Digit Biomark 2020;4:26–43
DOI: 10.1159/000506860
Received: November 11, 2019
Accepted: February 28, 2020
Published online: April 30, 2020

Research Reports – Research Article

Developing Smartphone-Based Objective Assessments of Physical Function in Rheumatoid Arthritis Patients: The PARADE Study

Valentin Hamy[a] Luis Garcia-Gancedo[a] Andrew Pollard[b] Anniek Myatt[b] Jingshu Liu[c] Andrew Howland[c] Philip Beineke[c] Emilia Quattrocchi[d] Rachel Williams[e] Michelle Crouthamel[f]

[a]Digital Biomarkers, R&D Development, GlaxoSmithKline, Stevenage, UK; [b]Tessella, Altran World Class Center for Analytics, Stevenage, UK; [c]Data Science, Medidata Solutions, New York City, NY, USA; [d]Discovery Medicine, R&D Development, GlaxoSmithKline, Stockley Park, UK; [e]Epidemiology, R&D Development, GlaxoSmithKline, Collegeville, PA, USA; [f]Digital Clinical Trials, R&D Development, GlaxoSmithKline, Collegeville, PA, USA

Keywords
Machine learning · iPhone sensor · Rheumatoid arthritis · Range of motion · Gait

Abstract
Background: Digital biomarkers that measure physical activity and mobility are of great interest in the assessment of chronic diseases such as rheumatoid arthritis, as it provides insights on patients' quality of life that can be reliably compared across a whole population. ***Objective:*** To investigate the feasibility of analyzing iPhone sensor data collected remotely by means of a mobile software application in order to derive meaningful information on functional ability in rheumatoid arthritis patients. ***Methods:*** Two objective, active tasks were made available to the study participants: a wrist joint motion test and a walk test, both performed remotely and without any medical supervision. During these tasks, gyroscope and accelerometer time-series data were captured. Processing schemes were developed using machine learning techniques such as logistic regression as well as explicitly programmed algorithms to assess data quality in both tasks. Motion-specific features including wrist joint range of motion (ROM) in flexion-extension (for the wrist motion test) and gait parameters (for the walk test) were extracted from high quality data and compared with subjective pain and mobility parameters, separately captured via the application. ***Results:*** Out of 646 wrist joint motion samples collected, 289 (45%) were high quality. Data collected for the walk test included 2,583 samples (through 867 executions of the test) from which 651 (25%) were high quality. Further analysis of high-quality data highlighted links between reduced mobility and in-

Valentin Hamy
Digital Biomarkers
R&D Development, GlaxoSmithKline
Gunnels Wood Road, Stevenage SG1 2NY, Hertfordshire (UK)
valentin.x.hamy@gsk.com

Digit Biomark 2020;4:26–43
DOI: 10.1159/000506860

creased symptom severity. ANOVA testing showed statistically significant differences in wrist joint ROM between groups with light-moderate (220 participants) versus severe (36 participants) wrist pain ($p < 0.001$) as well as in average step times between groups with slight versus moderate problems walking about ($p < 0.03$). ***Conclusion:*** These findings demonstrate the potential to capture and quantify meaningful objective clinical information remotely using iPhone sensors and represent an early step towards the development of patient-centric digital endpoints for clinical trials in rheumatoid arthritis.

Introduction

Rheumatoid arthritis (RA) is a chronic autoimmune disease mainly characterized by persistent inflammation of large and small joints, with chronic pain, joint swelling, and deformities leading to disability and impaired ability to perform daily tasks [1–5]. RA is one of the most prevalent chronic inflammatory diseases [6]. The individual burden results from the musculoskeletal impairment, with progressive decline in physical function and quality of life. While existing measures of treatment efficacy, such as American College of Rheumatology 20 [7] or Disease Activity Score 28 [8] provide useful efficacy endpoints in clinical trials, RA symptoms fluctuate throughout the day and are variable within and between patients. As a result, the infrequent clinic visits involved in clinical trials present an incomplete view of disease activity. In addition, the evaluation of joint swelling and joint pain by a study physician introduces a high degree of subjectivity, limiting the ability to draw strong conclusions from the data. Hence, it is imperative to develop objective tools to measure RA disease progression that can be used at home with a frequency that captures the real-time disease variation, thereby gathering valuable insights into the effectiveness of new medicines as well as aspects of patients' daily lives which remain overlooked.

Activity- and mobility-specific tasks have been successfully used in several studies to collect "real-world" data, providing insights into the effect of therapies on the daily lives of patients suffering from RA as well as other conditions. In particular, iPhone-embedded sensors have been used to collect range of motion (ROM) data for wrist joint [9–11] in healthy subjects. Likewise, several studies have explored the use of smartphone-based sensor data collection to analyze gait [12–14]. The Patient Rheumatoid Arthritis Data from the Real World (PARADE) mobile software application (app) was developed in a successful attempt to assess the feasibility of using a ResearchKit-based iPhone app to conduct an end-to-end remote real-world evidence study [15]. The PARADE app included preliminary objective assessments of mobility leveraging iPhone embedded sensors and targeting two types of motion: wrist joint flexion-extension movement and walking.

This paper provides a description of the algorithms and machine learning models developed to analyze data generated in the PARADE study through these objective assessments. In both cases, data quality criteria were designed to identify low compliance with the instructions provided. Task-specific motion parameters were subsequently derived from high-quality data and compared with subjective data reported via questionnaires. The purpose of this data analysis was to examine the possibility of extracting clinically meaningful information from the data captured. Such novel, remote, mobility assessment tools are expected to become useful for gauging how the wrist joint movement and walk function are impacted by the degree of RA-induced joint pain and subsequent disability over time.

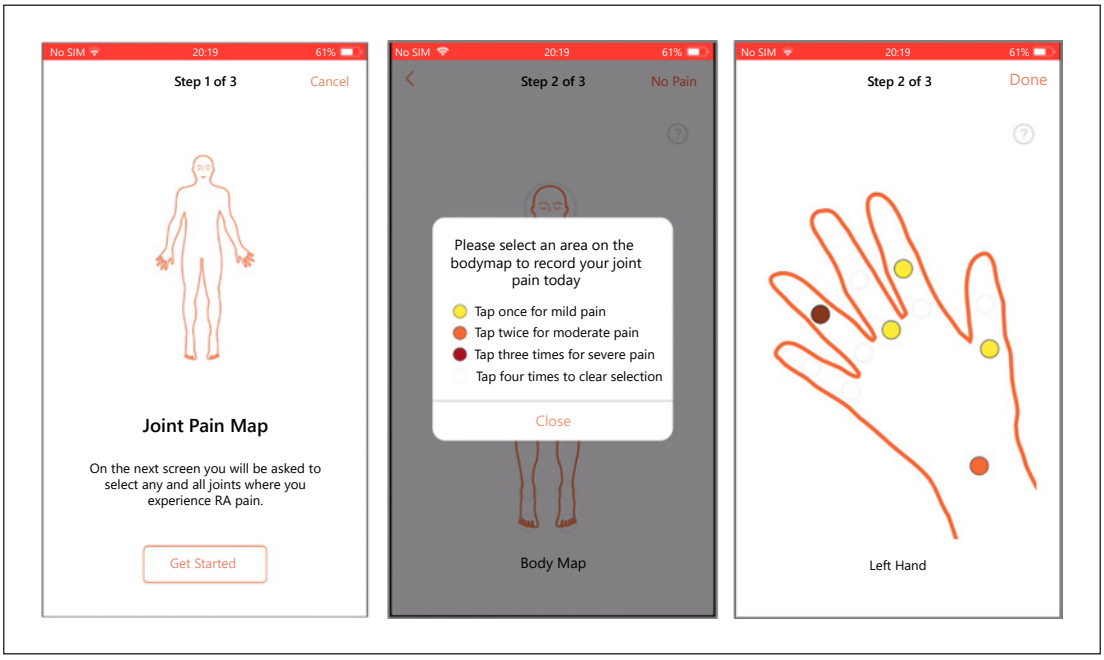

Fig. 1. Screenshots of the PARADE app showing the interactive joint-pain map tool utilized to record pain from 55 prespecified joints.

