



## ARTICLE OPEN



# Digital health technologies and machine learning augment patient reported outcomes to remotely characterise rheumatoid arthritis

Andrew P. Creagh<sup>1,2</sup><sup>✉</sup>, Valentin Hamy<sup>3</sup>, Hang Yuan<sup>1,2,4</sup>, Gert Mertes<sup>1,2,4</sup>, Ryan Tomlinson<sup>5</sup>, Wen-Hung Chen<sup>5</sup>, Rachel Williams<sup>5</sup>, Christopher Llop<sup>6</sup>, Christopher Yee<sup>6</sup>, Mei Sheng Duh<sup>6</sup>, Aiden Doherty<sup>1,2,4,7</sup>, Luis Garcia-Gancedo<sup>1,2,4,7</sup> and David A. Clifton<sup>1,7</sup>

Digital measures of health status captured during daily life could greatly augment current in-clinic assessments for rheumatoid arthritis (RA), to enable better assessment of disease progression and impact. This work presents results from wearABLE-PRO, a 14-day observational study, which aimed to investigate how digital health technologies (DHT), such as smartphones and wearables, could augment patient reported outcomes (PRO) to determine RA status and severity in a study of 30 moderate-to-severe RA patients, compared to 30 matched healthy controls (HC). Sensor-based measures of health status, mobility, dexterity, fatigue, and other RA specific symptoms were extracted from daily iPhone guided tests (GT), as well as actigraphy and heart rate sensor data, which was passively recorded from patients' Apple smartwatch continuously over the study duration. We subsequently developed a machine learning (ML) framework to distinguish RA status and to estimate RA severity. It was found that daily wearable sensor-outcomes robustly distinguished RA from HC participants (F1, 0.807). Furthermore, by day 7 of the study (half-way), a sufficient volume of data had been collected to reliably capture the characteristics of RA participants. In addition, we observed that the detection of RA severity levels could be improved by augmenting standard patient reported outcomes with sensor-based features (F1, 0.833) in comparison to using PRO assessments alone (F1, 0.759), and that the combination of modalities could reliably measure continuous RA severity, as determined by the clinician-assessed RAPID-3 score at baseline ( $r^2$ , 0.692; RMSE, 1.33). The ability to measure the impact of the disease during daily life—through objective and remote digital outcomes—paves the way forward to enable the development of more patient-centric and personalised measurements for use in RA clinical trials.

npj Digital Medicine (2024)7:33; <https://doi.org/10.1038/s41746-024-01013-y>

## INTRODUCTION

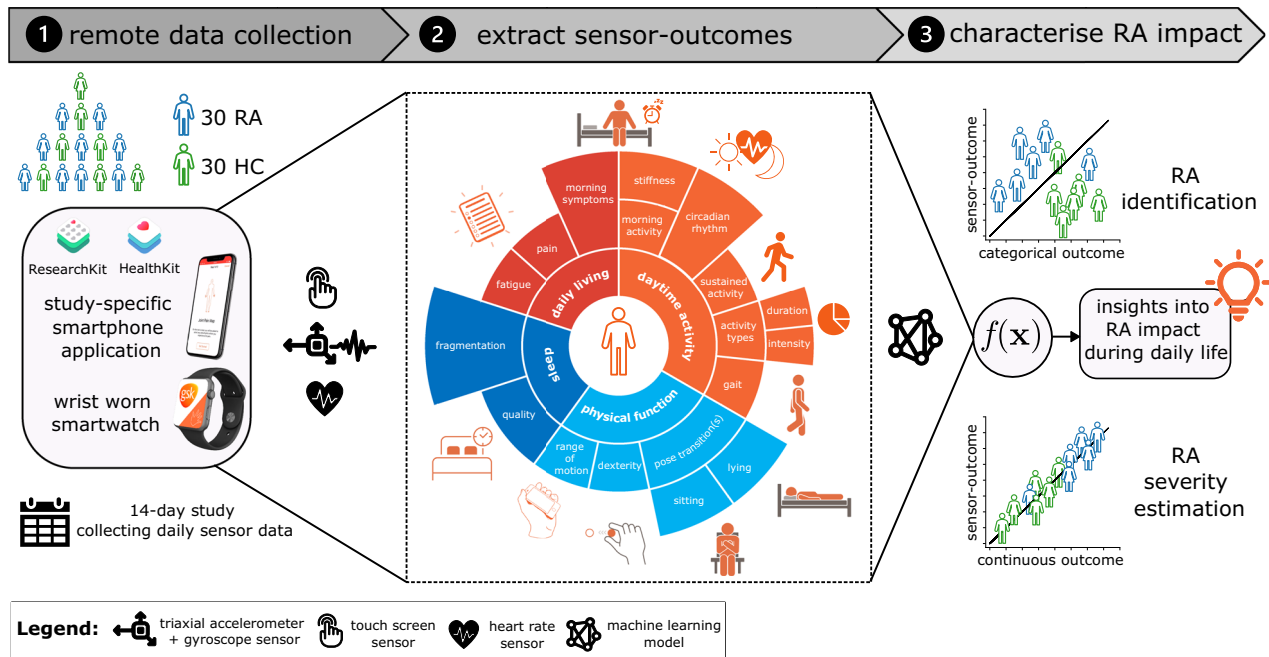
Rheumatoid arthritis (RA) patients follow subtle and unpredictable disease courses, patient-to-patient, with a progressive decline in physical function and quality of life and over time—often leading to disability and difficulty to perform many tasks of daily life<sup>1</sup>. RA symptoms include joint pain or tenderness, joint swelling, morning stiffness, reduction in joint range of movement (ROM), muscle pain, and fatigue<sup>1</sup>. Currently, the gold-standard methods to measure the impact of RA on daily life rely on infrequent clinical visits that may often occur every 3–4 months, with assessments depending on a combination of subjective clinician-determined scores<sup>2</sup> and patient-reported outcomes<sup>3</sup>. These have inherent limitations, however, in that they can be subjective and are prone to recall bias<sup>4,5</sup>. As such, there is a need to objectively measure the impact of RA on daily life<sup>6</sup>, remotely over a continuous period, rather than restricting assessments to only intermittent physician visits. In recent years, consumer-grade mobile applications (app.) and wearable devices have shown promise to objectively measure participants' symptoms during daily life<sup>7</sup>; these digital health technologies (DHT) tools<sup>8</sup> have shown to increase study engagement, improve patient convenience, streamline collection of PROs<sup>9</sup>, and potentially generate more frequent and accurate data that can characterise disease<sup>10</sup>. DHT have been shown to measure RA symptoms and functions, such as range of motion (ROM) and gait-specific metrics during prescribed “active” assessments<sup>11,12</sup>.

Other studies have shown how “passive” wearable actigraphy sensor-outcome measurements capture differences in RA physical activity (PA) in daily life, compared to healthy controls (HC)<sup>13</sup>, as well as to detect flaring of RA symptoms<sup>14</sup>.

