PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	A MULTI-CENTRIC OBSERVATIONAL STUDY PROTOCOL FOR THE TECHNICAL VALIDATION OF REAL-WORLD MONITORING OF GAIT
AUTHORS	Mazzà, Claudia; Alcock, Lisa; Aminian, Kamiar; Becker, Clemens; Bertuletti, Stefano; Bonci, Tecla; Brown, Philip; Brozgol, Marina; Buckley, Ellen; Carsin, Anne-Elie; Caruso, Marco; Caulfield, Brian; Cereatti, Andrea; Chiari, Lorenzo; Chynkiamis, Nikolaos; Ciravegna, Fabio; Del Din, Silvia; Eskofier, Björn; Evers, Jordi; Garcia Aymerich, Judith; Gazit, Eran; Hansen, Clint; Hausdorff, Jeffrey; Helbostad, Jorunn; Hiden, Hugo; Hume, Emily; Paraschiv-Ionescu, Anisoara; Ireson, Neil; Keogh, Alison; Kirk, Cameron; Kluge, Felix; Koch, Sarah; Küderle, Arne; Lanfranchi, Vitaveska; Maetzler, Walter; Micó-Amigo, M. Encarna; Mueller, Arne; Neatrour, Isabel; Niessen, Martjin; Palmerini, Luca; Pluimgraaff, Lucas; Reggi, Luca; Salis, Francesca; Schwickert, Lars; Scott, Kirsty; Sharrack, Basil; Sillen, Henrik; Singleton, David; Soltani, Abolfazi; Taraldsen, Kristin; Ullrich, Martin; Van Gelder, Linda; Vereijken, Beatrix; Vogiatzis, Ioannis; Warmerdam, Elke; Yarnall, Alison; Rochester, Lynn

VERSION 1 – REVIEW

REVIEWER	Karahanoglu, F Pfizer Inc
REVIEW RETURNED	24-May-2021

GENERAL COMMENTS	The authors present an excellent work, a detailed and focussed protocol, for accessing gait in multiple cohorts. The protocol includes all necessary tasks and measurements to serve as a robust validation. Since the work is presented as a review of protocol, it includes only high level explanation of study analyses. However, in order to get a good understanding of what this valuable data will help to achieve, I think some degree of elaboration is still needed. My comments are below: 1. There are multiple tasks and devices associated to the protocol, and those were explained in Table 4. It is clear what the reference and Tested devices associated with Mobility Tasks. However, it is not very clear what type of endpoints we can (and cannot) get from each device for those same mobility Tasks. I suggest the authors summarize the variables (e.g., endpoints) we would get from those reference and test devices. 2. Following comment #1, Table 5 summarizes DMOs, and I am assuming those are a subset of variables that will be tested? It is also not clear which devices provide those DMOs. I would suggest the authors to add a justification why they picked to test those DMOs. 3. Addition to previous comments above, Table #6, an addition of

	which devices/tasks will be tested would be informative to understand what references and tested devices will specifically be tested. 4. In Table #4, it is a bit confusing that INDIP is indicated as a Tested Device for Laboratory tasks, however, it switches to become a reference device for real world setting. Under what circumstances INDIP accuracy will be enough to be a reference system? What if the accuracy is very low based on laboratory testing compared to Stereophotogrammetry? 5. The DMOs are summarized per walking bouts, will this be the case for all tasks? In Lab and also real world settings? 6. What is the accuracy of Aeqora app to detect walking bouts and related gait parameters to be successfully used as a valid reference device? 7. What specific data quality checks will be performed to ensure data integrity? Perhaps a short discussion can be added for data management. 8. Adding to previous point, how will the missing data handled?

REVIEWER	Heesch, Kristiann Queensland University of Technology, Institute of Health & Biomedical I confirm the above for my review of this paper submitted to BMJ Open
REVIEW RETURNED	18-Jun-2021

GENERAL COMMENTS	The article describes the study protocol for a project to validate and determine acceptability of a method for monitoring mobility. It is a comprehensive evaluation plan and will be useful for future studies of mobility. The comments below are to help the authors improve the reporting of their plans. Line numbers refer to the line numbers presented on the left-side of the page. Some grammatical and punctuation errors were noted (e.g., 2nd line of Introduction: Change Organizations' to Organization's). Also noted were typos (e.g., "such as such as", line 38 Page 8) and some issues with sentence structure. Some writing uses American English and other writing uses British. Introduction The Introduction is one long paragraph and a shorter 2nd paragraph. The flow of the first paragraph is not always clear in the one paragraph. For clarity, dividing up the text into several paragraphs is recommended. For example, new paragraphs could start with "Existing mobility endpoints…"; "Poor gait,…"; "Wearable devices…"; and "In this paper…". Line 16 Page 9: The line says that the validation study aims to describe a protocol. Should this say that this paper describes the multiple-stage protocol for the technical validation study? Or perhaps the text should that 'the technical validation study aims to : a) verify …' Methods Lines 35-42 Page 12: The total sample size is 120 with 20 per cohort x 6 cohorts. How many patients are being recruited per site? For example, with five sites, is there a quota of 4 patients per cohort per site? Also, the sample size is said to be appropriate according to COSMIN guidelines, but based on the references provided for COSMIN, COSMIN standards were developed for patient-reported

	outcomes. While the current study is reporting on patient acceptance of the measure, the measure being validated is an objective measure. Therefore, the authors are asked to revisit their argument for the recruitment of a sample size of 20 people per cohort for a total of 120 to assess validity. Lines 53-54 Page 14: Please provide a sentence briefly describing the SP system and its validity/accuracy for readers not familiar with this system. This will also support the argument that it is an appropriate gold standard. Lines 3-19 Page 16: Please provide references for the validation of the Surface Test and Hallway Test in line with the text for the other tests. Line 41 page 16: Clarify if the habitual environment must be an indoors environment. It also appears to require an environment where a kitchen is available to the participant. Text later indicates that the location is indoors, but it is not explicitly stated. Line 3 Page 17: Are there data available on the accuracy/validity of the app for the purposes being used in the current study? If so, to support the use of the app, it would be useful to mention this information. Line 12-13 Page 18: What instructions do participants receive about wear time? For example, are they asked to wear it from the time they get out of bed to the time they get into bed or throughout the evening? Are they to remove it when in water? Do they separately log their wear time and reasons for removal? Are they given any instructions for when they are allowed to turn off their phone, say for charging? Line 57 Page 18: Are the 'researchers' here the same as the 'assessors' mentioned in the first sentence of the paragraph? If so, please change to 'assessors' to be clear that the questionnaires went to the assessors specifically, not to all researchers involved in the study. Last 2 paragraphs, Page 18 and first paragraph, Page 19: For completeness the authors are asked to consider including in supplementary material the following: interview questions being used in interviews with participants and with assessors and the 'bespoke questionnaire' being used with assessors. Line 19, Page 20: Please indicate what F1 means. The – looks like it is a minus sign so consider removing it, so it reads F1 score = The denominator for the formula for F1 – score is PPV + * Sensitivity. Is the + * a typo? Should it be PPV + Sensitivity? Tables and figures Tables: Please define your acronyms in a footnote so that the tables are readable on their own. In titles of tables, please spell out every word rather than use acronyms. Table 2: For the COPD inclusion criteria, I'm not quite sure which people are included based on smoking status. Is it correct that you are including all non-smokers and only current smokers or ex-smokers if they have smoked on average at least 10 pack years? So, you wouldn't be including smokers who currently or previously smoke/d less than that amount? Are non-smokers people who have never smoked? What is the definition of a healthy person or criteria used to determine whether a person is a healthy adult in addition to age? For example, are you excluding based on certain diseases? Table 5: Should the definition of the stride/step duration include step duration in addition to stride definition? Please clarify what a contact event and what an initial contact event and a final contact event are specifically (perhaps in footnote). Figure 1: I am not clear how this figure is useful. Consider cutting it.