Materials and Methods

Data Collection

A total of 399 participants were enrolled in the study and invited to report on various symptoms each week, using their own iPhone, over a period of 12 weeks. The PARADE study was run in the United States between July and November 2016. The app was available for free on the AppStore. Baseline population characteristics have been described in a previous publication [15]. Quality of life was assessed using the 5-level version of the EuroQoL, 5 dimensions (EQ-5D-5L) questionnaire [16, 17] including questions on mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. An interactive joint-pain map was designed specifically for the app to record the number of painful joints (from 55 pre-specified joints which may be affected in RA patients) and the associated pain severity. At weeks 1 and 8, participants were asked to score the pain in each joint depicted in a body map (Fig. 1) as 0 (no pain, set as default), 1 (mild pain), 2 (moderate pain), or 3 (severe pain).

In addition to traditional patient-reported outcome questionnaires, the study participants were invited to perform two pre-defined activities for objective data collection in order to assess the impact of the disease on the wrist joint flexion-extension motion and gait. In both tasks, raw data from the iPhone's inertial measurement unit were captured to monitor participants' motion as they performed the tests. Gyroscope and accelerometer data were collected at a sampling frequency of 10 Hz.

Wrist joint motion sensor data were collected at weeks 1 and 12. For this task, on-screen guidance was provided by means of a short video: participants were instructed to sit down and place their forearm at the edge of a standard size table, with palm facing up, and holding the iPhone in their hand, and to flex and extend their wrist joint to its maximum ROM. Specific warning was given not to attempt to do this test in case of severe wrist joint or hand arthritis. Participants who indicated that they thought their wrist joint or hand arthritis would allow

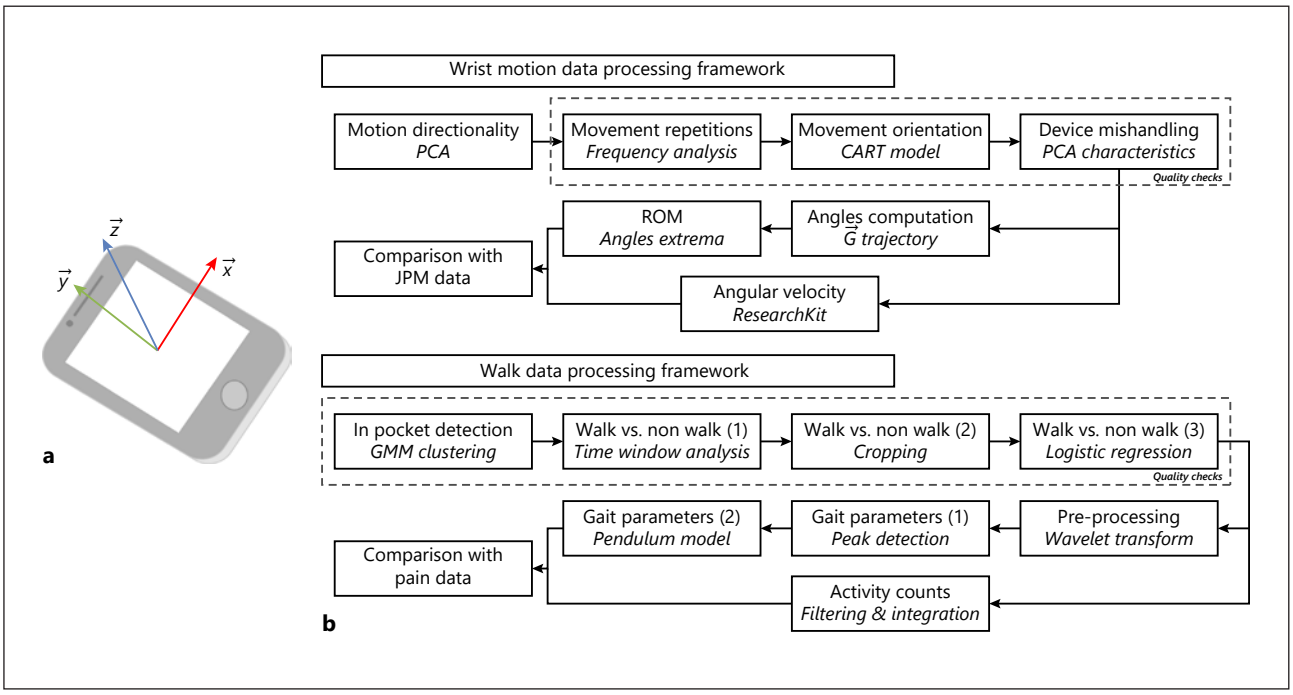

Fig. 2. Description of the iPhone's reference frame (**a**). Overview of the processing and analysis framework for data from both tasks (**b**). In each box, the upper part indicates the processing step while the lower text in italic describes the method used.

them to try the test were instructed to perform the task for 10 s with the iPhone held in the right hand first and then repeat it with the iPhone held in their left hand.

The walk task was designed in a similar fashion as in the mPower study [18] and consisted of three segments, each performed with the iPhone placed in the participant's pocket. Walk data were collected on a weekly basis over the whole study period. For this activity, participants were instructed to walk up to 10 steps in a straight line for 10 s (segment 1), turn around and stand still for 10 s (segment 2), and then walk up to 10 steps back for 10 s (segment 3). Audio instructions were provided to indicate the start and end of each segment. As for the wrist task, specific warning was given and only participants who indicated they could walk 10 steps without assistance were invited to complete the activity test.

Videos of the app screens detailing both objective task instructions and flow are available in the online suppl. material A (see www.karger.com/doi/10.1159/000506860 for all online suppl. material).

Data Analysis

The gravity vector in the iPhone's frame of reference – isolated from the accelerometer signal using gyroscope data – was used as a parameter of interest for motion quantification, as well as the rotation rate (for the wrist task only). Fig. 2a provides a description of the iPhone's axes naming convention for reference. All sensor data were collected in a real-world setting, without clinical supervision. Due to the high variability in the data, a number of task-specific checking steps were applied to identify and filter out samples where instructions had not been followed correctly while keeping all analyzable files. An overview of the analysis framework for data collected in each type of active task is shown in Fig. 2b. All analyses are presented from the "point of view" of the iPhone's frame of reference.