However, there remains a lack of sufficient evidence for how DHT can provide objective insights into the impact of therapies for RA, despite progress made in other disease areas<sup>15–22</sup>. Particularly, the benefit of sensor-outcomes generated from prescribed active assessments compared with passive monitoring has not yet been explored together. While digitised patient-reported outcomes (PROs) enable a patient the ability to regularly record their “subjective” experience of disease activity in remote settings<sup>23</sup>, it remains unclear how “objective” sensor-outcomes could provide additional insights that can augment PROs to better characterise the impact of RA on daily life. As part of this characterisation, the sensitivity of DHT to measure RA symptoms, such as the volume of remote data required and the number of sensor-outcome measurements needed, will also need to be determined. Finally, the application of DHT sensor-outcomes to monitor RA during daily life remains yet to be validated against standard in-clinic administered assessments of RA impact<sup>24</sup>.

In this study, we therefore aimed to investigate how active and passive sensor-based measurements should be combined using machine learning (ML) to distinguish RA status from healthy controls, to augment traditional patient self-reported outcome

<sup>1</sup>Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, UK. <sup>2</sup>Big Data Institute, University of Oxford, Oxford, UK. <sup>3</sup>Value Evidence and Outcomes (VEO), GSK, UK. <sup>4</sup>Nuffield Department of Population Health, University of Oxford, Oxford, UK. <sup>5</sup>Value Evidence and Outcomes (VEO), GSK, US. <sup>6</sup>Analysis Group (AG), Boston, MA, USA. <sup>7</sup>These authors jointly supervised this work: Aiden Doherty, Luis Garcia-Gancedo, David A. Clifton. <sup>✉</sup>email: [andrew.creagh@eng.ox.ac.uk](mailto:andrew.creagh@eng.ox.ac.uk)



**Fig. 1** Illustration detailing the objectives of this study. The wearABLE-PRO 14-day trial aimed to investigate how digital health technologies (DHT)—a wrist-worn Apple smartwatch and an iPhone device, with bespoke mobile apps.—could augment patient reported outcomes (PRO) to characterise the impact of rheumatoid arthritis (RA) during the daily life of 30 moderate-to-severe RA patients, compared to 30 matched healthy controls (HC). We explore the ability of machine learning (ML) models to (1) estimate categorical RA outcomes, such as identifying RA participants from healthy controls and (2) estimate continuous RA outcomes, such as RA severity, using a combination of PRO and sensor-outcomes.

(PRO) data, and to estimate standard in-clinic assessments of RA severity. Our work offers the first comprehensive evaluation of how sensor data captured during daily life can characterise RA status and severity, which represents an important first step towards the development of more sensitive and patient-centric measurements for use in RA clinical trials and real-world studies.

In order to investigate the objectives of this study, we performed the following set of analysis and experiments. We first illustrate the variety of sensor-based measurements that can be extracted from daily prescribed (active) smartphone-based assessments and (passive) smartwatch-based activity monitoring in an RA cohort. In this, we evaluate how smartwatch-based daily physical activity patterns can be remotely estimated using our bespoke deep convolutional neural (DCNN), pre-trained using multi-task self-supervised learning (SSL) on a large-scale open-source cohort. We next assess the ability of our sensor-based measurements to identify RA status from healthy controls and to distinguish RA severity levels. As part of our analysis, we also explore the volume of days and number of sensor-outcomes required to remotely distinguish RA status. Finally, we investigated the power of active and passive sensor-outcomes to augment routinely collected patient self-reported outcome (PRO) data to estimate RA severity—as measured by standard in-clinic assessments of RA, such as the RAPID-3<sup>25</sup>.

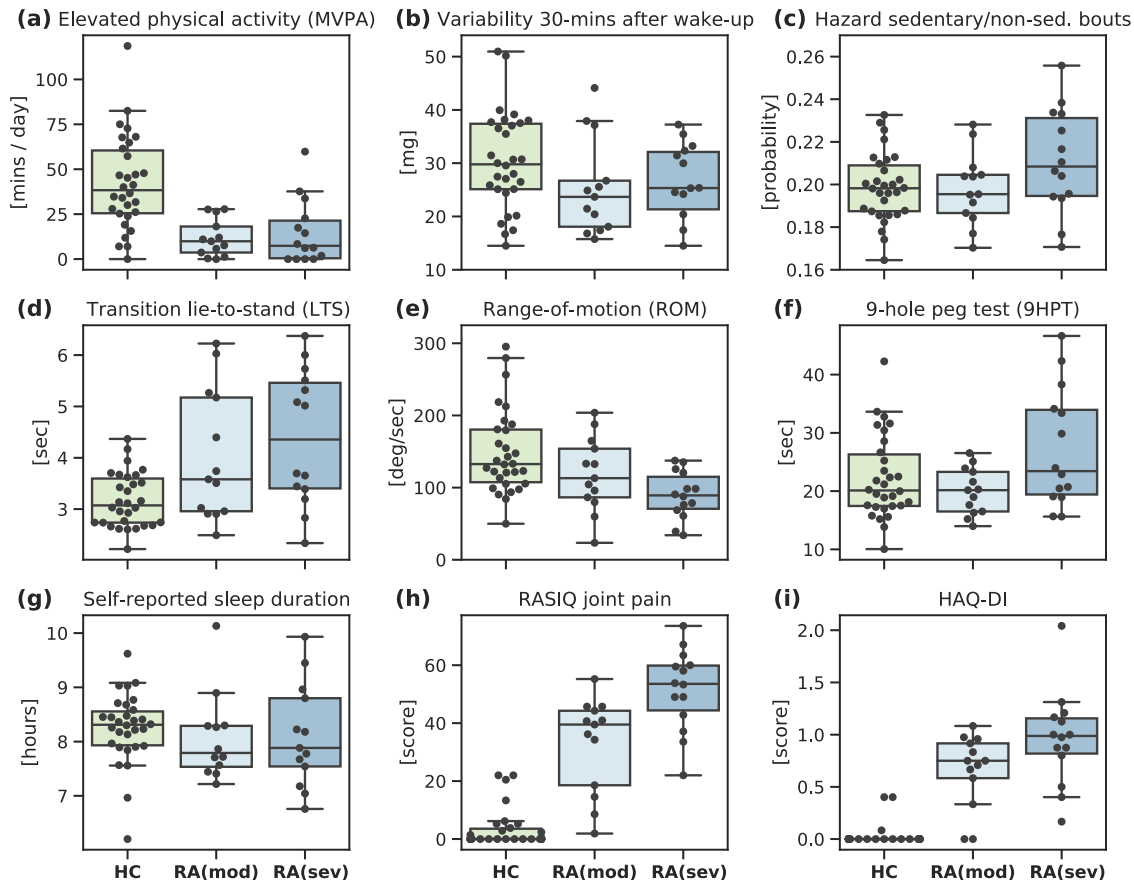
## RESULTS

The GSK wearABLE-PRO study (GSK212295) was a 14-day observational study which investigated how DHT tools could objectively measure the impact of RA on participants' daily lives. Digital wearable devices—a wrist-worn Apple Watch for passive monitoring and an iPhone, integrated with a bespoke mobile app. which prescribed daily guided assessments—collected high-frequency, objective sensor data in 30 RA patients and 30 matched Healthy Controls (HCs). Figure 1 provides an illustrative overview of the objectives of this study. Sensor-based measures of

physical function, mobility, dexterity, and other RA specific symptoms were extracted from daily prescribed (active) iPhone guided tests using a combination of bespoke algorithms and proprietary algorithms developed by Apple ResearchKit, for instance, a wrist-range of motion exercise, a walking assessment, a nine-hole peg test, as well as two pose transition-based mobility exercises, lie-to-stand (LTS) and sit-to-stand (STS). In addition, continuous (passive) actigraphy was recorded from participants' Apple smartwatch over the study duration in order to characterise daily activity patterns and sleep. In order to illustrate the various characteristics of RA we are interested in assessing, we have grouped measurements in Fig. 1 into four domains: physical function, daytime activity, daily living, and sleep; denoting particular types of measurements which may attribute to each domain. Note: this manuscript details a sub-study of wearABLE-PRO; trial design, feasibility, participant adherence, and other primary related study outcomes are reported in Hamy et al.<sup>26</sup>. Two RA participants withdrew immediately after enrolling in the study. Data from these participants were not collected, leaving 28 RA participants, 28 matched HCs, and 2 unmatched HCs for a total of 58 participant

