehensive Battery of Gait Characteristics in Healthy Older Adults and Parkinson's Disease: Toward Clinical and at Home Use. IEEE J Biomed Health Inform. 2016;20(3):838-47.

9. Hickey A, Del Din S, Rochester L, Godfrey A. Detecting free-living steps and walking bouts: validating an algorithm for macro gait analysis. Physiological measurement. 2017;38(1):N1-n15.

10. Studenski S, Perera S, Patel K, Rosano C, Faulkner K, Inzitari M, et al. Gait speed and survival in older adults. Jama. 2011;305(1):50-8.

11. Walther BA, Moore JL. The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. Ecography. 2005;28(6):815-29.

12. ISO I. 5725-1: 1994, Accuracy (trueness and precision) of measurement methods and results-Part 1: General principles and definitions. International Organization for Standardization, Geneva. 1994.

13. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychological methods. 1996;1(1):30.

14. Wilcoxon F. Individual comparisons by ranking methods. Biometrics Bulletin, 1, 80-83. Reprinted in the Bobbs-Merrill Reprint Series in the Social Sciences, S-541. 1945.

15. Giavarina D. Understanding Bland Altman analysis. Biochemia medica. 2015;25(2):141-51.

16. Najafi B, Aminian K, Paraschiv-Ionescu A, Loew F, Bula CJ, Robert P. Ambulatory system for human motion analysis using a kinematic sensor: monitoring of daily physical activity in the elderly. Biomedical Engineering, IEEE Transactions on. 2003;50(6):711-23.

17. Tietsch M, Muaremi A, Clay I, Kluge F, Hoefling H, Ullrich M, et al. Robust Step Detection from Different Waist-Worn Sensor Positions: Implications for Clinical Studies. Digital biomarkers. 2020;4(1):50-8.

18. Del Din S, Galna B, Godfrey A, Bekkers EMJ, Pelosin E, Nieuwhof F, et al. Analysis of Free-Living Gait in Older Adults With and Without Parkinson's Disease and With and Without a History of Falls: Identifying Generic and Disease-Specific Characteristics. J Gerontol A Biol Sci Med Sci. 2019;74(4):500-6.

19. Del Din S, Godfrey A, Galna B, Lord S, Rochester L. Free-living gait characteristics in ageing and Parkinson's disease: impact of environment and ambulatory bout length. J Neuroeng Rehabil. 2016;13(1):46.

20. Shah VV, McNames J, Harker G, Mancini M, Carlson-Kuhta P, Nutt JG, et al. Effect of Bout Length on Gait Measures in People with and without Parkinson’s Disease during Daily Life. Sensors. 2020;20(20):5769.

21. Del Din S, Yarnall AJ, Barber TR, Lo C, Crabbe M, Rolinski M, et al. Continuous Real-World Gait Monitoring in Idiopathic REM Sleep Behavior Disorder. Journal of Parkinson's disease. 2020;10(1):283-99.

22. Kang H. The prevention and handling of the missing data. Korean journal of anesthesiology. 2013;64(5):402-6.

23. Rousseeuw PJ, Hubert M. Robust statistics for outlier detection. Wiley interdisciplinary reviews: Data mining and knowledge discovery. 2011;1(1):73-9.

24. Cousineau D, Chartier S. Outliers detection and treatment: a review. International Journal of Psychological Research. 2010;3(1):58-67.

5. Appendices

Appendix A: Sample size evaluation for TVS study

Sample size evaluation for the TVS: Sample size based on number of walking bouts estimations required for mobilise-d TVS.

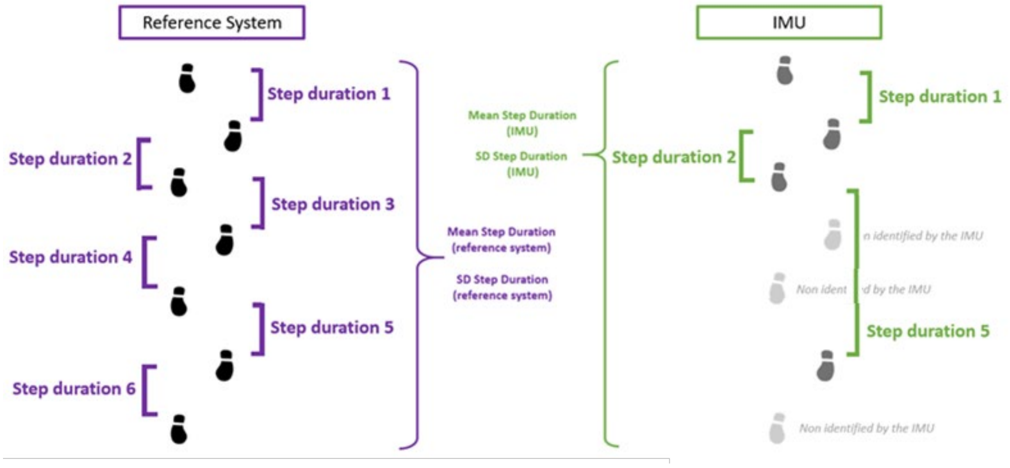
Objective: to validate the estimation of DMOs in real life vs gold standard measures