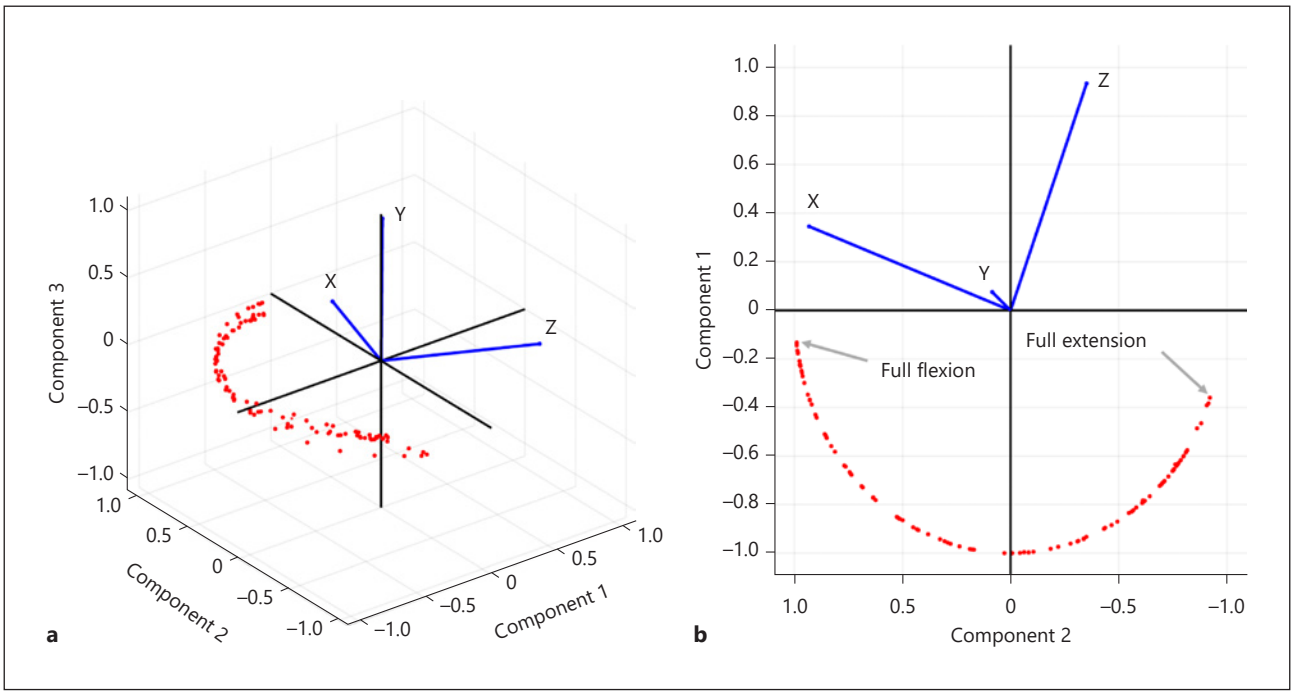

Fig. 3. Example principal component analysis of wrist joint motion with 3-dimensional representation of the data highlighting its nonplanar nature (**a**) and projection onto the plane that best represents the flexion-extension motion (**b**). Red dots indicate the trajectory of the gravity vector over time in the iPhone's frame of reference represented by the blue axes, while black axes represent the principal components.

Wrist Joint Motion

Motion Directionality

The primary challenge in analyzing the wrist motion data was to identify the main directionality of the recorded movement. In the ideal case, the motion would be performed in a robotic way including only rotation of the iPhone about $\vec{y}$ (see Fig. 2a). Such motion is very unlikely to be observed in the real world and especially in RA patients who are prone to experiencing pain and discomfort due to their disease, e.g., through joint deformities. Thus, it is expected that a limited proportion of the wrist motion will consist of pronation-supination (i.e., rotation about $\vec{x}$) and radial-ulnar deviation (i.e., rotation about $\vec{z}$). To account for this effect, principal component analysis (PCA) was applied to the trajectory of the gravity vector over time in the iPhone's frame of reference, thus allowing for derivation of the main wrist joint motion directionality. Assuming reasonable compliance with the task instructions, this main motion directionality captured in the plane defined by the first two principal components output by PCA will mainly reflect flexion-extension motion (see Fig. 3). Projecting the data points onto such plane enables the extraction of angles for subsequent computation of the flexion-extension ROM. Importantly, the iPhone orientation at the beginning of the task may differ depending on the positioning and shape of the participant hands and the direction may not exactly align with the direction of the gravity vector. Consequently, the detected neutral wrist orientation is likely to vary between different executions of the task both within and between subjects. For this reason, separate measures of range of flexion or range of extension cannot be estimated in a way that would allow for comparison between samples. Instead, a single measure was used for the ROM, defined as the median of angle differences between consecutive full flexion and full extension (see Fig. 3b). Additional details on the assessment of validity of the wrist ROM computation are available in the online suppl. material B. Both

ROM and wrist angular velocity were computed for all analyzable samples, i.e., that passed the checks described in the following paragraphs.

Movement Repetitions

Reliable measurement of the wrist joint ROM requires the full flexion-extension motion to be completed at least once within the allotted 10 s. This time duration, however, allows for multiple repetitions of the movement resulting in periodic variation of the gravity vector orientation in the iPhone's reference frame. The frequency of this periodic signal was measured as half of the number of local extrema. Samples with relatively high frequency signal imply that the movement was likely to have been performed too fast for the full ROM to be reached. Test data were collected from 5 healthy volunteers with no wrist motion problems who repeatedly performed the task with increasing numbers of movement repetitions. The ROMs computed from these test data were used to investigate the frequency range resulting in reduced estimates.

Movement Orientation and Device Mispositioning

Assessment of movement in the flexion-extension direction was the main target for the task, hence any motion including high proportions of supination-pronation and/or radial-ulnar deviation (see formalism in the online suppl. material C) should not be considered for analysis. Furthermore, as defined in on-screen guidance for the task, participants were instructed to place their iPhone in their hands, palm facing up. While enabling the measurement of wrist joint motion, this gives participants freedom to further increase the iPhone inclination using fingers or elbow when reaching their maximum wrist angle capacity, leading to the introduction of a measurement bias. In order to account for these extra motions, 300 samples were selected at random from the study dataset and manually labeled as analyzable or non-analyzable (using 3D modeling of the iPhone's motion). A set of features were subsequently computed for each sample, including:

- Maximum range of angles for data points projected onto the plane defined by the second and third PCA components (i.e., reflecting pronation-supination).
- Pearson's correlation between changes in $\vec{G}.\vec{x}$ and $\vec{G}.\vec{z}$ over time (i.e., reflecting whether $\vec{x}$ and $\vec{z}$ follow similar trajectories in the reference frame of the laboratory as impacted by radial-ulnar deviation), where $\vec{G}$ is the gravity vector in the iPhone's reference frame as represented in Fig. 2a.
- Angles in flexion-extension with respect to the initial wrist orientation (i.e., reflecting fingers and/or elbow motion).

The dataset formed of the computed features and associated high-/low-quality labels defined a training set for a classification and regression tree (CART) model. The model was pruned by reduction on a 5-fold cross-validation error and applied to the full wrist motion dataset for classification as high or low quality. A specific restriction on the angle between the third PCA component was also introduced to account for instances where the iPhone was not held properly.

Walk Test

A correctly performed walk test means the participant was walking with the iPhone in their pocket when instructed to do so by the app (segments 1 and 3), and not walking during the resting phase (segment 2). Upon initial inspection of the collected data, two types of non-compliance with these instructions could be identified: (i) task performed while the iPhone was not in a pocket, (ii) participants not walking during the walking phases or walking during the non-walking phase. The former case made it unreliable to compare samples, for example if a participant was holding the iPhone in their hand and swinging

their arm as they were walking, the intensity of the activity recorded by the iPhone's sensors may appear comparatively larger than if the iPhone was in the pocket. Therefore, these samples had to be discarded. In the latter case, the samples could still be used but the label assigned to them (walking/not walking) based on the type of segment needed to be updated. The following pre-processing steps were applied to identify instances where these situations occurred.

Phone Location

If the iPhone was in a pocket, the angle between $\vec{z}$ and $\vec{G}$ would be expected to be close to 90°. If the angle was closer to 180°, i.e., iPhone facing upwards, it is more likely that the participant had the iPhone out of their pocket and was looking at the screen. Unsupervised machine learning algorithms including: 2- and 3-means as well as 2- and 3-components Gaussian mixture model (GMM) was used to define clusters in the feature space defined by the mean and standard deviation for this angle over time.

Walking versus Not Walking

Each performance of the task produced one data file per segment, thus providing an initial classification between walking (forth or back) and non-walking (i.e., turn around and stand still). In this stage, data from both walking segments were combined into a single category. However, such initial classification needed to be refined.

As a first step, local averaging of raw acceleration signals was applied using a sliding time window (2 s with 1 s overlap). Samples in which the user acceleration continuously exceeded a magnitude threshold of 0.05 g for a duration of at least 6, 8, or 10 s remained labeled as walking and were trimmed so that only the signal of interest could be kept. The selection of the best of these three values for the time duration threshold is described in the Results section. Samples not matching this criterion were re-labeled as not walking. The magnitude threshold value selection was based on a trade-off between excluding signals with a low amplitude throughout the whole spectrum – obtained by Fourier transform of the acceleration signal – and including signals with low amplitude except for a peak in the 1.5–2 Hz region (typically associated with walking [19]). This trade-off was performed using visual inspection of the amplitude Fourier spectra.

Following the initial re-labeling based on signal characteristics, subsequent data re-classification was performed to further refine such labels. This could be achieved by training a classifier using the imperfectly labeled dataset and then reassign labels to the same data using the trained model. This approach was justified as a small proportion of samples would be labeled incorrectly after time window analysis. Logistic regression was chosen as classifier due to its flexibility and its characteristic soft assignment to classes. Re-classification consisted of three stages:

- Training using imperfect labels obtained from previous time window analysis (overfitting was avoided using cross-validation).
- Use model to estimate each sample's probability to correspond to walking (i.e., soft assignment).
- Selection of an appropriate cut-off probability value using the model's receiver operating characteristic curve and update labels.