	Figure 3: I recommend expanding the legend for the image to be clearer about what is being shown, to explain unusual words and acronyms, and to avoid acronyms where possible. Include how the image of the device links to the images of the shoes. The legend should help the reader understand the figure without going back into the article text. Figure 4: I recommend expanding the legend to be clearer about what is being shown. The legend should help the reader understand the images without going back into the article text. Figure 5: For clarity, I would suggest that the final image be given the letter (G) and then in the text clarify that the reader is to look at (G) for the 8 steps listed. For the 8 steps, add where the person starts the process. Also, add what the 3 red x's signify.

REVIEWER	Bogen, Bård Western Norway University of Applied Sciences
REVIEW RETURNED	01-Jul-2021

GENERAL COMMENTS	Thank you for this opportunity to review your validation protocol. This is a thorough description of the validation process. Further, this is important work from a large number of distinguished centers. Real-world mobility assessments will likely play a great role in future health care, and I believe that this paper is suitable for BMJ Open. I have some detailed and minor comments below: Abstract: P7, lines 9-10: This is a bold and sweeping statement, particularly as it does not seem very well backed up in the introduction. Reference #4 is relevant here, but it cannot see that "insensitive" and "resource-intensive" is addressed specifically? Perhaps the first two lines in the abstract could be modified slightly, without diminishing the importance of your work? P7, line 25: Should "IMI" be explained? P8, line 8: Are the tests extremely challenging? If so, should this be addressed a bit further in the study? Methods: P9, line 36: Should the weight and dimensions of the IMU be reported? P10, table 1: I may have missed this, but do you also plan validation of the barometer and the temperature sensor? P10, line 23: I suggest rephrasing the sentence to "These short static acquisitions can be performed simply by using etc" P 11, table 2: Would it be helpful to list abbreviations below the table? P12, line 49: "Characteristics" instead of "characterizations"? P13, line 38: In the trial registration, it says that the study will be finished in April 2021. It is of course completely understandable if things take longer than anticipated, even without a pandemic. Perhaps this is not very important. P14, table 4: Do you have any thoughts about using INDIP as a reference? What if it turns out that it does not perform well against SP? P14, line 57: On the illustration picture, there are only three markers on the device. Also, the figures do not have numbers. P15. Line 10: I suggest rephrasing to "…which will establish the specific level of accuracy for each system." P18, line 5-6: Perhaps this has been stated clearly elsewhere, but how do you handle sleep, swimming, showering etc? P18, line 35: A comma is missing between "comfort" and "perceived"

	P19, line 35: By “reliability”, do you mean ICCs here? P20, lines 7-19: Perhaps you could be even clearer with regards to sensitivity and specificity etc: Is the reference system then the gold standard and the IMU the “diagnostic” device? Should there be any cut-offs for this, to determine what is true and what is false? P20, line 22: Is relative difference something else than Intraclass correlation coefficients? P20, line 30: I assume you will use independent t-tests? And if so, the non-parametric alternative would probably be Mann-Whitney?

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

Dr. F Karahanoglu, Pfizer Inc

Comments to the Author:
The authors present an excellent work, a detailed and focussed protocol, for accessing gait in multiple cohorts. The protocol includes all necessary tasks and measurements to serve as a robust validation. Since the work is presented as a review of protocol, it includes only high level explanation of study analyses. However, in order to get a good understanding of what this valuable data will help to achieve, I think some degree of elaboration is still needed. My comments are below:

Thank you for appreciating the work and the suitability of the protocol. We genuinely appreciated your insightful and constructive comments, which we have addressed at our best.

1. There are multiple tasks and devices associated to the protocol, and those were explained in Table 4. It is clear what the reference and Tested devices associated with Mobility Tasks. However, it is not very clear what type of endpoints we can (and cannot) get from each device for those same mobility Tasks. I suggest the authors summarize the variables (e.g., endpoints) we would get from those reference and test devices.

Thank you for this excellent suggestion. We have now amended Table 5 to clarify the list of DMOs that will be specifically calculated from each device.

able5 List of digital mobility outcomes (primary and secondary DMOs) thatwill be analysed as part of the TVS.

Variables	DMOs (units)	Definition	DMO Attainable			
			DynaPort MM+	SP System	INDIP	Aeqora
Walking Bout (WB) *walking sequence containing at ast two consecutive strides of both feet. Start and end of a WB are ermined by a resting period or any her activity (non-walking period).*	Number of WBs (count)	Based on the identification of gait as an activity (yes/no) to a sample level of 0.1 s	✓	✓	✓	
	WB Start (s)	Start of WB	✓	✓	✓	
	WB End (s)	End of WB	✓	✓	✓	
	WB Duration (s)	Time between start and the end of WB	✓	✓	✓	
Stride/ Step Duration (SD) *fers to the duration (time intervals) f strides, calculated as the time in between two non-consecutive (alternate) initial contacts.*	Stride Duration (s)	Duration between two non-consecutive (alternate) initial contact events	✓	✓	✓	
	Step Duration (s)	Duration between two consecutive initial contact events	✓	✓	✓	
Cadence (CE)	Cadence (steps/minute)	Steps performed within a minute	✓	✓	✓	
Stride Length (SL)	Mean Stride Length (m)	Average stride length within a WB	✓	✓	✓	
l-World - Walking-Speed (RWS)	Gait speed (m/s)	Velocity, average stride speed within a WB	✓	✓	✓	✓
Turning	Number of Turns	Overall number of turns performed in a WB based on the identification of turns (yes/no) to a sample level of 0.1 s	✓	✓	✓	
	Turn Start (s)	Start of each turn within the WB	✓	✓	✓	
	Turn End (s)	End of each turn within the WB	✓	✓	✓	
	Turn Duration (s)	Time between the start and the end of the turns within the WB	✓	✓	✓	
	Maximal Turn Angle (deg)	Maximal angle achieved in the turn	✓	✓	✓	
Height Estimation	Elevation Change (m)	Difference between the minimal and maximal height or elevation for the complete walking bout detected for inclinewalking	✓	✓	✓	✓
Left/Right Identification	Laterality (label)	Left or Right category, indicating the foot with which the initial contact is performed	✓	✓	✓	
Additional Secondary Outcomes (SO)	Number of Final Contact Events (counts)	Correct identification of Final Contact events	✓	✓	✓	
	Final Contact Event (s)	Instant of time at which each final contact event is performed within a walking bout	✓	✓	✓	