### Assessing smartwatch-based daily physical activity patterns

The daily physical activity of RA participants and healthy controls were estimated with a deep convolutional neural network (DCNN) that was first pre-trained on 100,000 participants in the publicly available UK Biobank, following a multi-task self-supervised learning (SSL) methodology<sup>27</sup>, which was subsequently fine-tuned on the free-living Capture-24 dataset<sup>28</sup> of < 150 participants to determine broad activity patterns of interest {sleep, sedentary, light physical activity, moderate-to-vigorous physical activity (MVPA)}<sup>29,30</sup> and fine-grained activity prediction labels {sleep, sitting/standing, mixed, vehicle, walking, bicycling}<sup>28</sup>. In this study, we build upon our previous work by adding a temporal dependency to the “DCNN (SSL)” through a hidden markov model



**Fig. 2 Ability of individual sensor-outcomes to distinguish between RA status and RA severity levels.** Comparison of the average feature distributions per participants, between healthy controls (HC), RA (moderate) and RA (severe) groups for: **a–c** selection of passively collected smartwatch features; **d–f** selection of guided test collected smartphone features; and **g–i** selection of patient self-reported outcomes recorded on the smartphone application. For all examples shown, medians were significantly different between HC and RA groups: One-way ANOVA determined from the Kruskal-Wallis H-test,  $p < 0.001$ . deg degrees, HAQ-DI Health Assessment Questionnaire-Disability Index, min minutes, mg milli-gravity acceleration units, MVPA moderate-to-vigorous physical activity, RASIQ GSK RA symptom and impact questionnaire, sed sedentary, sec seconds.

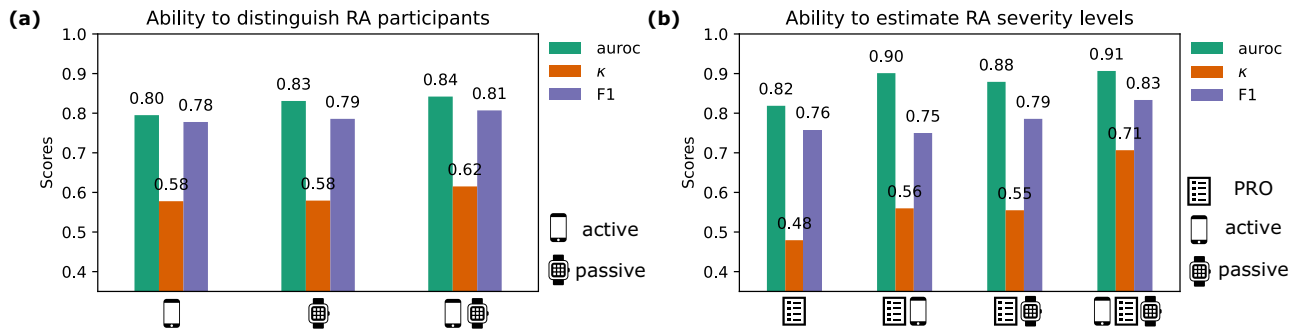
(HMM), which was appended to obtain a more accurate sequence of predicted activities over the continuous study period. It was found that the “DCNN (SSL) + HMM” improved broad activity estimation in Capture-24 ( $\kappa$ ,  $0.862 \pm 0.088$ ; F1,  $0.815 \pm 0.103$ ) as compared to a baseline random forest (RF) + HMM approach ( $\kappa$ ,  $0.813 \pm 0.108$ ; F1,  $0.775 \pm 0.117$ )<sup>28</sup>. Next, the fine-tuned “DCNN (SSL) + HMM” model transformed the raw Apple smartwatch sensor data in wearAble-PRO to determine participants’ daily activity patterns over the 14-day study period, for example, the time spent walking, the frequency of exercise, the length and quality of sleep, and other RA-specific measures, such as morning stiffness. Activity predictions were qualitatively evaluated over the entire RA and HC study population and demonstrated face validity (see Supplementary Figs. 1 and 2 for additional details).

#### Analysis of sensor-outcomes to distinguish RA status and severity levels

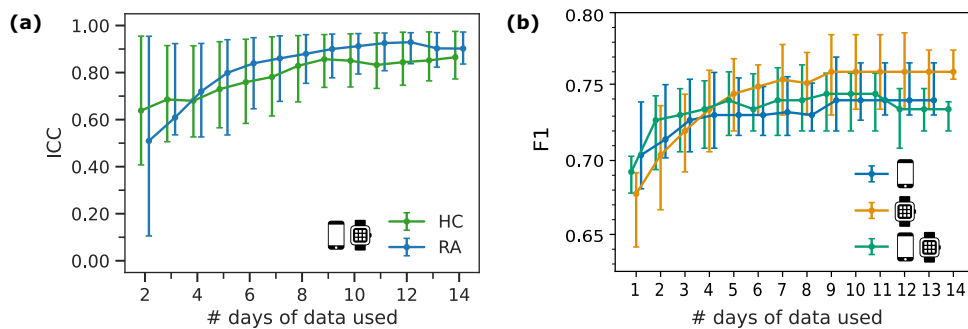
The raw smartphone and smartwatch data recorded during the (active) guided test exercises, and passively during the participants’ daily life, respectively, were summarised as sensor-outcome features. Univariate analysis demonstrated that a total of 153 (93%) sensor-based features (passive,  $n = 131$  (94%); active,  $n = 22$  (88%)) displayed significantly different medians (after post-hoc correction for multiple comparisons) between HC and RA severity groups (Kruskal-Wallis H test,  $p < 0.05$ ). A further 47 (34%) passive features, compared to 6 (24%) active features, were also

significantly different (Mann-Whitney U test,  $p < 0.05$ ) between healthy and RA participants. Figure 2 compares the (fortnightly) average feature distributions between healthy controls (HC), RA (moderate) and RA (severe) participants for a selection of examples of passively collected smartwatch features (Fig. 2a–c) and active guided test sensor features (Fig. 2d–f) and a selection of patient self-reported outcomes recorded on the smartphone application (Fig. 2g–i).