The following features were used to identify walking: (i) the peak of the amplitude spectrum of the vertical component of the acceleration signal, for which a peak in frequency is expected in the 1.5–2 Hz region, (ii) the peak of the amplitude spectrum of the angle between $\vec{y}$ and $\vec{G}$, as this angle is expected to oscillate following the leg's pendulum-like motion.

Digit Biomark 2020;4:26–43
DOI: 10.1159/000506860

Gait Parameterization

For all analyzable walk samples (i.e., labeled both as "in-pocket" and "walking" after re-classification), a number of processing steps were applied to the signal, starting with the extraction of initial foot contacts (or heel strikes) as well as final contacts (or toe offs) using McCamley's method [20]. This method consists of integrating the vertical component of the accelerometer signal which is then differentiated using the continuous wavelet transform with a Gaussian wavelet kernel. Peak detection applied to the resulting signal provided the occurrences of initial and final contacts corresponding to minima and maxima, respectively. Once extracted, the detected peaks were used to compute step times and derive the number of steps. A pendulum model [21, 22] was subsequently used to compute estimates of the average step length, assuming that the iPhone was located at a height equal to half the participant's height. Additional details on the assessment of validity of the gait parameters computation are available in the online suppl. material B.

The average number of activity counts per second [23] was also computed to provide a measure of how vigorously the participant was walking throughout the test.

Comparison to Pain Data

All the measures derived from each of the objective data collection tasks were subsequently compared with specific pain scores independently collected via the app questionnaires. As the wrist joint motion test predominantly mobilizes a single joint, ROM measures were directly compared with the corresponding entries in the joint-pain map (i.e., right or left wrist). However, because walking is a much more complex type of activity involving several joints (including ankles, knees, hips, and back/spine), parameters related to gait cannot be compared to the pain score recorded for a single joint. For this reason, more generic outcomes from questionnaires available through the app and targeting symptoms relevant to walking were used for comparison. These include:

- Pain: current level of pain on a scale of 1 (no pain) to 10 (worst pain imaginable).
- Pain/discomfort: current level of pain/discomfort on a scale of 1 (no pain) to 5 (extreme pain) from EQ-5D-5L.
- Mobility: current level of mobility on a scale of 1 (no problem walking about) to 5 (unable to walk about), also from EQ-5D-5L.

Both the EQ-5D-5L and joint-pain map instruments have been psychometrically evaluated at group level in RA. The pain scale had not been psychometrically evaluated at the time of the study.

The mobile application allowed for large variability (up to 1 week) in time difference between completion of the various tasks and questionnaires. To account for this and ensure that pain levels were reflecting the participant's experience at the time of the activity task, comparison was only performed between questionnaire and sensor data that had been collected within a maximum period of 2 h of each other.

Results

The total volume of data collected for PARADE's objective tasks included 376 wrist test attempts (29 not completed due to hand arthritis) and 1,086 walk test attempts (49 not completed due to inability to walk without assistance) representing 26% of ideal case full compliance across the 399 participants.

Wrist Joint Motion

In total, 646 samples were collected from 287 participants throughout the study period (week 1 and week 12), including 570 samples at week 1.

Movement Repetitions

Test data analysis showed that reduced measures of ROM were obtained for samples including 6 or more repetitions of the full movement. Therefore, samples where the frequency fell below 0.1 Hz (i.e., at least one full flexion-extension) or above 0.55 Hz (i.e., at most 5.5 repetitions) were excluded from further analysis.

Movement Orientation and Device Mispositioning

Cut-off values for the different orientation-related features as derived from the CART model training to define high quality were: maximum pronation-supination range of 50°, radial-ulnar range resulting in Pearson's correlation coefficient $r \geq 0.94$, and a limitation on the flexion-extension over time not to exceed 100° in more than 30% of the data points. Additionally, samples where the third PCA component deviated from $\vec{y}$ by more than 35° were also excluded to avoid bias due to device mispositioning. The classification error of the CART model on the training set was 4% and visual inspection of the remaining samples classification did not highlight any obvious issue.

Comparison to Pain Data

It was found that 289 (257 at week 1) samples satisfied all three QC criteria, representing 45% of the total data volume. 101 (35%) out of the 287 participants who took part in the test provided analyzable data for at least one of the 2 weeks.

ROM data were compared with wrist joint pain information provided through the joint-pain map. In total there were 233 week 1 analyzable samples with joint-pain map data available (i.e., participants had completed both tasks within the 2-hour time window). Such comparisons could not be performed using week 12 data because wrist joint specific pain information was not captured at that time point.

Fig. 4 shows boxplots of wrist joint ROM and maximum angular wrist velocity for each reported wrist joint-pain level at week 1, with no distinction between right and left joint. The lower ROMs measured by the iPhone sensors for severe pain levels are consistent with clinical expectation. Maximum angular velocity also appears to decrease with increasing pain level. There was a statistically significant difference between severity groups as determined by one-way analysis of variance (ANOVA) for both wrist joint ROM ($F(3,229) = 6.59$, $p < 0.001$) and maximum wrist angular velocity ($F(3,229) = 10.87$, $p < 0.0001$). Tukey's honestly significant difference (HSD) post hoc test subsequently showed that both measures in the severe wrist joint-pain group significantly differ from the other groups. Moreover, the maximum angular velocity in the moderate wrist joint-pain group also significantly differs from the group with no pain reported (indicated as "N/A" in Fig. 4). A similar repartition of high- and low-quality data across wrist pain groups was observed as summarized in the tables available in the online suppl. material D.

Walk Test

In total, the test was attempted in 867 instances (including 854 walk-forth, 863 walk-back, and 866 turn-around-and-stop), resulting in 2,583 single-segment samples collected from 316 participants throughout the whole study period (week 1 to week 12), including 867 samples at week 1.

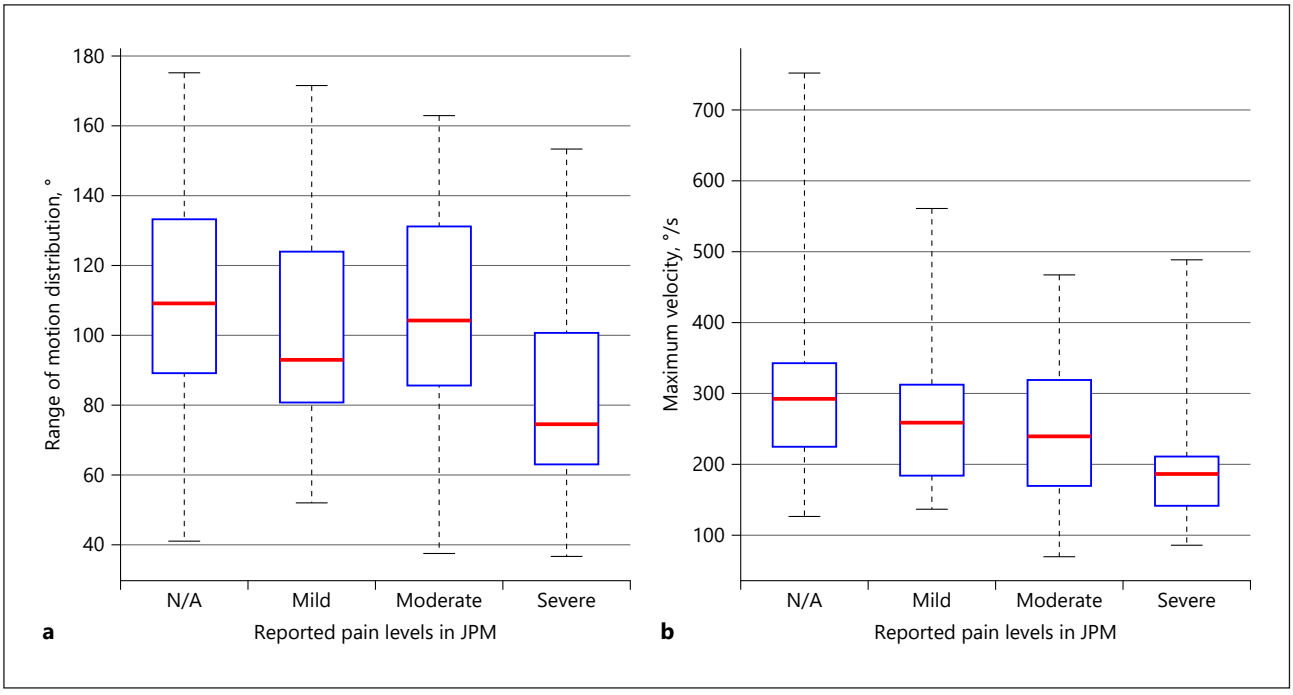

Fig. 4. Boxplots for ROM (**a**) and maximum angular velocity (**b**) with respect to wrist pain level reported in the joint-pain map at week 1. In each box the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points.