2. Following comment #1, Table 5 summarizes DMOs, and I am assuming those are a subset of variables that will be tested? It is also not clear which devices provide those DMOs. I would suggest the authors to add a justification why they picked to test those DMOs.

The DMOs that have been selected and listed in the table are those that can be potentially calculated from a single IMU located on the lower trunk and that have been previously shown to have potential clinical relevance in the cohorts of interest. At the end of this study, we expect to have a reduced list of DMOs that will only include those that have been proven to be valid. These will then undergo a full clinical validation, aiming to prove their validity as biomarkers in regulatory drug trials. We have now added this information in the text. The amended version of table 5, following comment 1 should hopefully clarify also which device will provide them.

3. Addition to previous comments above, Table #6, an addition of which devices/tasks will be tested would be informative to understand what references and tested devices will specifically be tested.

Agreed. As per previous comments, this has now been implemented in Table 5.

4. In Table #4, it is a bit confusing that INDIP is indicated as a Tested Device for Laboratory tasks, however, it switches to become a reference device for real world setting. Under what circumstances INDIP accuracy will be enough to be a reference system? What if the accuracy is very low based on laboratory testing compared to Stereophotogrammetry?

Thanks for this very relevant comment. 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 [1-3], 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 [4,5]. Spatial parameters (i.e. stride length) are estimated from the IMUs 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 [1]. 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) [2].

In a very recent study in which the same IMUs 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 walking conditions, respectively [3]. Finally, regarding the estimation of the spatial parameters, it has been shown in the study conducted by Hannink et al. [6], that the OFDRI technique was the best performing among the double integration methods for mobile gait analysis tested in their study.

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% in the estimate of the walking speed, and 2-3% in the estimate of the stride length. 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.

We have now added the following sentence in the paper:

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 [39,42,43].The final assembled system in its fully synchronized configuration, developed to address the requirements of this study, is expected to perform equivalently and as such we can anticipate mean absolute percentage errors of 1% on the stride duration, between 2% and 3% in the estimate of the stride length.*

*Ref in this doc [1-3].

5. The DMOs are summarized per walking bouts, will this be the case for all tasks? In Lab and also real world settings?

Yes. The walking bout is our smallest window of observation for all tasks and all conditions. This information has now been added to the paper:

In all tasks and observations, continuous variables (e.g. cadence, real-walking-speed) will be summarised with descriptive statistics for the values obtained within walking bouts (mean and standard deviation).

6. What is the accuracy of Aeqora app to detect walking bouts and related gait parameters to be successfully used as a valid reference device?

Thanks for this excellent question concerning the accuracy of the Aeqora App in detecting walking bouts (WB) and measuring the contextual factors (RWS, elevation change and indoor/outdoor).

In terms of detecting WB, we do not have any Gold Standard to determine the Apps WB sensitivity. The subjects are not required to have the phone on them at all times and are instructed only to carry it with them when they leave their home. Therefore, when there is a WB detected by the McR/INDIP sensors that is not detect by Aeqora it is unclear whether this is a false negative, as the subject may not be carrying the phone. It is possible to examine at the Apps WB specificity, where the phone detects activity when no WB is detected by the McR/INDIP sensors. However, having activity detection sensitivity which leads to such false positives is not an issue in the Mobilise-D use case, as data from the false positive activity will not be used. Ideally, the GPS device would be also be continuously carried/active (and potentially integrated into one movement monitoring device), however there are the issues of battery life, size and privacy concerns to be addressed.

There have been a few empirical studies that measure GPS derived walking speed accuracy, however these almost entirely perform the measurements on a single fixed journey. However, GPS accuracy largely depends on the WB location and length. Different locations influence the degree of GPS noise caused by signal occlusion and reflection. However, confidence in the speed calculation will increase for longer, continuous walking bouts, as these are necessary to provide enough GPS points to smooth out individual GPS point errors, and with consistent direction of travel between

points. Basically, when the subject is purposefully walking from point A to B, i.e. for exercise or transport. Elevation accuracy has the added issue of vertical measurement noise, which depends on the elevation resource, in the UK Ordinance Survey Terrain5 provides an average RMSE of around 2 metres, while the worldwide SRTM data used for other countries is around 5-9 m, although these are measurement of absolute altitude error rather than local errors required for calculating elevation change.

Identification of indoor/outdoor walking is seen as one of the key contextual factors in influencing walking behaviour, it is possible to remotely observe the indoor/outdoor classifications by examining the journeys on a map. This method was used to perform error analysis to determine the cause of observed errors, two main causes where identified: missing data (actual buildings are not in the data) leading to falsely identifying outdoor journeys, and journeys on the boundaries between indoor and outdoor. A more systematic error analysis will be undertaken to determine the degree to which the calculate probability of indoor/outdoor agrees with the manual assessment.

Last but not least, the use of the App in conjunction with the Beacon will allow us to isolate the subset of walking bout in which the participant was certainly using a walking aid.

The following sentence has now been included in the paper:

The above contextual factors and the use of walking aids will be included in the analyses to determine the extent to which they affect variation in the DMOs, although the degree of correlation will be adversely affected by the issues in accurately measuring context that are associated with missing data and GPS accuracy.

7. What specific data quality checks will be performed to ensure data integrity? Perhaps a short discussion can be added for data management.

All participants within this study are allocated a participant ID and all data captured is associated with this ID. Each site maintains a study key which matches the participant ID to the participant's name. This study key is stored securely at site and will be destroyed at the end of the project (or retained longer as per local guidelines). We are capturing data from a variety of sources; electronic patient reported outcomes (ePRO), electronic clinician reported outcomes (eClinRO), and mobility data via the use of motion capture systems (including IMUs). All of this data is ingested into our central staging platform, e-Science Central (e-SC) which is hosted on AWS. We then use an extract-transfer-load (ETL) process to extract the relevant data and load into a normalised Data Warehouse (DW) for query and analysis. There is no Personal Identifiable Information (PII) uploaded to e-SC or stored in the DW.