Main estimate: ICC(2,1) of DMOs measured with one system vs the same DMOs measured with another system

Participants: 6 groups (healthy, PD, MS, COPD, CHF, PFF)
→ sample size for each cohort

Design "TVS condition 3 (2.5 hours assessment)" see Figure:
During 2.5 h patients wear a single sensor (SS) and a multi-sensor system (INDIP) while doing free-living activities.
DMOs are thus measured by the INDIP (used as a reference system) and SS (system to be validated).
During those 2.5 h, patients will perform a number of bouts (ranging from 3 to 139, mean≈50 based on preliminary data from TVS) and DMOs (eg step duration, walking speed) are measured at each bout.

➔ Sample size refers to number of bouts – not number of patients



Sample size estimations

Using the following Stata code:
sampicc <p1> <reps> [, Alpha(#) Power(#) Sample(#) Width(#) CI]

where
p1= desired ICC(2,1) coefficient (≥0.7)
reps=2 (two tests)
Alpha=0.05
Beta=0.9 (to be conservative, instead of the traditional 0.8, because we are testing many parameters)
Width= confidence interval around ICC(2,1), we look for a range from 0.1 to 0.4 (we do not only want to reject the null (that ICC=0) but also want to see it with precision ie small CI)

Eg: sampicc 0.7 2, alpha(0.05) power(0.9) w(0.1) ci

Number of observations (i.e. bouts)*patient required

	Minimum ICC(2,1) to identify as statistically significant		
Desired width of CI	**0.6**	**0.7**	**0.8**
0.1	630	401	200
0.2	158	101	51
0.3	71	45	23
0.4	40	26	13

Total observations by disease based on min and max number of bouts * 20 patients

Data from TVS first patients 2.5h acquisitions (INDIP)					Estimations of totals in 20 patients	
Cohort	Participants	n WB min	n WB max	Total number of bouts	Mean number of bouts/patient	Expected total n of bouts in 20 'average' patients
COPD	3	38	92	222	74	1480
MS	11	15	112	597	54	1080
PD	15	3	139	613	41	820
PFF	8	7	111	388	49	980

Interpretation – sample size evaluation:
The current analysis confirms that 20 participants for PD, MS, PFF, HA, 17 participants for COPD and 15 for CHF is a sufficient sample size.

Appendix B: Templates for statistical outputs.

TVS-SAP Appendix (attached excel file): the attached file presents the template of the results for the statistical analysis that will be evaluated for each assessment (e.g. Laboratory assessment and each comparison: SS vs. SP or SS vs. INDIP; Real world 2.5 hour assessment for the comparison SS vs. INDIP). Each DMO will be reported in a different sheet.

Results will be provided for each of participant present in the TVS dataset (e.g. 1001, 1002, etc.) and also on a disease cohort level (e.g. Group 1 (e.g. PD), etc.). Results are averaged over a test for those DMOs in which a single value is provided per WB. In the case of DMOs with multiple values per WB, the average over the available WBs will be reported, after averaging over all the DMOs values within the WB. This Appendix B only presents an example of the results template for a single assessment (e.g. Laboratory assessment, Real world 2.5 hour assessment).

Appendix C: Example of a DMO results dataset used for the visualisation tool: results exploration and plot generation.

Example of Dataset for visualisation tool (attached .csv file): the attached file presents the template of a DMO results dataset (cadence in the example) that will be loaded on the visualisation toolbox platform for data exploration and plot generation purposes. Each raw corresponds to a single WB for each participant, columns represent all available information (clinical, demographic, statistical performance metrics, type of assessment, etc.). The file includes, for each DMO, results from the Lab and 2.5 hour assessments.

Each DMO will be reported in a different .csv file.

Note:

All WP2 deliverables can also be found in the following SharePoint folder:

https://newcastle.sharepoint.com/sites/IMI-MOBILISE-D/Shared%20Documents/Forms/AllItems.aspx?viewid=c4ca292e%2D4858%2D483a%2D97b8%2Dc2b7fdecd3a7&id=%2Fsites%2FIMI%2DMOBILISE%2DD%2FShared%20Documents%2FMobilise%2DD%20Work%20Packages%2FWork%20Package%202%2FReports%20and%20Deliverables%2FDeliverables%2FSubmitted