Phone Location

Amongst the different clustering algorithms tested, the 3-component GMM provided the most sensible results, allowing for curved decision boundaries for the two clearly identifiable clusters in the data distribution represented in Fig. 5a. The model output is shown in Fig. 5b. The cluster made of red circles is near a mean of 180° and corresponds to cases where the iPhone was facing up for the entire segment, while the cluster made of green squares appears closer to a mean of 90° and corresponds to cases where the iPhone was near-vertical (likely in the pocket) for the entire test segment. The remaining points in the blue cluster labeled as "undefined" are likely to be cases in which the iPhone was in the participant's pocket only during a limited part of the segment. Points where the standard deviation is large can be interpreted as samples where the participants moves (and rotates) the iPhone during the recording phase, possibly to take the iPhone out of their pocket and/or put it back into their pocket. Visual inspection of the gravity vector trajectory for samples from each of the clusters identified by the model corroborated these interpretations.

According to the 3-component GMM, 29.9% of the samples contain data collected with the iPhone out of the pocket, 41.4% contain data collected with the iPhone in the pocket, and 28.8% of the samples contain undefined data. Whilst the participants who followed the instructions to place the iPhone in the pocket form the largest cluster, a number seem to have held the iPhone in their hand or otherwise out of their pocket during the task.

Walking versus Not Walking

The time threshold values used for time window analysis were tested to optimize the trade-off between including samples of sufficient walking duration for subsequently extracted gait parameters to be meaningful, and including as many samples as possible in the selection.

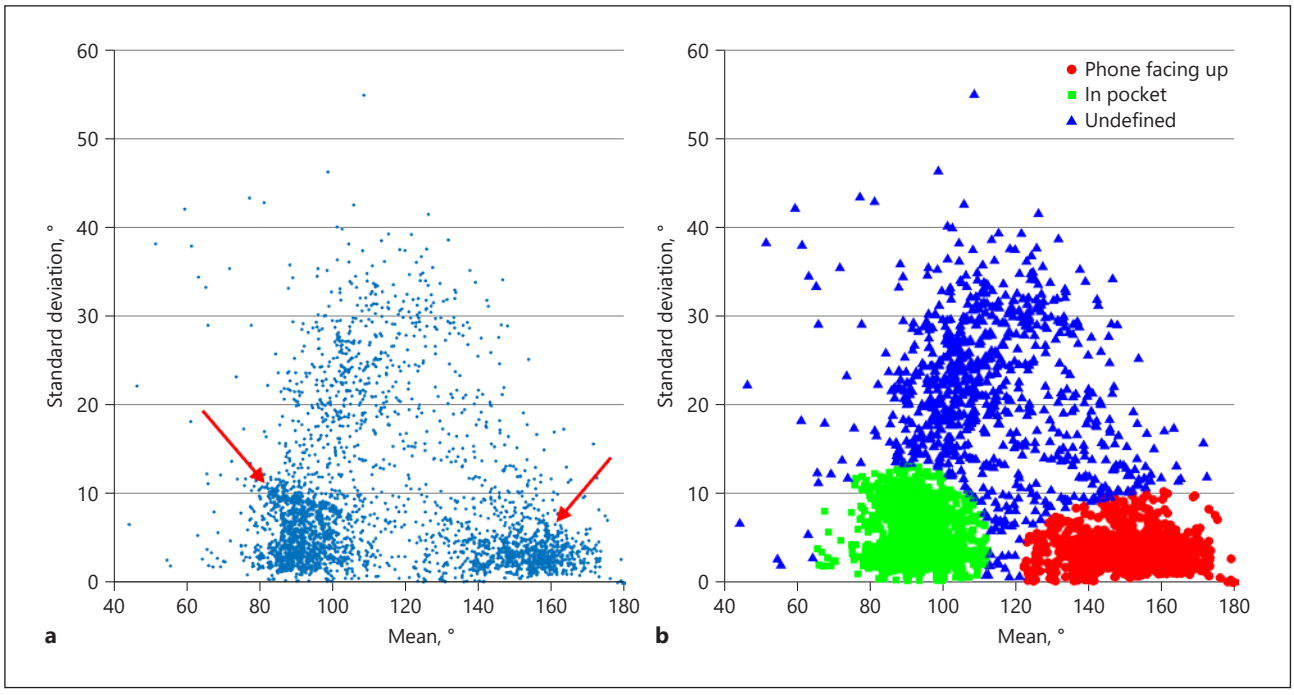

Fig. 5. Mean and standard deviation (over time) of the angle between $\vec{z}$ and $\vec{G}$, with two clusters visually identifiable indicated by arrows (**a**). The 3-component Gaussian mixture output shows 3 clusters for iPhone in pocket, iPhone facing up, and undefined (**b**).

The 10 s threshold was not kept as it only included 15.4% of the samples labeled as in pocket. Setting the threshold at 6 or 8 s resulted in a similar number of samples included with 25.9 and 25.2%, respectively.

The relabeling resulting from the time window analysis efficiently identified most samples incorrectly labeled as "walking" (details of the Fourier spectra inspection are available in the online suppl. material C). However, a number of samples remain mis-labeled, which suggests that a subsequent re-classification stage with logistic regression was appropriate.

The logistic regression classifier was trained on the "in-pocket" data. Based on the soft class assignments output by the model, a probability threshold of 41% was chosen, thus determining a particular tuning of the classifier. This choice corresponds to the optimal operating point on the receiver operating characteristic curve (details available in the online suppl. material C) and was guided by sensitivity analysis. It is a trade-off between high confidence that the participant performed the test correctly and inclusion of participants who walked less vigorously. This relatively low value prioritizes the latter over the former, as RA patients are likely to have more restricted mobility due to their disease.

Gait Parameterization

It was found that 651 samples satisfied both QC criteria, representing 25% total data volume. 111 (35%) out of the 316 participants who took part in the test provided analyzable data for at least one of the twelve data collection time points.