We use both web-based forms on e-SC and an application from ERT to capture the electronic clinical outcome assessments (eCOAs). The e-SC forms written in JSON (Java Script Object Notation)

provide storage of event data, and support for data validation via the use of JSON schema (a standard used to describe data structures and to enforce various validation rules). This schema validates the data uploaded and is also used to automatically create data entry forms. This use of schema and automatically generated forms has allowed basic data entry and verification to be performed rapidly on all data that is gathered directly into the e-SC platform. ERT provided a tablet-based eCOA application with a web-based backup version. This provided assessors with mobility, flexibility and minimised risk of data entry. Both e-SC and ERT systems employ error-handling at source which alert the assessors of incorrect data entry (e.g. min/max boundaries, required/optional fields). Data captured at source on paper was copied, signed, and scanned, then uploaded to e-SC as a certified copy.

The motion capture data are also transferred to e-SC and stored in an unmodified form and are associated with the relevant patient automatically. These data are either uploaded directly to e-SC via the e-SC portal or transferred via Application Programming Interface (API). Algorithms being developed and benchmarked will be used to process these files and extract Digital Mobility Outcomes (DMOs) which will be loaded into the DW using the ETL process outlined earlier.

On completion of the data collection, each participant folder on e-SC can contain a set of observations (structured JSON data items as described above) and collections of arbitrary data files (for example, movement data files uploaded by McRoberts). Both e-SC and ERT provide reporting interfaces and the Data Management Team run weekly data quality and compliance reports. Any protocol deviations are logged and presented to the study sponsor, any missing data is queried with the site and actioned accordingly, and any incorrect data entries are corrected via a Data Change Request (DCR) process. DCRs are logged and regularly reviewed, with the Study Management Group (SMG).

This information has now been summarised and added in the data management section as it follows:

In particular, we will use both web-based forms on e-SC and an application from ERT (partner in the project) to capture the electronic clinical outcome assessments (eCOAs). The e-SC forms provide storage of event data, and support for data validation and basic data entry and verification. Both e-SC and ERT systems employ error-handling at source which alert the assessors of incorrect data entry (e.g. min/max boundaries, required/optional fields). Data captured at source on paper will be copied, signed, and scanned, then uploaded to e-SC as a certified copy. The motion capture data will also be transferred to e-SC and stored in an unmodified form. These data will be either uploaded directly to e-SC via the e-SC portal or transferred via an Application Programming Interface (API). The algorithms being developed and benchmarked will be used to process these files and extract and store the DMOs.

8. Adding to previous point, how will the missing data handled?

Participants' data that may be missing or unavailable for specific assessments/ tests (e.g. test within the Laboratory assessment due to failure of the recording device or other technical issues) will be identified and reported/described. Assuming that data are missing completely at random, a complete case approach will be used [7]. 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.

The following sentence has now been included in the paper:

If participants missed data from one assessment (e.g., one of the tasks in the laboratory) or observation (e.g. 2.5hs), their remaining available data will still be included in the analyses. Assuming that data are missing completely at random, a complete case approach will be used to handle missing data [60].*

*Ref in this doc [7].

Reviewer: 2
Dr. Kristiann Heesch, Queensland University of Technology

Comments to the Author:
The article describes the study protocol for a project to validate and determine acceptability of a method for monitoring mobility. It is a comprehensive evaluation plan and will be useful for future studies of mobility. The comments below are to help the authors improve the reporting of their plans. Line numbers refer to the line numbers presented on the left-side of the page.

Thank you for appreciating how this study will be useful for future research in the field. Your comments were extremely useful and well received and we have now implemented them as described in the following answers.

Some grammatical and punctuation errors were noted (e.g., 2nd line of Introduction: Change Organizations' to Organization's). Also noted were typos (e.g., "such as such as", line 38 Page 8) and some issues with sentence structure. Some writing uses American English and other writing uses British.

Thank you. The typos have now been amended and language used has been reviewed to ensure consistency with British English.

Introduction
The Introduction is one long paragraph and a shorter 2nd paragraph. The flow of the first paragraph is not always clear in the one paragraph. For clarity, dividing up the text into several paragraphs is recommended. For example, new paragraphs could start with "Existing mobility endpoints…"; "Poor gait,…"; "Wearable devices…"; and "In this paper…".
Line 16 Page 9: The line says that the validation study aims to describe a protocol. Should this say that this paper describes the multiple-stage protocol for the technical validation study? Or perhaps the text should that 'the technical validation study aims to : a) verify …'

Thank you for the suggestion. We have now amended the introduction following your recommendation.

Methods

Lines 35-42 Page 12: The total sample size is 120 with 20 per cohort x 6 cohorts. How many patients are being recruited per site? For example, with five sites, is there a quota of 4 patients per cohort per site? Also, the sample size is said to be appropriate according to COSMIN guidelines, but based on the references provided for COSMIN, COSMIN standards were developed for patient-reported outcomes. While the current study is reporting on patient acceptance of the measure, the measure being validated is an objective measure. Therefore, the authors are asked to revisit their argument for the recruitment of a sample size of 20 people per cohort for a total of 120 to assess validity.

Thank you for this very relevant comment. We do agree with the reviewer that the COSMIN guidelines are not entirely appropriate for this study but given the novelty of the data and the lack of previous real world validation study, we decided to consider them as a suitable initial reference. We had indeed planned from the beginning to revise our sample size calculation using data from the initial observations to establish the validity of this assumption. Since the DMOs are measured at walking bout level and not at patient level, this analysis will be based on the effective number of walking bouts observed during the 2.5 hours. While we do not believe it's relevant to publish this preliminary data as part of this protocol paper, we are happy to share with the reviewer the information that this analysis has now been performed and its results confirmed that the adopted sample size is suitable for the purpose.

We have now added the information about the plan for the sample size revision in the paper:

Given the novelty of the data, rather than on a power calculation the sample size of 120 has been initially defined according to Consensus-based Standards for the selection of health Measurement Instruments guidelines for measurement properties (COSMIN [21]). This sample size, however, will be refined after 50% of the data collection. Given the DMOs are measured at walking bout level and not at patient level, in this analysis we will use the effective number of walking bouts observed during the 2.5 hours to perform the power calculation. We will based this analysis on a desired ICC coefficient ≥0.7, with Alpha=0.05 and Beta=0.9, and an aimed confidence interval of 0.1. Based on this review, more participants may be recruited.*

*Ref in this doc [8].

Lines 53-54 Page 14: Please provide a sentence briefly describing the SP system and its validity/accuracy for readers not familiar with this system. This will also support the argument that it is an appropriate gold standard.