In order to explore the ability of many wearable sensor-outcomes to distinguish symptoms of RA from otherwise healthy individuals, and therefore measure the impact of RA during daily life, we devised a number of multivariate classification-based experiments. First, we investigated the performance of regularised logistic regression (LR) to differentiate RA participants from healthy controls using both passively collected activity monitoring features and guided test exercise features. Comparing model performance between sources (Fig. 3a), passive activity monitoring-based sensor features better distinguished RA participants using fortnightly averaged features (F1, 0.786) versus active (guided test) features (F1, 0.778). It was found that 12 subjects were misclassified using active-only models and 12 for passive-only, with just 4/12 (33%) of the same subjects incorrectly identified by both sources, 3 of which were the same HC participants. Combining active and passive wearable sensor features yielded in the highest performing models to distinguish RA participants overall, for example, using fortnightly averaged features from both sources (F1, 0.807) (for further expansion of



**Fig. 3 Ability of combined sensor-outcomes to distinguish between RA status and RA severity levels.** Comparison of **a** RA identification (RA vs. HC) performance and **b** RA severity level estimation (RA (mod) vs. RA (sev)), using patient reported outcomes (PRO) and combined PRO (list icon), active (smartphone icon), and passive (smartwatch icon) sensor-based outcomes in the wearAble-PRO study. auroc area under the receiver operator curve,  $\kappa$  Cohen's Kappa statistic, F<sub>1</sub> macro-F1 score.



**Fig. 4 The number of days of sensor-data required to remotely characterise RA impact.** Comparison of **a** the minimal amount of days of data needed to distinguish RA status, as measured by the F<sub>1</sub> score across 5-fold cross validation (CV), between active (smartphone icon), passive (smartwatch icon), and combined (smartphone & smartwatch icons) feature sources; **b** the feature (test-retest) reliability, as measured by the intraclass correlation coefficient (ICC), between RA participants and HC across the study duration (14 days); F<sub>1</sub> scores and ICCs suggest that model performance and feature reliability stabilises once more than 7 days of data are used per participant.

results, see Supplementary Table 4). It should also be noted that linear logistic regression was found to perform comparatively to non-linear ensembles of decision trees, a Random Forest (RF) model and Extreme Gradient Boosted Trees (XGB)—as such this work subsequently opted to explore simple linear models for further analysis (see Supplementary Table 5).

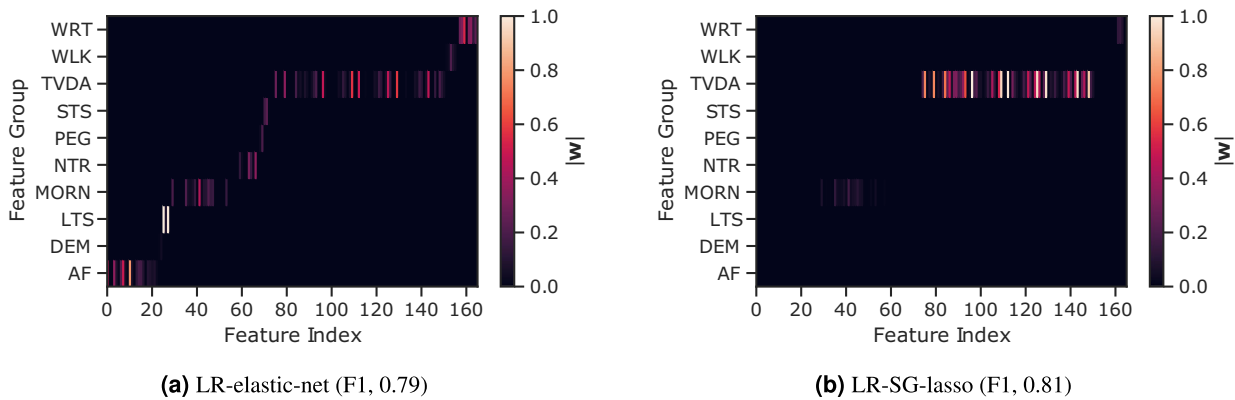
This study next investigated the ability of multiple sensor-based outcomes to augment PRO data in order to stratify RA severity levels. In wearAble-PRO, participants were denoted as having moderate or severe RA based on baseline clinician-assessed RAPID-3 scores. Following similar procedure to RA identification, LR regularised models were investigated in order to distinguish RA (mod) and RA (sev) as binary classification tasks using fortnightly averaged study data. The benefit of incorporating additional sensor-based outcomes to patient (self-) reported outcomes is presented in Fig. 3b (expanded in Supplementary Table 6). It was observed that the linear combination of PRO assessments could accurately stratify RA symptom severity (F<sub>1</sub>, 0.759). The fusion of PRO data and sensor-based outcomes improved RA severity level estimation further with the addition of active (F<sub>1</sub>, 0.750) or passive (F<sub>1</sub>, 0.786) sources. Finally, the amalgamation of PRO outcomes with both active and passive sensor-based outcomes resulted in the most accurate RA severity level estimation (F<sub>1</sub>, 0.833)—an improvement of 10% compared to PRO outcomes alone (Fig. 3b). For additional information on the selected PRO + sensor-outcomes, we refer the reader to Supplementary Table 3.

#### Estimating the volume of days and number of sensor-outcomes required to remotely distinguish RA status

In wearAble-PRO, participants performed daily guided test exercises—resulting in daily sensor features—and continuously

recorded Apple Watch sensor data were summarised as daily activity monitoring-based features, over the 14-day study period. In this work, we aimed to determine the minimal number of days of sensor data required to build a stable and robust estimate of disease status in RA participants compared to HC over the 14-day study period. Figure 4a represents an experiment exploring the (observation-wise) out-of-sample RA classification performance as a function of varying the number of non-contiguous days of data that are averaged per participant. Evaluated over 500 randomly sampled permutations of non-contiguous days, results (median + IQR) indicated that RA prediction stabilised once more than 7 non-contiguous days of data were used per participant. Furthermore, we found that averaging daily feature values over weekly and fortnightly periods improved model performance. However, it was observed that model performance using weekly-averaged features was often similar to fortnightly averaged (we also refer the reader to Supplementary Table 4).

To investigate feature consistency and reproducibility, the intraclass correlation coefficient (ICC) for each feature was evaluated over the study duration (14 days). ICCs were calculated for each feature using  $n = [2, 3, \dots, 14]$  days of data per participant, individually for HC and RA participants. Higher ICCs suggest a high degree of similarity on the performance of each task over the course of the study, and lower coefficients mean that participants tended to perform the task differently each day of the study. ICCs for HCs ranged from 0.582 to 0.854, while those for RA participants ranged from 0.424 to 0.897. Figure 4b depicts the median + interquartile range (IQR) of ICC values for the LR-elastic net retained active + passive features. Intra-rater reliability analyses suggest that feature reliability stabilises to good (ICC=0.75–0.9) and



**Fig. 5** The number of sensor-outcomes required to remotely distinguish RA status. Comparison of features selected between regularised logistic regression (LR) models for: **a** elastic-net (F1, 0.79) and **b** SG-lasso (F1, 0.81). The SG-lasso promotes group-wise sparsity (i.e., regularising the number of feature domains) and within-group sparsity (i.e., regularising the number of features per domain), achieving a similar performance to LR elastic-net, while selecting a fewer number of domains and features. Feature importance, denoted as the mean LR coefficient value ( $w$ ) over cross-validation, are illustrated by colour intensity. Feature domains: AF activity fragmentation, DEM demographics, LTS lie-to-stand assessment, MORN morning stiffness, NTR night-time restlessness, PEG 9-hole peg test, STS sit-to-stand assessment, TVDA total volume of daytime activity, WLK walking assessment, WRT wrist assessment.

excellent ( $ICC > 0.9$ ) once more than 7 contiguous days of data were used per participant.