Gait parameters computed for analyzable samples are presented in Fig. 6. Mean values for the average step length, step time, and number of steps are 0.66 m, 0.7 s, and 10.4, respectively. Also presented in Fig. 6b is the comparison between average step length and average counts of activity showing a moderate correlation (Pearson correlation coefficient r = 0.56,

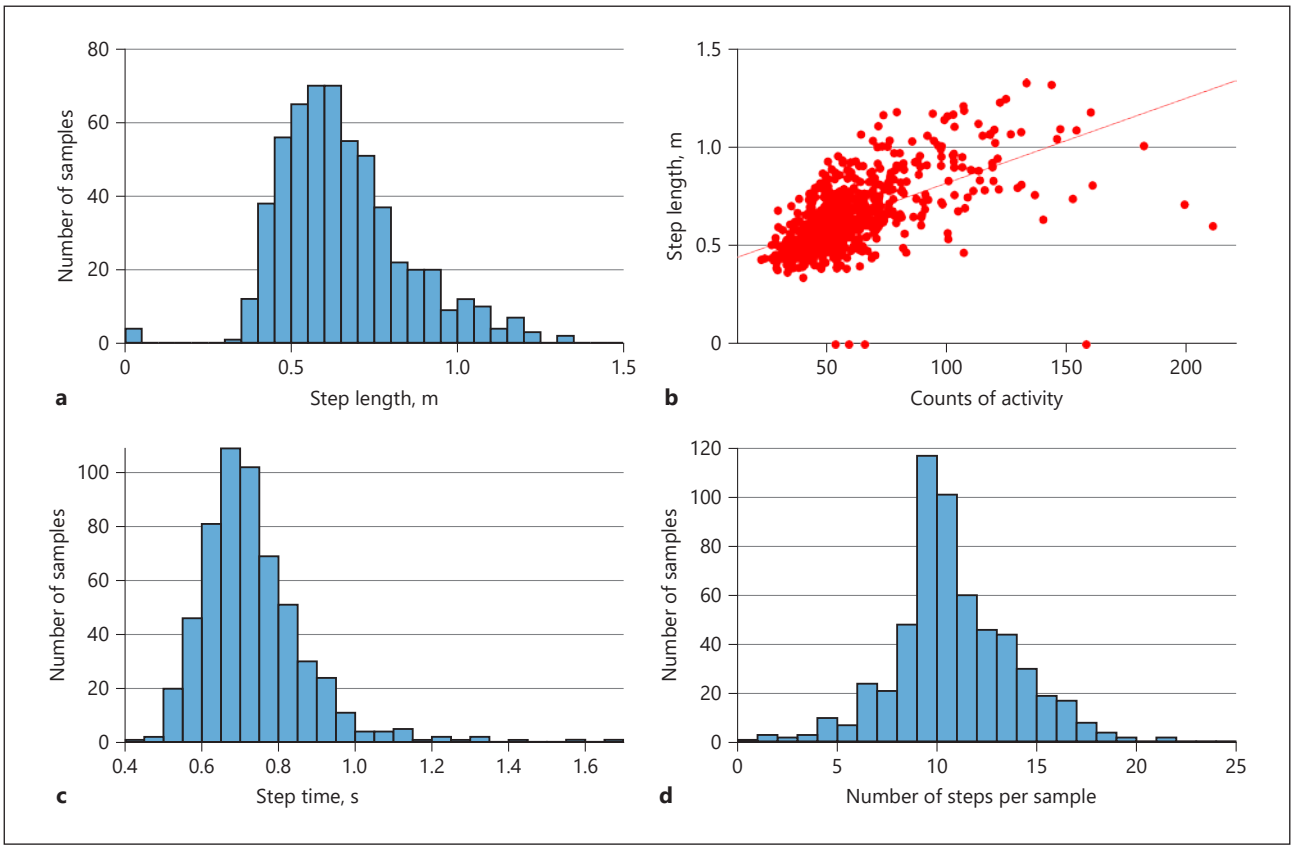

Fig. 6. Distributions of average step length (**a**) along with the comparison to counts of activity (**b**; a red line marks the linear fit to the data), average step times (**c**), and number of steps across all analyzable samples (**d**).

$p < 1e{-}50$). This trend is in agreement with the intuitive expectation that the more vigorous the walking, the longer the steps. The distribution of activity counts is centered around a mean of 62.73 (median 55.73, interquartile range 24.42) counts per second, which is consistent with values previously reported for slow walking in healthy individuals [23, 24]. Step frequency, derived from the step times, has a mean and median value of 1.4 Hz (interquartile range 0.33), which is consistent with the expected frequency of walking (typically in the range of 1.5–2 Hz in healthy individuals).

Comparison to Questionnaire Data

In total 440 and 201 samples were available for comparison with pain scores and scores from the EQ-5D-5L (i.e., participants had completed the task and questionnaires within the 2-hour time window), respectively. Interestingly, one participant indicated they were unable to walk but did performed the walk test. However, the corresponding samples did not match the requirement for the time difference between completion of the objective task and the EQ-5D-5L. This provides evidence of symptom variability and justifies the use of the 2-hour time window.

Encouraging trends were observed as illustrated in Fig. 7. In particular, step time appears to slightly increase with increasing pain, pain/discomfort, and mobility scores. Also, the step length appears to slightly decrease with increasing scores. Step frequency and velocity logically follow the same trend as inversely proportional to step time and – for velocity only – proportional to step length, hence these are not included in Fig. 7 for the sake of clarity. There

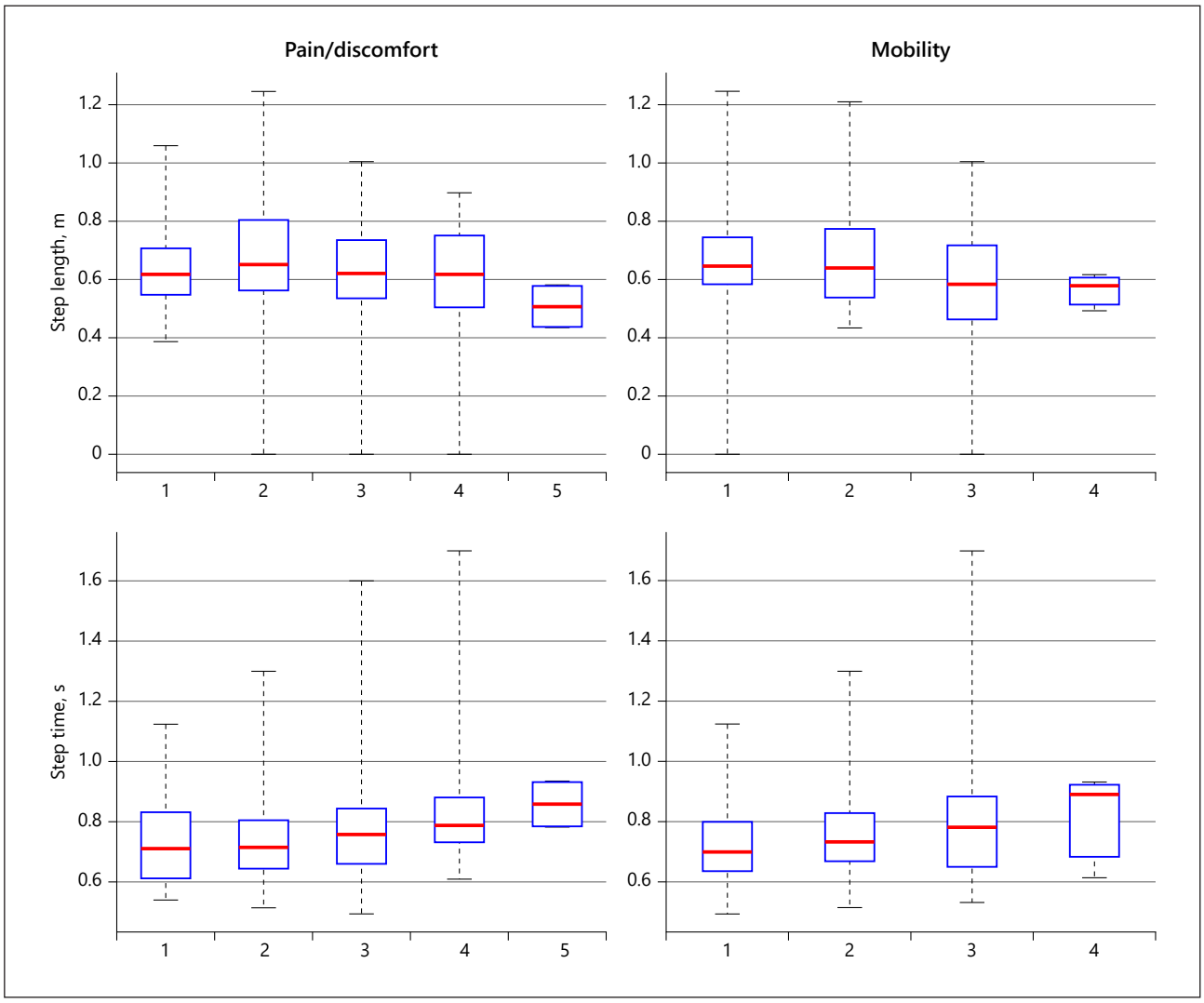

Fig. 7. Boxplots for step length and step time with respect to pain/discomfort and mobility scores reported through the app questionnaire. Step length and step time, respectively, appear to slightly decrease and increases with increasing scores. In each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points.

was no statistically significant difference between pain groups as determined by one-way ANOVA for step length and step time nor between pain/discomfort groups for step length. However, there was a statistically significant difference between pain/discomfort groups for step time ($F(4,195) = 2.71$, $p < 0.04$), with Tukey's HSD showing that step time in the pain/discomfort score 2 group significantly differed from the pain/discomfort score 4 group. Lastly, there was a statistically significant difference between mobility groups for both step length ($F(3,196) = 2.83$, $p < 0.04$) and step time ($F(3,196) = 3.06$, $p < 0.03$). Tukey's HSD showed that both step measures in the mobility score 3 group significantly differed from the mobility score 1 group. In the case of step length, the mobility score 3 group also significantly differs from the mobility score 2 group. A similar repartition of high- and low-quality data across pain/discomfort and mobility groups was observed as summarized in the tables available in the online suppl. material D.