Thanks for the suggestion. The following paragraph has been added in the methods section.

SP systems provide a measurement of the instantaneous position of points in a 3D measurement volume, by means of a set of cameras, each of which can capture the 2D trajectories of markers that are attached to the object of interest. The trajectory reconstructions are affected by systematic and random instrumental errors, normally minimised to a few millimetres via ad hoc calibration procedures and filtering and smoothing techniques [30].*

*Ref in this doc [9].

Lines 3-19 Page 16: Please provide references for the validation of the Surface Test and Hallway Test in line with the text for the other tests.

While the other tasks have been inspired by standard tests for mobility capacity assessment, these two tests have been specifically designed for the purpose of simulating the effect that different real-world surfaces (e.g. carpet) and the presence of obstacles could have on the accuracy if the DMOs estimates. As such, they have not been validated before.

The following information has now been included in the text:

Two novel additional tests were also included to simulate confounding factors that could be encountered in the real word:

Line 41 page 16: Clarify if the habitual environment must be an indoors environment. It also appears to require an environment where a kitchen is available to the participant. Text later indicates that the location is indoors, but it is not explicitly stated.

The definition of the real-world environment is not prescribed. We realised that the list of suggested tasks might have been misleading. We have hence now added this information more specifically in the paper:

[…] It will be performed in a habitual environment (home/work/community/outdoor) chosen by the participants, without specific restrictions.

[..] To capture the largest possible range of activities during this assessment, participants will be guided by the following list of activities to be included, if relevant for their chosen environment:

Line 3 Page 17: Are there data available on the accuracy/validity of the app for the purposes being used in the current study? If so, to support the use of the app, it would be useful to mention this information.

As per answer to Reviewer1 (please refer to that for more details), the accuracy of the App is highly dependent on the context of observation and as such difficult to establish a-priori. We do not have any Gold Standard to determine the Apps sensitivity in Walking Bout detection, since the subjects are not required to have the phone on them at all times and are instructed only to carry it with them when they leave their home. Nonetheless, it is possible to examine at the Apps WB specificity, since having activity detection sensitivity which leads to such false positives is not an issue in the Mobilise-D use case, as data from the false positive activity will not be used.

Identification of indoor/outdoor walking is a key contextual factor in influencing walking behaviour, and these can be accurately detected. Last but not least, the use of the App in conjunction with the Beacon will allow us to accurately identify and isolate the subset of walking bout in which the participant was certainly using a walking aid, and quantify the effect of this factor on RWS as measured from the single device.

The following sentence has now been included in the paper:

The above contextual factors and the use of walking aids will be included in the analyses to determine the extent to which they affect variation in the DMOs, although the degree of correlation will be adversely affected by the issues in accurately measuring context that are associated with missing data and GPS accuracy.

Line 12-13 Page 18: What instructions do participants receive about wear time? For example, are they asked to wear it from the time they get out of bed to the time they get into bed or throughout the evening? Are they to remove it when in water? Do they separately log their wear time and reasons for removal? Are they given any instructions for when they are allowed to turn off their phone, say for charging?

The participants are asked to wear and carry the Dynaport at all times for the full 7 day period (including at night). As the Dynaport is not waterproof, they are instructed to remove it for showering, bathing, using a sauna and swimming and reattach it afterwards. Participants are asked to keep the mobile phone switched on at all times and charged (they were provided with a wireless rapid charger to facilitate this operation). They are instructed to carry it with them whenever possible when inside the house and all the times when leaving their homes.

The following sentence has now been added in the paper:

The participants will be asked to wear the DynaPort MM+, and to carry a mobile phone equipped with the Aeqora App. Bluetooth beacons will also be used to track the use of walking aids. The participants will wear the Dynaport MM+ at all times (including at night). As this device is not waterproof, they will be instructed to remove it for showering, bathing, using a sauna and swimming

and reattach it afterwards. They will be asked to keep the mobile phone charged, switched on at all times and to carry it with them whenever possible, especially when leaving the house.

Line 57 Page 18: Are the 'researchers' here the same as the 'assessors' mentioned in the first sentence of the paragraph? If so, please change to 'assessors' to be clear that the questionnaires went to the assessors specifically, not to all researchers involved in the study.

Yes, they are the same. We have now amended the text as suggested.

Last 2 paragraphs, Page 18 and first paragraph, Page 19: For completeness the authors are asked to consider including in supplementary material the following: interview questions being used in interviews with participants and with assessors and the 'bespoke questionnaire' being used w

ith assessors.

The requested information has now been included in the supplementary material.

Line 19, Page 20: Please indicate what F1 means. The – looks like it is a minus sign so consider removing it, so it reads F1 score = The denominator for the formula for F1 – score is PPV + * Sensitivity. Is the + * a typo? Should it be PPV + Sensitivity?

Thanks for noticing the typo and for the editing suggestion. The formula has now been corrected:

$$F1\ score\ = 2 * \frac{Positive\ Predicted\ Value * Sensitivity}{Positive\ Predicted\ Value + Sensitivity}.$$

Tables and figures

Tables: Please define your acronyms in a footnote so that the tables are readable on their own. In titles of tables, please spell out every word rather than use acronyms.

Thanks for the suggestion - this has now been implemented.

Table 2: For the COPD inclusion criteria, I'm not quite sure which people are included based on smoking status. Is it correct that you are including all non-smokers and only current smokers or ex-smokers if they have smoked on average at least 10 pack years? So, you wouldn't be including smokers who currently or previously smoke/d less than that amount? Are non-smokers people who have never smoked? What is the definition of a healthy person or criteria used to determine whether a person is a healthy adult in addition to age? For example, are you excluding based on certain diseases?

We thank the reviewer for having spotted a mistake in Table 2, which indeed generated confusion. In the study we include either current or ex-smokers with a primary diagnosis of COPD (please see revised Table 2 below). Ex-smokers should have a smoking history equivalent to at least 10 pack-years. Patients that have never smoked or have smoked less than 10 pack-years are excluded. The reliance on pack-years is appropriate for enrolment in clinical trials of COPD, which typically rely on a minimum of a 10 pack-year history [10]. This is standard procedure in clinical trials in order to exclude patients who have COPD for reasons other than smoking (i.e. Alpha-1 antitrypsin deficiency that is an inherited disorder that may cause lung disease). Furthermore, we are including current smokers with a primary diagnosis of COPD independently of the number of pack-years. Non-smokers are people who have never smoked. For the COPD cohort we include people with a primary diagnosis of COPD (post-bronchodilator forced expiratory volume in the first second (FEV1) to forced vital capacity (FVC) ratio <0.70. People with FEV1/FVC > 0.70 are excluded because a normal ratio for a healthy person (i.e.: without respiratory disease) is considered to be 70-80% in adults. Furthermore, respiratory patients with a primary respiratory disease other than COPD (e.g. asthma) are excluded.