In order to evaluate the number of sensor-outcomes required to remotely distinguish RA status, we compared various feature regularisation techniques, lasso ( $\ell_1$ ), ridge ( $\ell_2$ ), elastic-net ( $\ell_1 + \ell_2$ ), and sparse-group lasso, using fortnightly (i.e., study duration) averaged features. It was found that introducing sparsity through regularisation improved classification performance. In addition, active and passively recorded sensor-based features could be grouped into domains, based on the guided test they were extracted from, or the perceived functional domain of daily activity they were assumed to assess. Introducing group-wise sparsity with the sparse-group lasso (SG-lasso), regularising on the number of groups (i.e., the feature domains) and the coefficients within each group, resulted in the highest RA participant identification performance (F1, 0.807), compared to lasso ( $\ell_1$ , F1, 0.772), ridge ( $\ell_2$ , F1, 0.792), and elastic net ( $\ell_1 + \ell_2$ , F1, 0.792) regularisation (for expansion of results, see Supplementary Table 5). The features and groups selected by each regularisation technique are illustrated in Fig. 5, represented as the mean LR coefficient value  $w$  over CV per each feature and feature domain (coefficient values have been normalised between 0 and 1 to benefit comparison between models). Examining the feature sparsity of elastic-net ( $\ell_1 + \ell_2$ ) (Fig. 5a), it was observed that features from multiple domains were selected. In contrast, the SG-lasso, as shown in Fig. 5b, selected mostly passive activity-based smartwatch features—TVDA with some morning stiffness measures—to distinguish RA status. Group sparsity penalised simultaneously selecting from multiple feature domains, where within group-sparsity regularised the feature coefficient values within the selected domains. Using fewer domains and less features, the SG-lasso was able achieve similar performance to LR elastic-net, even marginally improving performance (F1, 0.807). For further details on the features extracted, and selected, we refer the reader to the Supplementary Methods.

#### Estimating in-clinic RA severity scores from PRO and sensor-based outcomes

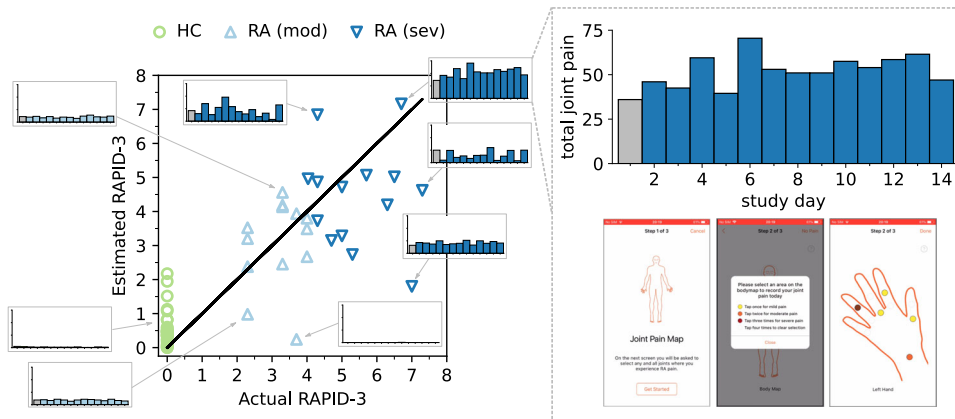
Rheumatoid arthritis severity levels were denoted by a clinician administered RAPID-3 assessment<sup>25</sup> at baseline in the wearABLE-PRO study. The RAPID-3—a “rapid” and easy to administer questionnaire—is also validated against more exhaustive assessments for RA, such as the disease activity score 28 (DAS28) and clinical disease activity index (CDAI) in clinical trials and clinical care<sup>25</sup>. In this work, we aimed to establish how the combination of

PRO and sensor-based outcomes could stratify continuous RAPID-3 RA severity. Note: HC subjects were assigned a RAPID-3 score of zero at baseline. Through multivariate modelling, using LR elastic-net, it was determined that PRO and sensor-based features could accurately estimate RAPID-3 scores to within 1 point ( $r^2$ , 0.69; MAE, 0.94; RMSE, 1.33), an improvement compared to using PRO measures alone ( $r^2$ , 0.63; MAE, 1.16; RMSE, 1.45). The association between actual and PRO + sensor-outcome estimated RAPID-3 scores was found to be good-to-excellent ( $r > 0.75$ ), Pearson’s  $r = 0.60$ ,  $p < 0.001$ ; Spearman’s  $\rho = 0.83$ ,  $p < 0.001$ .

Participants in wearABLE-PRO were also administered a twice-daily interactive Joint Pain Map (JMAP) questionnaire on their iPhone<sup>11</sup>, in order to more precisely record and localise perceived pain. Participant model-estimated RAPID-3 scores were further interpreted through detailed inspection of the daily smartphone-based patient-reported joint pain map (JMAP) total scores—an external validation measure, which was not included as a predictor in the model—as expanded in Fig. 6. The JMAP score, defined as the sum of all individual joint pain scores per recording, was intended as a coarse measure to holistically capture participants’ overall level of perceived pain, in addition to validated PRO assessments. Higher JMAP scores indicate higher levels of pain experienced. It was observed that RAPID-3 estimations were reliable and robust, in that they faithfully characterised RA participant’s perceived level of symptoms, through the JMAP. For example, in Fig. 6, the RA (sev.) participant with consistently the largest reported degree of pain across the 14-day study exhibited the highest actual RAPID-3 score (6.7), which was closely estimated by the model at 7.1. JMAP scores further enabled additional explanation of model performance, especially with respect to RAPID-3 estimations that were not reflective of actual RAPID-3 scores. For instance, the RA (mod) participant with the lowest estimated RAPID-3 score (0.2) actually reported zero pain experienced over the 14-day study duration, despite a RAPID-3 assignment of 3.7 at baseline. Non-zero estimated RAPID-3 scores for some HC could also often be contextualised, due to these participants frequently self-reporting low-levels of pain in their JMAP (i.e., non-zero JMAP entries) over the study period, despite being healthy. As such, it was determined that PRO and sensor-based RAPID-3 estimates could reliably reflect participant’s RA symptoms over the study.