Digit Biomark 2020;4:26–43
DOI: 10.1159/000506860

Discussion/Conclusion

This study addresses the challenge of deriving meaningful information on functional ability from objective sensor data collected remotely from patients with RA. It is based on the PARADE study in which all participants' data were collected entirely through an iPhone app and which targeted two types of physical functions: wrist joint flexion-extension and walking. The principal finding of the described work is the highlighted link between objective measurements and the participant-reported information on pain, discomfort, and mobility. This suggests that the collected data do convey clinically meaningful information. The development of task-specific schemes for extraction of quantitative measures from raw sensor data yielded encouraging results despite challenges related to data quality and adherence.

A number of studies using smartphone apps including the combination of questionnaires and objective tasks to assess the impact of a disease have been reported. Gait assessment tests have been deployed in research on Parkinson's disease for both observational [18] and interventional [12] studies. In particular, several studies on RA investigated smartphone sensor data collected during walking over a short distance using features such as peak frequency, autocorrelation peak, and coefficient of variance from the acceleration signal [13, 14]. In the present case, it was chosen to extract gait parameters directly related to steps as these seem more meaningful and appropriate for direct comparison to clinical data. However, the use of a specific model to estimate step length involves assumptions, for example on the participants' leg length.

Unlike walking, few studies focused on wrist ROM assessment. While this work relied on the computation of angles using information on the gravity vector orientation, other approaches essentially involved direct on-screen reading of angles from built-in iPhone apps [9, 11] or third-party applications [10]. The choice of the gravity vector in the iPhone's reference frame as a representation of orientation in space was motivated by the access to the vertical direction while keeping the same axis convention, as opposed to other representations such as attitude quaternion, which is specified with respect to a laboratory reference frame with arbitrary directions for $\vec{x}$ and $\vec{y}$.

The level of compliance to the objective tasks deployed in this study was moderate (26%). Similar challenges around compliance were observed in mPower, which also was a fully remote, digital observational study [18]. However, higher compliance was reported in other recent studies with different settings. In Perraudin et al. [25], 56% of the administered repeated sit-to-stand task were performed as instructed. The study included supervised performance of the objective task during face-to-face onsite visits at the beginning and at the end of the 4-week study period. This may partly explain the larger proportion of analyzable data in comparison to that observed in PARADE, bearing in mind the difference in the nature of the objective tasks (i.e., wrist joint flexion-extension/walking versus sit-to-stand). Likewise, Lipsmeier et al. [12] reported 61% adherence to active tests in their study, where the initial in-clinic visit, during which training was provided, possibly contributed to this high rate.

The performance of the algorithms used for wrist ROM computation and gait analysis were assessed through comparison with gold standard goniometer measures and GAITRite® mat parameters, respectively. Details of those experiments are available in the online suppl. material B and include quantitative information on the algorithms' accuracy and applicability. While these experiments were useful to show the feasibility of deriving meaningful information from sensor data, a more thorough validation, including for instance larger numbers of subjects, is necessary.

Generating accurate and truthful objective measures requires robust data collection schemes and high-data quality. In that sense, this study has several limitations. The variability

observed in the sensor data was a challenge. Objective activity tasks were performed as expected in 45% of wrist motion data samples and 25% of walk data samples. The corresponding volume of analyzable data was sufficient to run a comparison with symptom severity data and the random distribution of low-quality data did not seem to bias the outcome. However, it is desirable to obtain higher proportions of high-quality data as this is likely to result in stronger signals. The low success rate in meeting quality criteria is likely to be related, at least in part, to the absence of clinical supervision and/or training: task-specific (as opposed to free living) data acquisition conducted in a fully unsupervised manner requires that extremely clear and extensive guidance is provided to participants. Preliminary tests run in a controlled clinical environment could be considered for the deployment of similar objective activity tasks in the future. Such tests may provide information on how participants interpret the instructions, which may then be used to improve guidance. The choice of the pocket as the iPhone's location during the walk task may be the cause of less reliable gait parameters estimates as opposed to, for example, attaching the device to the participant's back or hip with a belt [22]. It was found that a significant number of participants (at least 29.9%) did not perform the test with the phone in their pocket for the whole task duration, possibly due to a misunderstanding of the instructions. This limitation may be addressed in future studies by requesting participants to place their iPhone on a different body location, including the use of a strap around the participant's leg or waist, or using a wearable device to collect the data. Further improvement of gait parameters estimates may be obtained by setting up a clearer target in terms of number of steps (instead of allowing for "up to" 10 steps) or time. Non-analyzable samples were also found in the wrist task dataset. Improvement in consistency of execution regarding this type of test may be obtained by adding further detailed guidance (e.g., to specify how not to hold the iPhone at the beginning or during the test) to improve clarity.

Additionally, quality control algorithms developed for this study rely on several thresholds defined empirically and optimized for the study population. Although cross-validation and subsequent results inspection were used, the machine learning models applied for quality control were directly trained and applied to single datasets, which limits generalization to other study datasets. Ideally these thresholds should be prospectively defined through supervised tests based upon ground truth observations. In-depth validation accounting for as many edge cases (e.g., with respect to age, morphology, and disease severity) as possible is required to refine the parameterization of such algorithms. If successful, these quality control algorithms may be incorporated to future versions of the objective tasks, either as remote monitoring tools or as real-time prompts displayed on screen and notifying participants that the task is not being performed as expected.

Sensor data analysis in PARADE was limited by the overall high level of attrition and the observational nature of the study (i.e., no medical intervention) [15]. In this context, estimating change over time in the derived measures would be irrelevant and longitudinal assessment was therefore not included as a study objective. However, the definition of concepts such as minimum clinically important difference for objective measures including wrist ROM or gait parameters would be of interest in future studies with the appropriate settings. In an attempt to reduce attrition, future work may include reminders sent through the app to participants who do not comply with the schedule of activities.

A further limitation pertains to the availability of participant-reported data collected at the same time points as the objective test. The frequency at which tasks were assigned was defined to minimize the burden to participants at a specific time point. Some assessments such as the joint-pain map and the wrist joint motion test were administered in such a way that corresponding data collection overlapped at baseline, enabling investigation of correlations. Future work may consider synchronized collection of objective data and participant-

reported information to maximize the relevance of comparison. This is particularly important in diseases such as RA where symptoms fluctuation is high.

A possible extension of the data analysis work presented in this paper may leverage machine learning models to explore the predictive power of mobility parameters and joint-pain map on disease severity. The joint-pain map contains the most direct indication of the level of pain in each joint. It may be beneficial to consider combinations of the pain levels of multiple joints reported through the joint-pain map that would better reflect the mechanisms involved in different types of movement. Again, machine learning techniques could be used to define the relevant weighting to generate the combination providing the best prediction of overall pain scores.