Reference

Table 2 has now been amended consistently;

~~*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)*

Table 5: Should the definition of the stride/step duration include step duration in addition to stride definition? Please clarify what a contact event and what an initial contact event and a final contact event are specifically (perhaps in footnote).

Figure 1: I am not clear how this figure is useful. Consider cutting it.

We believe that this figure summarises the complexity of the study and the specific device agnostic and multi-faceted approach that we have chosen to follow. As such, we have decided to keep it.

Figure 3: I recommend expanding the legend for the image to be clearer about what is being shown, to explain unusual words and acronyms, and to avoid acronyms where possible. Include how the image of the device links to the images of the shoes. The legend should help the reader understand the figure without going back into the article text.

This has now been expanded as:

Figure 3 – Illustration of the adopted marker set configuration. Markers were located on the right (RHEEL) and left (LHEEL) heels, toes (RTOE, LTOE) and on the INDIP units located on the right and left foot (RINDIP, LINDIP). Two additional reference markers were asymmetrically attached to the side of the foot to favour automatic recognition (RREF. LREF). Four additional markers were located on the DynaPort MM+ sensor (DYNAY, DYNAO, DYNAX, DYNAREF).

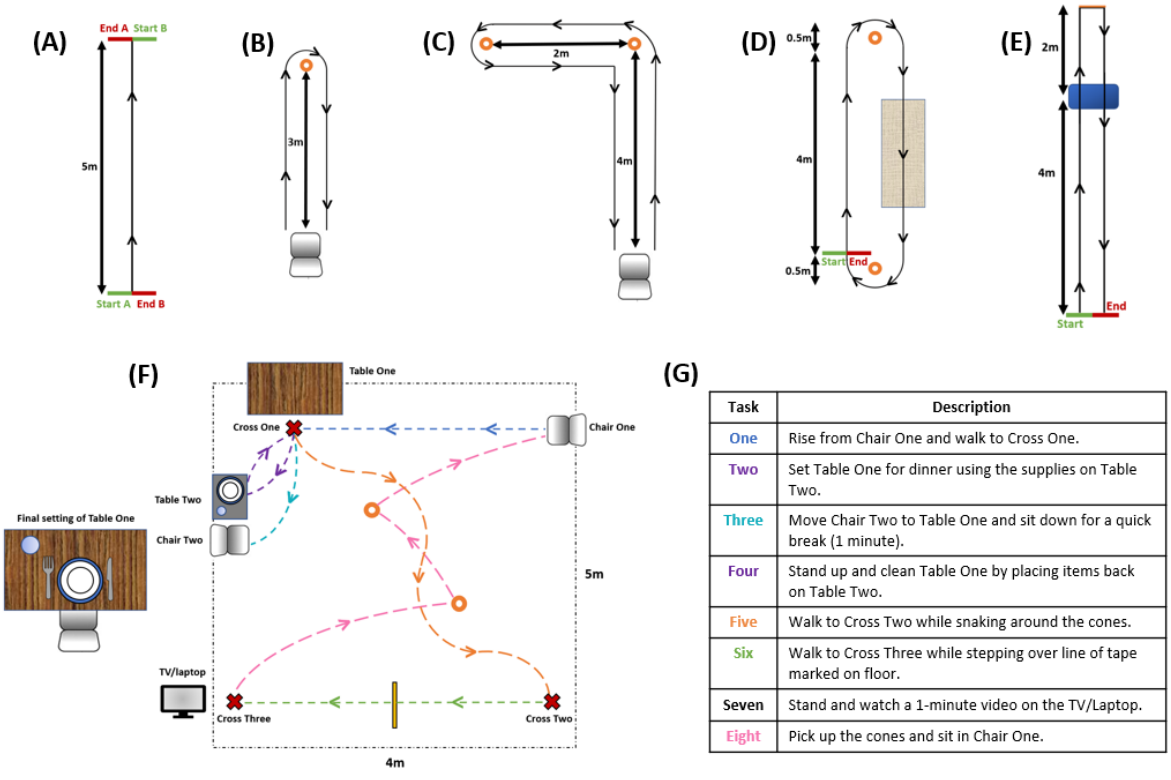
Figure 4: I recommend expanding the legend to be clearer about what is being shown. The legend should help the reader understand the images without going back into the article text.

Thanks for the suggestion. We have now expanded the legend:

Figure 4 –Different components of the INDIP system. The figure on the left shows the pressure insoles and the connectors that link them to the distance sensors and the inertial modules. The picture on the right shows how the same system is then attached to the participant's foot and leg.

Figure 5: For clarity, I would suggest that the final image be given the letter (G) and then in the text clarify that the reader is to look at (G) for the 8 steps listed. For the 8 steps, add where the person starts the process. Also, add what the 3 red x's signify.

Amended as suggested:



Task	Description
One	Rise from Chair One and walk to Cross One.
Two	Set Table One for dinner using the supplies on Table Two.
Three	Move Chair Two to Table One and sit down for a quick break (1 minute).
Four	Stand up and clean Table One by placing items back on Table Two.
Five	Walk to Cross Two while snaking around the cones.
Six	Walk to Cross Three while stepping over line of tape marked on floor.
Seven	Stand and watch a 1-minute video on the TV/Laptop.
Eight	Pick up the cones and sit in Chair One.

Figure 5 – Diagrams of the selected tasks: (A) Straight Walking Test, (B) Timed Up and Go, (C) L-Test, (D) Surface Test, (E) Hallway Test, F) Schematic of the Daily living activities, G) Description of the eight tasks performed during the Daily Living activities.

Reviewer: 3
Dr. Bård Bogen, Western Norway University of Applied Sciences

Comments to the Author:
Thank you for this opportunity to review your validation protocol. This is a thorough description of the validation process. Further, this is important work from a large number of distinguished centers. Real-world mobility assessments will likely play a great role in future health care, and I believe that this paper is suitable for BMJ Open. I have some detailed and minor comments below:

Thank you for appreciating the importance of this study and for the very useful and constructive comments. We have now implemented them as indicated below.

Abstract:
P7, lines 9-10: This is a bold and sweeping statement, particularly as it does not seem very well backed up in the introduction. Reference #4 is relevant here, but it cannot see that "insensitive" and "resource-intensive" is addressed specifically? Perhaps the first two lines in the abstract could be modified slightly, without diminishing the importance of your work?

We agree with the reviewer that in the attempt of being concise we might have too blunt here. The sentence has been toned down as following:

Existing mobility endpoints based on functional performance, physical assessments and patient self-reporting often lack of sensitivity, limiting their utility in clinical practice.