#### DISCUSSION

Our findings in the wearABLE-PRO study demonstrate how digital health technology (DHT) captured sensor-outcomes, recorded



**Fig. 6 The ability of remote PRO + sensor-outcomes to estimate in-clinic determined RA severity scores.** Scatter plot of baseline RAPID-3 scores  $y$  versus predicted  $\hat{y}$  scores per subject, using elastic net with PRO + sensor-outcomes, over cross-validation (CV). Participant model-estimated RAPID-3 scores can be further interpreted through detailed inspection of the daily smartphone-based patient-reported joint pain map (JMAP) total scores—which was not included as a predictor in the model. Higher JMAP scores indicate higher levels of pain experienced. Additional interpretability, through the JMAP, demonstrated that PRO + sensor-based outcome estimation of the RAPID-3 could reliably reflect patient’s perceived daily RA symptoms. Note: Baseline JMAP total scores, recorded on the same day as the baseline RAPID-3, are denoted in grey; the JMAP y-axis scale is the same among all subplots. HC subjects were assigned a RAPID-3 score of zero at baseline. A black line represents perfect predictions ( $r^2$ , 0.692; MAE, 0.938; RMSE, 1.333).

from smartphone-based active tests, and continuously collected passive smartwatch-based monitoring, could characterise meaningful aspects of rheumatoid arthritis (RA) impairment and physical function impacting daily life. Remotely collected wearable sensor-outcomes could distinguish RA status from healthy controls—demonstrating further improved performance when combining the sensor-data from both devices—and how objective sensor-outcomes could augment patient (self-) reported outcomes to remotely estimate RA severity. Furthermore, by the half-way point of the wearAble-PRO study (day 7), a sufficient volume of data had already been collected to reliably distinguish the characteristics of RA participants. This work provides the first comprehensive evaluation how remote and objective digital sensor-outcomes enrich our ability to understand the impact of RA on daily life between clinical visits.

In this work, we detailed how raw data collected from smartphone and smartwatch sensors can be transformed into sensor-based outcomes that are reflective of disease status. In concurrence with previous studies, many remotely collected smartphone sensor-outcomes distinguished RA participants and RA severity levels. For example, it was observed that joint ROM features differentiated HC and RA groups—a similar finding to our previous work<sup>12</sup>—and that RA participants were less mobile, taking longer to move between positions (as measured during the lie-to-stand exercise)—as previously shown by Andreu-Perez et al.<sup>31</sup>. Continuously collected smartwatch sensor data, known as passive monitoring, allowed the measurement of aspects of RA daily life, such as physical activity, sleep, and other RA specific symptoms, such as morning stiffness, or night-time restlessness. In this study we trained an activity recognition model on the free-living capture-24 dataset to estimate daily activity patterns in the wearable-pro population. Leveraging the latest advances in self-supervised learning (SSL) allowed our model to be pre-trained on 100,000 participants with 700,000 days of diverse, unlabelled wearable sensor data in the uk biobank<sup>27</sup>, which combined with HMM temporal smoothing, significantly improved activity prediction compared to our previous established RF-HMM based methods<sup>28,30</sup>. Our SSL DCNN+HMM model enabled a more robust and fine-grained estimation of daily activity patterns beyond traditional acceleration magnitude levels<sup>13,14</sup>, which we proposed could allow a richer characterisation of PA and sleep in RA activity monitoring revealed distinct differences distinguishing RA status, for example the daily percent of the day in moderate-to-vigorous

physical activity, and similar features, were significantly lower in the RA population compared to healthy controls—a similar finding by Prioreshi et al.<sup>13</sup>, and an observation people with RA regularly self-report<sup>32</sup>. Other specific RA symptom measurements, like morning stiffness or disrupted sleep, were evident in certain RA participants. For example, the mean acceleration value  $> 30$  [mins] after wake-up were lower in RA—also a similar finding to Keogh et al.<sup>33</sup>—or that the number of movement episodes during night-time sleep distinguished some specific RA participants. We also observed that after collecting 7 days of sensor-data in the wearAble-PRO study, a sufficient volume of data had already been recorded to reliably distinguish RA participants from a healthy population; participant feature reliability (as measured ICC values) stabilised at good-to-excellent levels, maximal identification performance of RA participants plateaued, and that there was no additional benefit to averaging over a fortnight’s worth of data versus a week. Therefore it is recommended that considering at least one week’s worth of sensor data is collected, it might be more beneficial to gather less data from a greater number of participants, rather than greater duration of sensor data from the same participants.

Our work is the first study to combine active smartphone and passive wearable measurements to distinguish RA status and measure variations in RA severity. While models trained on only passive features tended to marginally outperform models trained solely on active guided test features, combining both active + passive features led to the best performance in RA identification for all models investigated. Interestingly, it was found that different subjects were misclassified by active versus passive models. For example, 12 subjects were misclassified using active-only models and 12 for passive-only, with just 4/12 (33%) of the same subjects incorrectly identified by both sources, 3 of which were the same HC participants. In addition, further experiments with the LR-SG-lasso determined that only activity monitoring domain features were mainly needed in order to distinguish RA participants from health controls. This indicates that we sometimes do not need to prescribe all guided test assessments, or to parse all activity feature domains, but that a small number of prescribed assessments can be sufficient to characterise RA status. For example, including only the lie-to-stand assessment rather than also prescribing the similar, and highly correlated, sit-to-stand assessment in future studies; or removing the prescribed walking assessment (shown to have little predictive value in the

weARable-PRO study), and using passive daily life walking predictions generated from the activity recognition model instead, which could reduce patient burden. Finally, we also found that combining patient-reported outcomes (PRO) and objective sensor-outcomes could better capture RAPID-3-based RA severity at baseline than PROs alone; most estimated RAPID-3 scores correctly stratified participants across severity levels from healthy to moderate to severe RA, suggesting that sufficient information to characterise RA disease severity could be reflected in the remote monitoring outcomes derived in the 14-day weARable-PRO study. To the best of the authors knowledge, this offers the first evaluation and insight how remote monitoring outcomes in daily life can estimate in-clinic administered assessments of RA impact.

There are a number of limitations that must be considered in the weARable-PRO study. Despite rich individual level measurements, the study recruited a relatively small sample size (HC,  $n = 30$ ; RA,  $n = 30$ ). As such, a degree of variability and uncertainty existed in constructing cross-validated models to distinguish RA participants, RA severity levels, or estimate the in-clinic RAPID-3 assessment. Extrapolation of results aimed at generalising RA is therefore not possible without the availability of larger cohorts and further external validation. In addition, this study only recruited RA patients with moderate-to-severe levels of disease activity; future studies should also aim to characterise patients with lower levels of disease activity or those in remission. There were also limitations associated with modelling a clinician-administered assessment, or clinical labels formulated from in-clinic assessments. For instance, the RAPID-3 was assessed at baseline, with participants recalling the prior week, yet the PRO and sensor-based features were calculated as averages over subsequent 14-day trial period from baseline. As such, the baseline RAPID-3 may not have precisely reflected the participant's disease status recorded earlier, due to the underlying mutability and heterogeneity of RA symptoms over short periods of time. The subjectivity of PRO predictors should also be considered, for instance, pain or perceived quality of sleep is relative, and some healthy participants recorded experiencing pain or affected sleep in PRO questionnaires. As a result, some PRO values influenced HC RAPID-3 predictions greater than zero, i.e., indicating the presence of RA symptoms—albeit non-zero estimated RAPID-3 predictions for HCs were generally low ( $< 2$ ).

The weARable-PRO study typifies how continuously collected patient self-reported and sensor-based outcomes may more closely reflect participant perceived and experienced symptoms that impact daily life. While in-clinic assessments are considered the gold-standard means of assessing disease severity in RA, it is clear that remotely collected, continuous, patient-centric measurements generated from PRO and sensor-based outcomes offer promising insights that can undoubtedly augment in-clinic assessments for RA. We believe that our work—the first comprehensive evaluation how remote sensor data can augment traditional PRO measures to estimate clinician-determined RA severity—helps inform future DHT study design to better characterise the impact of RA on daily life, ultimately to expand the use of DHT to develop more sensitive, and patient-centric, endpoints in RA clinical trials and real-world studies.