In summary, this work presented methodologies developed for the analysis of iPhone-captured time-series data collected during objective activity tasks on wrist motion and walking. Movement-specific processing based on various learning algorithms was successfully applied to assess data quality, extract measures of interest, and compare these to participant-reported pain. This feasibility study for fully remote real-world data collection resulted in encouraging trends suggesting that iPhone sensor data can provide meaningful information on functional ability in patients suffering from RA, including sensitivity to disease severity. Such information may contribute to a better understanding of patients' overall quality of life. Further investigation and improvement to data acquisition schemes would be needed to confirm these observations. This includes prospective testing and changes to the design of the objective tasks to reduce variability and therefore increase data quality along with output measures' accuracy.

Acknowledgement

The authors are very grateful to all participants who used the app, for their time and generous input to the study. The authors also wish to acknowledge the contributions of Possible Mobile, who developed the PARADE app.

Statement of Ethics

This study (study number 205718) was sponsored by GSK (GlaxoSmithKline plc., London, United Kingdom) and approved by the Quorum institutional review board (Quorum Review Inc., Seattle, WA, USA). Electronic informed consent was obtained from study participants, in compliance with Title 21 of the Code of Federal Regulations Part 11 (21CP11). GSK designed the study; contributed to the collection, analysis, and interpretation of the data; and supported the authors in the development of the manuscript. GSK is committed to publicly disclosing the results of GSK-sponsored clinical research that evaluates GSK medicines and, as such, was involved in the decision to submit the manuscript for publication.

Disclosure Statement

At the time of the study, V.H., L.G.-G., E.Q., R.W., and M.C. were employees of GSK (GlaxoSmithKline plc., London, United Kingdom) and held shares; at the time of submission, E.Q. and M.C. are no longer with GSK. A.P., A.M., J.L., A.H., and P.B. have no conflicts of interest to declare.

Funding Sources

This study was entirely funded by GSK (GlaxoSmithKline plc., London, United Kingdom).

Author Contributions

V.H. edited the manuscript, supervised the whole data analytics work, generated code for the wrist ROM computation, and processed the study data. L.G.-G., E.Q., and P.B. provided input to the data analytics work and supervised the methodology design. J.L. and A.H. generated code for the wrist data analysis (including pre-processing and ROM computation). A.M. and A.P. designed and developed the methodology for walk test data processing (including pre-processing and gait parametrization). M.C., R.W., and E.Q. designed and managed the PARADE study. All authors provided input and comments to the manuscript.

References

1 Carr A, Hewlett S, Hughes R, Mitchell H, Ryan S, Carr M, et al. Rheumatology outcomes: the patient's perspective. J Rheumatol. 2003 Apr;30(4):880–3.
2 Hewlett S, Smith AP, Kirwan JR. Measuring the meaning of disability in rheumatoid arthritis: the Personal Impact Health Assessment Questionnaire (PI HAQ). Ann Rheum Dis. 2002 Nov;61(11):986–93.
3 Hewlett S, Smith AP, Kirwan JR. Values for function in rheumatoid arthritis: patients, professionals, and public. Ann Rheum Dis. 2001 Oct;60(10):928–33.
4 Cutolo M, Villaggio B, Otsa K, Aakre O, Sulli A, Seriolo B. Altered circadian rhythms in rheumatoid arthritis patients play a role in the disease's symptoms. Autoimmun Rev. 2005 Nov;4(8):497–502.
5 Katz PP. The impact of rheumatoid arthritis on life activities. Arthritis Care Res. 1995 Dec;8(4):272–8.
6 Cross M, Smith E, Hoy D, Carmona L, Wolfe F, Vos T, et al. The global burden of rheumatoid arthritis: estimates from the global burden of disease 2010 study. Ann Rheum Dis. 2014 Jul;73(7):1316–22.
7 Felson DT, Anderson JJ, Boers M, Bombardier C, Chernoff M, Fried B, et al.; The Committee on Outcome Measures in Rheumatoid Arthritis Clinical Trials. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. Arthritis Rheum. 1993 Jun;36(6):729–40.
8 Prevoo ML, van 't Hof MA, Kuper HH, van Leeuwen MA, van de Putte LB, van Riel PL. Modified disease activity scores that include twenty-eight-joint counts. Development and validation in a prospective longitudinal study of patients with rheumatoid arthritis. Arthritis Rheum. 1995 Jan;38(1):44–8.
9 Kim TS, Park DD, Lee YB, Han DG, Shim JS, Lee YJ, et al. A study on the measurement of wrist motion range using the iPhone 4 gyroscope application. Ann Plast Surg. 2014 Aug;73(2):215–8.
10 Pourahmadi MR, Ebrahimi Takamjani I, Sarrafzadeh J, Bahramian M, Mohseni-Bandpei MA, Rajabzadeh F, et al. Reliability and concurrent validity of a new iPhone® goniometric application for measuring active wrist range of motion: a cross-sectional study in asymptomatic subjects. J Anat. 2017 Mar;230(3):484–95.
11 Modest J, Clair B, DeMasi R, Meulenaere S, Howley A, Aubin M, et al. Self-measured wrist range of motion by wrist-injured and wrist-healthy study participants using a built-in iPhone feature as compared with a universal goniometer. J Hand Ther. 2019 Oct - Dec;32(4):507–14.
12 Lipsmeier F, Taylor KI, Kilchenmann T, Wolf D, Scotland A, Schjodt-Eriksen J, et al. Evaluation of smartphone-based testing to generate exploratory outcome measures in a phase 1 Parkinson's disease clinical trial. Mov Disord. 2018 Aug;33(8):1287–97.
13 Yamada M, Aoyama T, Mori S, Nishiguchi S, Okamoto K, Ito T, et al. Objective assessment of abnormal gait in patients with rheumatoid arthritis using a smartphone. Rheumatol Int. 2012 Dec;32(12):3869–74.
14 Nishiguchi S, Yamada M, Nagai K, Mori S, Kajiwara Y, Sonoda T, et al. Reliability and validity of gait analysis by android-based smartphone. Telemed J E Health. 2012 May;18(4):292–6.
15 Crouthamel M, Quattrocchi E, Watts S, Wang S, Berry P, Garcia-Gancedo L, et al. Using a ResearchKit smartphone app to collect rheumatoid arthritis symptoms from real-world participants: feasibility study. JMIR Mhealth Uhealth. 2018 Sep;6(9):e177.
16 Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res. 2011 Dec;20(10):1727–36.
17 Janssen MF, Pickard AS, Golicki D, Gudex C, Niewada M, Scalone L, et al. Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: a multi-country study. Qual Life Res. 2013 Sep;22(7):1717–27.
18 Bot BM, Suver C, Neto EC, Kellen M, Klein A, Bare C, et al. The mPower study, Parkinson disease mobile data collected using ResearchKit. Sci Data. 2016 Mar;3(1):160011.

19 Ji T, Pachi A. Frequency and velocity of people walking. Struct Eng. 2005;84(3):36–40.
20 McCamley J, Donati M, Grimpampi E, Mazzà C. An enhanced estimate of initial contact and final contact instants of time using lower trunk inertial sensor data. Gait Posture. 2012 Jun;36(2):316–8.
21 Zijlstra W, Hof AL. Assessment of spatio-temporal gait parameters from trunk accelerations during human walking. Gait Posture. 2003 Oct;18(2):1–10.
22 Silsupadol P, Teja K, Lugade V. Reliability and validity of a smartphone-based assessment of gait parameters across walking speed and smartphone locations: Body, bag, belt, hand, and pocket. Gait Posture. 2017 Oct;58:516–22.
23 Rowlands AV, Stiles VH. Accelerometer counts and raw acceleration output in relation to mechanical loading. J Biomech. 2012 Feb;45(3):448–54.
24 Treuth MS, Schmitz K, Catellier DJ, McMurray RG, Murray DM, Almeida MJ, et al. Defining accelerometer thresholds for activity intensities in adolescent girls. Med Sci Sports Exerc. 2004 Jul;36(7):1259–66.
25 Perraudin CG, Illiano VP, Calvo F, O'Hare E, Donnelly SC, Mullan RH, et al. Observational study of a wearable sensor and smartphone application supporting unsupervised exercises to assess pain and stiffness. Digit Biomark. 2018 Oct;2(3):106–25.