P7, line 25: Should "IMI" be explained?

Thank you for the suggestion. We have now spelled out in the paper that IMI stands for Innovative Medicine Initiative.

P8, line 8: Are the tests extremely challenging? If so, should this be addressed a bit further in the study?

The tests can indeed become challenging for highly impaired participants. The protocol does indeed include the possibility for the patients to take rests whenever needed and they are clearly instructed about the fact that they can stop at any time and do not necessarily have to complete the entire protocol if it is too daunting for them.

The following sentence has now been added in the text:

Patients will be given regular opportunities for rest periods and will be asked to communicate if they require any additional breaks or would like to stop the assessment at any point. Use of arm rests for the TUG, L-Test and SDA, as well as handrails for the hallway test are permitted when needed.

Methods:
P9, line 36: Should the weight and dimensions of the IMU be reported?

Thank you for the suggestion. We have now included this information in the text:

Table 1 shows the sensing characteristics of the device used in this study, the Dynaport MM+ (dimensions: 106.6 x 58 x 11.5mm, size: 55 grams).

P10, table 1: I may have missed this, but do you also plan validation of the barometer and the temperature sensor?

We had originally reported this information for contextualisation, but these sensors will not be tested. We have now removed the two lines from the table to avoid confusion.

P10, line 23: I suggest rephrasing the sentence to "These short static acquisitions can be performed simply by using etc"

Changed as suggested.

P 11, table 2: Would it be helpful to list abbreviations below the table?

Thank you for the suggestion - These have now been added.

P12, line 49: "Characteristics" instead of "characterizations"?

Changed as suggested.

P13, line 38: In the trial registration, it says that the study will be finished in April 2021. It is of course completely understandable if things take longer than anticipated, even without a pandemic. Perhaps this is not very important.

The reviewer is absolutely correct. When the study was registered, we were not expecting to have to deal with the pandemic…. In the current version of the paper, we anticipated finishing recruitment in July instead. Unfortunately, we have now had to postpone this to September. The paper has been amended to include this further change and explain the discrepancy.

P14, table 4: Do you have any thoughts about using INDIP as a reference? What if it turns out that it does not perform well against SP?

As per answer to Reviewer 1, we are confident that the INDIP will perform adequately, since its individual components and the associated algorithms for the estimates of the DMOs have already been extensively validated in previous studies on various healthy and pathological cohorts [1-3] The final assembled system in its fully synchronized configuration, developed to address the specific project' s needs, is expected to perform equivalently and as such we can anticipate mean absolute percentage errors of 1% on the stride duration, 2% in the estimate of the walking speed, and 2-3% in the estimate of the stride length.

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.

We have now added the following sentence in the paper:

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 [39,42,43]. The final assembled system in its fully synchronized configuration, developed to address the requirements of this study, is expected to perform equivalently and as such we can anticipate mean absolute percentage errors of 1% on the stride duration, between 2% and 3% in the estimate of the stride length.*

*Ref in this doc [1-3].

P14, line 57: On the illustration picture, there are only three markers on the device. Also, the figures do not have numbers.

Thank you for the suggestion - These have now been added.

P15. Line 10: I suggest rephrasing to "…which will establish the specific level of accuracy for each system."

Changed as suggested.

P18, line 5-6: Perhaps this has been stated clearly elsewhere, but how do you handle sleep, swimming, showering etc?

This is a very relevant point, which we have now clarified including the following sentence:

The participants will wear the Dynaport MM+ at all times (including at night). As this device is not waterproof, they will be instructed to remove it for showering, bathing, using a sauna and swimming and reattach it afterwards.

P18, line 35: A comma is missing between "comfort" and "perceived"

Thanks for spotting this! Included.

P19, line 35: By "reliability", do you mean ICCs here?

Thank you for the comment. We have clarified and changed reliability to "concurrent validity" which will indeed be quantified using the Intraclass Correlation Coefficient ICC (2,1). We have amended the text reflecting this.

The data analysis will determine criterion validity (including selected performance metrics and criterion (concurrent) validity metrics

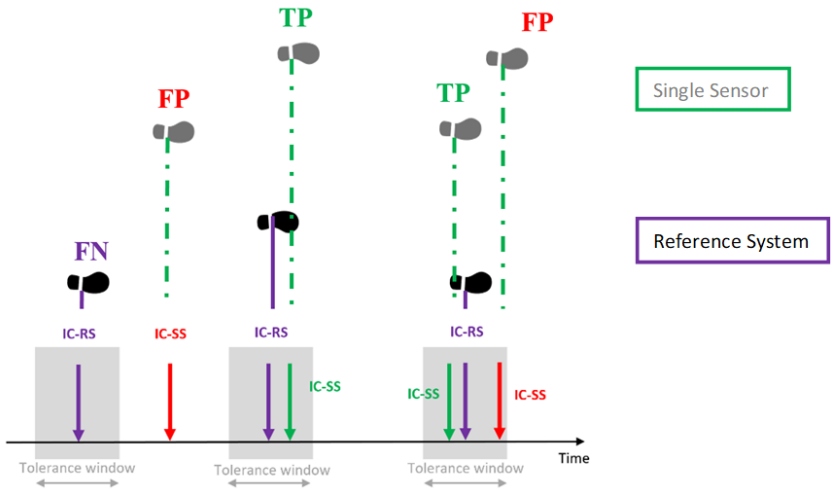
P20, lines 7-19: Perhaps you could be even clearer with regards to sensitivity and specificity etc: Is the reference system then the gold standard and the IMU the "diagnostic" device?

Should there be any cut-offs for this, to determine what is true and what is false?

To evaluate performance metrics (sensitivity, specificity, accuracy, etc.) of events with a start, duration and end (e.g. walking bouts) the reference system will be the gold standard and the IMU will be compared against it. In this case, each sample of the IMU signal will be compared to the corresponding sample of the reference system signal, in order to identify TP, TN, FP and FN.

Example of how each sample (0.01 seconds) is identified as TP, TN, FP and FN and compared between the IMU single sensor (SS) and reference system (RS) for the identification/detection of walking bouts. 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.

TP, FP and FN will be evaluated using a cut-off tolerance window defined as a fixed interval of 0.5 seconds [11]. The tolerance window will be centred on each event detected by the reference system. Events detected by the IMU within the tolerance window will be identified as TP, while outside the tolerance window as a FP, etc.



This figure presents how False Negative (FN), False Positive (FP) and True Positive (TP) events are identified based on the comparison of initial contact (IC) events detected by the IMU single sensor (IC-SS in red) versus the ones detected by the reference system (IC-RS in green). FN, FP and TP are defined with respect to the selected temporal tolerance window, in grey.