## METHODS

### Dataset

Remotely collected smartphone and smartwatch sensor data was obtained from the GSK study title: Novel Digital Technologies for the Assessment of Objective Measures and Patient Reported Outcomes in Rheumatoid Arthritis Patients: A Pilot Study Using a Wrist-Worn Device and Bespoke Mobile App. (212295, weARable-PRO)<sup>26</sup>. This observational study followed 30 participants

**Table 1.** Population demographics, in-clinic, and selected patient self-reported outcomes, as assessed at baseline, where the mean  $\pm$  standard deviation across the population is reported.

	HC <sup>a</sup> ( $n = 28$ )	RA (mod) <sup>b</sup> ( $n = 13$ )	RA (sev) <sup>c</sup> ( $n = 15$ )	$p^1$
<b>Demographics</b>				
Age, years	58.4 $\pm$ 9.9	56.9 $\pm$ 11.4	60.4 $\pm$ 7.1	0.33
Female, n (%)	25 (89%)	11 (84%)	14 (93%)	0.92
BMI	25.8 $\pm$ 4.6	31.1 $\pm$ 5.9	31.7 $\pm$ 8.6	0.96
<b>In-clinic Outcome(s)</b>				
RAPID-3	0 $\pm$ 0	3.2 $\pm$ 0.7	5.3 $\pm$ 1.1	< 0.001
<b>Patient Reported Outcome(s)</b>				
HAQ-DI	0 $\pm$ 0	0.63 $\pm$ 0.36	1.03 $\pm$ 0.42	< 0.01
RASIQ-pain	3.1 $\pm$ 6.7	32.1 $\pm$ 20.8	56.2 $\pm$ 11.6	< 0.01
RASIQ-stiffness	5.9 $\pm$ 9.5	33.9 $\pm$ 18.9	51.6 $\pm$ 10.2	< 0.05
RASIQ-impact	47.3 $\pm$ 5.0	53.9 $\pm$ 5.1	50.8 $\pm$ 7.6	0.33
FACIT	49.2 $\pm$ 2.9	38.9 $\pm$ 4.3	31.9 $\pm$ 7.6	< 0.05
PROMIS-sleep	49.6 $\pm$ 2.8	52.7 $\pm$ 4.2	52.4 $\pm$ 4.3	0.83
PROMIS-pain	42.2 $\pm$ 4.8	54.2 $\pm$ 7.29	58.8 $\pm$ 4.6	0.09
JMAP total pain <sup>2</sup>	0.20 $\pm$ 0.5	13.5 $\pm$ 13.9	18.8 $\pm$ 13.7	0.23

<sup>1</sup> $p$ ,  $p$ -value calculated from Mann Whitney U-test comparing severe vs. moderate RA participants;

<sup>2</sup>Note: self-reported JMAP is not a validated PRO in RA;

<sup>a</sup>Matched HC to RA participants only;

<sup>b</sup>RA participants with baseline RAPID-3: 6.1–12.

<sup>c</sup>RA participants with baseline RAPID-3: > 12;

diagnosed with moderate-to-severe RA and 30 matched HCs over 14 days. The population demographics, in-clinic, and relevant patient self-reported outcomes, as assessed at baseline, are reported in Table 1. RA participants were denoted as displaying moderate disability, RA (mod), or severe disability, RA (sev), as determined by their baseline RAPID-3 score. Note: Two RA participants withdrew immediately after enrolling in the study. Data from these participants were not collected, leaving 28 RA participants, 28 matched HCs, and 2 unmatched HCs for a total of 58 participants. All study information, informed consent, study questions and instructions for conducting the guided tests were first drafted in the form of a survey instrument. The survey instrument was then programmed into the mobile app. All documentation including the study protocol, any amendments, and informed consent procedures, were reviewed and approved by Reliant Medical Group's IRB. All participants provided written informed consent before any study procedures were undertaken. The study was conducted in accordance with the International Committee for Harmonisation principles of Good Clinical Practice and the Declaration of Helsinki. We refer the reader to Hamy et al.<sup>26</sup> for further study details. In addition, participant requirement and data collection are outlined in the accompanying Supplementary Methods material.

**Sensor-based data collection.** The Apple Watch and iPhone were used to collect high frequency raw sensor data from predefined, (active) guided tests on a daily basis. Participants were prescribed daily to perform five iPhone-based assessments: WRT, a wrist range of motion (ROM) exercise<sup>12</sup>; WLK, a 30-second walking exercise<sup>12</sup>; PEG, a digital 9-hole peg test<sup>34</sup>; STS, a sit-to-stand transition exercise<sup>31,35</sup>; and LTS, a lie-to-stand transition exercise<sup>31,35</sup>. A brief overview of the guided tests prescribed in weARable-PRO are presented in Supplementary Table 8. In addition, the Apple Watch was used to continuously collect background sensor data (denoted passive data), as the

participants went about their daily activities. Participants were asked to maintain a charge on both the Apple Watch and the iPhone, so that interruptions to monitoring and data transfer were kept to a minimum. Since night-time activity was also monitored, while participants were asleep, it was requested that charging should be done during the day, in a way that fit the participants' schedules (e.g., charging in the morning while getting ready for the day). For more details on the activity monitoring features, see Supplementary Table 9.

**Patient-reported outcomes.** Patient-reported outcomes (PRO), most often self-report questionnaires, were administered to assess disease activity, symptoms, and health status and quality of life from the patients' perspective<sup>36,37</sup>. The wearAble-PRO study administered a selection of validated PRO measures for RA in complement to bespoke digital PRO assessments—that are validated in clinical trials, where the questions, response options, and the general approach to assessment were standardised for all participants. PROs were recorded on days 1, 7, and 14 of data collection. The PRO assessments administered to participants are outlined in Supplementary Table 7.

### Smartwatch-based estimation of daily life patterns

In order to generate unobtrusive measures characterising physical activity and sleep in RA participants during daily life, the raw Apple Watch actigraphy (i.e., accelerometer) sensor data was transformed through a human activity recognition (HAR) sensor processing and deep convolutional neural network (DCNN) pipeline. Figure 7 illustrates how a deep convolutional neural network (DCNN) can transform raw Apple smartwatch sensor data to estimate a participant's daily activity patterns in the wearAble-PRO study using self-supervised learning (SSL). The construction of this pipeline yielded unobtrusively measured summary features of physical activity and sleep for RA participants, computed daily during normal life.

A deep convolutional neural network (DCNN) with a ResNet-V2 architecture was first pre-trained following a multi-task self-supervised learning (SSL) methodology on 100,000 participants, each participant contributing 7 days yielding roughly 700,000 person days of data, in the open-source UK biobank<sup>27</sup>. The SSL pre-trained model was then fine-tuned to perform activity recognition as a downstream task in the Capture-24 dataset.