We have now modified thins section sentence as it follows:

Using the gold standard as a reference, True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN) will be identified for the DMOs identified from the single device using a cut-off tolerance window defined as a fixed interval of 0.5 seconds [59] and centred on each event detected by the reference system. The following performance metrics will then be calculated: […]*

*Ref in this doc [11].

P20, line 22: Is relative difference something else than Intraclass correlation coefficients?

Thanks for asking this question, which made us realise that we were not clear in the text. Relative difference is indeed a different concept than ICC.

Relative difference (relative error) is defined as follows:

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

Absolute difference (absolute error) is defined as follows:

$$Absolute\ Error\ (DMO) = \ |DMO\ estimated\ by\ IMU - DMO\ estimated\ by\ Reference\ System|$$

We have now included this information in the paper.

P20, line 30: I assume you will use independent t-tests? And if so, the non-parametric alternative would probably be Mann-Whitney?

We will use paired t-tests, as we want to quantify the difference between two measures within the same subject, and the Wilcoxon signed-rank test as its non-parametric equivalent.

We have now modified the sentence as it follows:

Statistically significant differences between the DMOs quantified by the IMU and those by the reference system, parametric (paired t-test) or non-parametric (Wilcoxon signed-rank test) tests will be performed depending on the normality of the distribution of the DMOs. Data distribution will be visually inspected with histograms, and normality tested with the Shapiro Wilk test.

References

1 Trojaniello D, Cereatti A, Pelosin E, *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. Journal of neuroengineering and rehabilitation. 2014 Dec;11(1):1-2. doi:10.1186/1743-0003-11-152

2 Bertoli M, Cereatti A, Trojaniello D, *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 Dec;17(1):1-4. doi:10.1186/s12938-018-0488-2

3 Rossanigo R, Caruso M, Salis F, *et al*. 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 Jun 23 (pp. 1-6). IEEE., doi: 10.1109/MeMeA52024.2021.9478604.

4 Hausdorff JM, Ladin Z, Wei JY. Footswitch system for measurement of the temporal parameters of gait. *Journal of biomechanics*. 1995 Mar 1;28(3):347-51. doi:10.1016/0021-9290(94)00074-E

5 Catalfamo P, Moser D, Ghoussayni S, Ewins D. Detection of gait events using an F-Scan in-shoe pressure measurement system. Gait & posture. 2008 Oct 1;28(3):420-6. doi:10.1016/j.gaitpost.2008.01.019

6 Hannink J, Ollenschläger M, Kluge F, *et al.* Benchmarking foot trajectory estimation methods for mobile gait analysis. *Sensors.* 2017 Sep;17(9):1940. doi:10.3390/s1709 1940.

7 Kang H. The prevention and handling of the missing data. *Korean journal of anesthesiology.* 2013 May;64(5):402. doi:10.4097/kjae.2013.64.5.402

8 COSMIN. https://www.cosmin.nl/ (accessed 18 Feb 2021).

9 Chiari L, Della Croce U, Leardini A, *et al.* Human movement analysis using stereophotogrammetry: Part 2: Instrumental errors. *Gait & posture* 2005; **21;**197-211.

doi:10.1016/j.gaitpost.2004.04.004

10 Pleasants RA, Rivera MP, Tilley SL, *et al*. Both duration and pack-years of tobacco smoking should be used for clinical practice and research. *Annals of the American Thoracic Society*. 2020 Jul;17(7):804-6.. DOI: 10.1513/AnnalsATS.202002-133VP

11 Tietsch M, Muaremi A, Clay I, *et al*. Robust Step Detection from Different Waist-Worn Sensor Positions: Implications for Clinical Studies. *Digital biomarkers.* 2020;4(1):50-8. doi:10.1159/000511611

VERSION 2 – REVIEW

REVIEWER	Karahanoglu, F Pfizer Inc
REVIEW RETURNED	15-Sep-2021

GENERAL COMMENTS	I have no further comments, thank you very much for this great work.

REVIEWER	Heesch, Kristiann Queensland University of Technology, Institute of Health & Biomedical
REVIEW RETURNED	13-Aug-2021

GENERAL COMMENTS	The authors have done very well at addressing my queries. Thank you. I just have now three points to clarify based on the latest version. Line numbers refer to the line numbers presented on the left-side of the page in the version of the paper in which changes have been highlighted in yellow. Page 57 line 37: Should 'July 2020' be 'July 2021'? Page 68 lines 22-24: The authors are asked to clarify whether participants with missing data are included. One sentence indicates that participants with missing data from one assessment will be included in analyses, but the next sentence says that a complete case approach will be used, which suggests to me that participants with missing data are not included. Table 5, Page 69 line 20: Should the first column say 'of strides/steps, calculated as…'?

REVIEWER	Bogen, Bård Western Norway University of Applied Sciences
REVIEW RETURNED	09-Aug-2021

GENERAL COMMENTS	Thank you for your replies to my comments.

VERSION 2 – AUTHOR RESPONSE

Thanks you for the few additional suggestions from Reviewer 2.
Please our answers below:

Reviewer: 2
Dr. Kristiann Heesch, Queensland University of Technology

Comments to the Author:
The authors have done very well at addressing my queries. Thank you. I just have now three points to clarify based on the latest version. Line numbers refer to the line numbers presented on the left-side of the page in the version of the paper in which changes have been highlighted in yellow.

1) Page 57 line 37: Should ‘July 2020’ be ‘July 2021’?
Thanks for double checking. The study did indeed start in July 2020

2) Page 68 lines 22-24: The authors are asked to clarify whether participants with missing data are included. One sentence indicates that participants with missing data from one assessment will be included in analyses, but the next sentence says that a complete case approach will be used, which suggests to me that participants with missing data are not included.

We agree with the Reviewer that the paragraph on treatment of missing data is unclear and thank them for noticing. We need to distinguish between missing (skipping) a context and missing data. As the study includes several contexts of assessment (ie. laboratory, real-world 2.5 hours, and real-world 7 days), a participant may skip one of these assessments and yet would be included for the rest – since the collected data in remaining contexts will allow testing validity. For each of the contexts, we will use a complete case approach assuming that missing data (e.g., in a given variable/patient) is missing at random.

We have modified the corresponding paragraph as follows:
If participants miss data from do not participate in one assessment (e.g., one of the tasks in the laboratory) or observation (e.g., 2.5hs), their remaining available data corresponding to remaining assessments/observations will still be included in the analyses. Within each of the contexts/assessments/observations, and assuming that data are missing completely at random, a complete case approach will be used to handle missing data [61].

3) Table 5, Page 69 line 20: Should the first column say ‘of strides/steps, calculated as…’?
Thanks for spotting this. The table has now been amended as suggested.