The Capture-24 study is a manually labelled, free-living dataset—that is reflective of real-world environments—and is available for training an activity recognition model to be applied to the wearAble-PRO study. In Capture-24, actigraphy data was collected for 24-h from 132 healthy volunteer participants with a Axivity AX3 wrist-worn device as they went their normal day. Activity labels provided by photographs automatically captured roughly every 30 seconds by a wearable camera for each participant. Capture-24 was labelled with 213 activity labels, standardised from the compendium of physical activities<sup>29</sup>. Activity labels were then summarised into a small number of free-living behaviour labels, defining activity classes in Capture-24.

There are two major labelling conventions used within Capture-24 that the model was trained to predict, defined as broad activity: {sleep, sedentary, light physical activity, moderate-to-vigorous physical activity (MVPA)}<sup>29,30</sup>, and fine-grained activity: {sleep, sitting/standing, mixed, vehicle, walking, bicycling}<sup>28</sup>.

HAR model predictions are essentially independent—meaning that the sequence of activities over each 30 s epoch incorporates no temporal information epoch-to-epoch, for instance how the previous epoch prediction affects the current, or next, activity prediction. In order to add temporal dependency to the “DCNN (SSL)” model, a Hidden Markov Model (HMM) was implemented in a post-processing step to obtain a more accurate sequence of predicted activities over the continuous 14-day data collection period as per Willetts, et al.<sup>28</sup>.

This Capture-24 fine-tuned “DCNN (SSL) + HMM” model was then implemented to estimate daily activities in wearAble-PRO study data. For additional information of the HAR deep network, SSL, and other related information, we refer the reader to our previous work<sup>27</sup>. Further results relating to the “DCNN (SSL)” models are outlined in the Supplementary Table 1. The sensor processing pipeline developed for the Apple Watch in the wearAble-PRO study is outlined in Supplementary Fig. 5 and within the accompanying Supplementary Methods.

### Extraction of sensor-based outcomes

Wearable sensor-based features were derived from the smartphone during the active guided tasks and passively from the smartwatch during daily life. “Active” features, extracted from smartphone sensor-based measurements during the prescribed guided tests, aimed to capture specific aspects of RA physical function, related to pain, dexterity, mobility and fatigue<sup>12</sup>. In addition “passive” features were extracted from smartwatch sensor-based measurements, collected continuously in the background over the 14-day period. Daily activity predictions from the ML SSL model were summarised into general features measuring activity levels, period, duration and type of activity, as well as sleep detection and sleeping patterns. Furthermore, devised under the guidance of Rheumatologists, additional activity monitoring features specifically aimed at characterising well-known RA symptoms were also developed, such as morning stiffness and night-time restlessness.

The Supplementary Methods also detail algorithms used to extract active and passive features in the wearAble-PRO study. For a full list of extracted sensor-based features in wearAble-PRO, we refer the reader to Supplementary Table 9.

### Statistical analysis

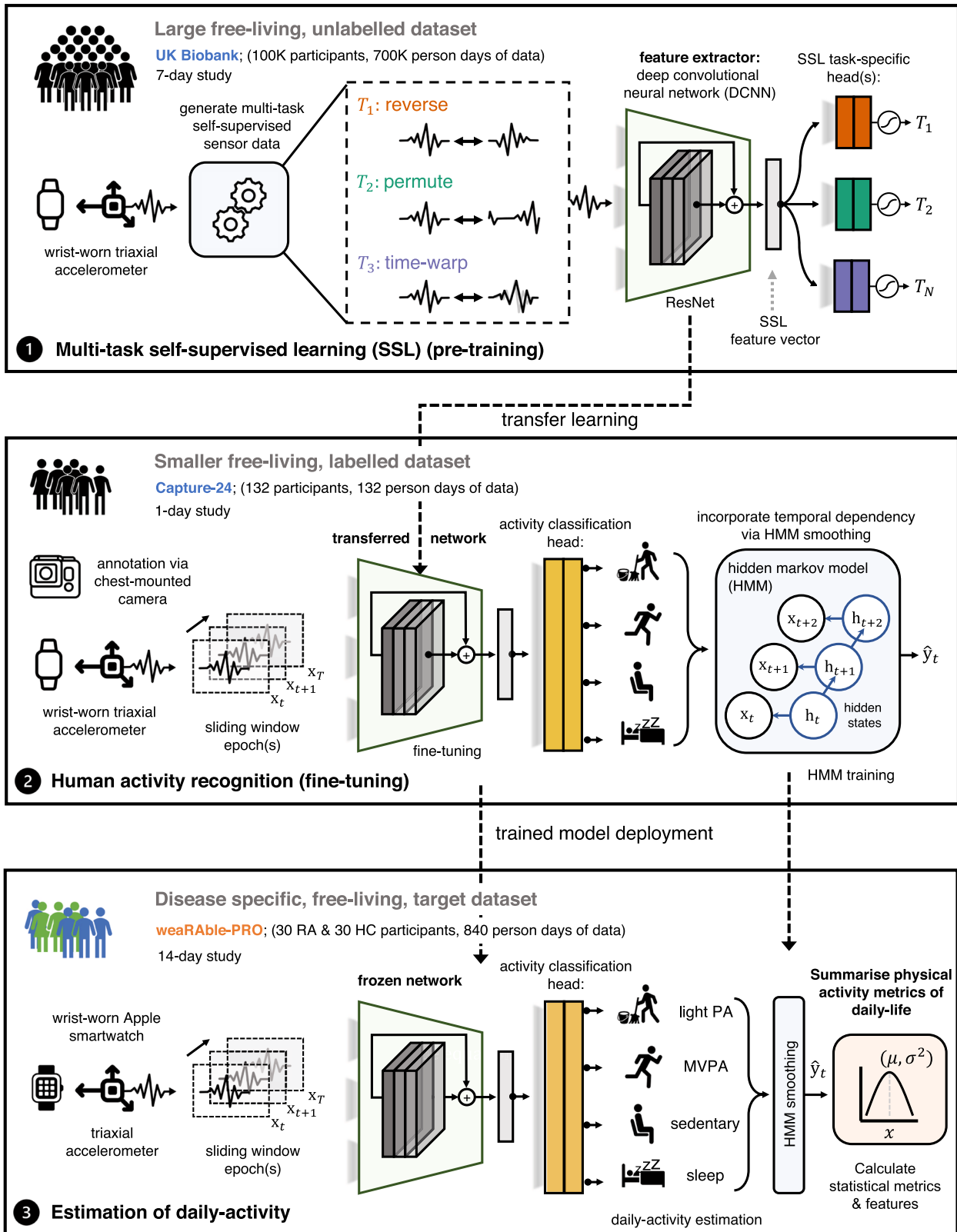
**Univariate testing.** Pair-wise differences groups between groups, for example HC vs. RA, or RA (mod) vs. RA (sev) were analysed for the equality in population median using the non-parametric Mann-Whitney U test (MWUT)<sup>38–40</sup>. One-way analysis of variance (ANOVA) tests were also used to assess differences between medians of multiple groups, for example HC vs. RA (mod) vs. RA (sev) were assessed using the Kruskal-Wallis (KWt) test by ranks<sup>41</sup>. The Brown-Forsythe (BF) test by (absolute deviation) of medians was used to investigate if various groups of data have been drawn with equal variances<sup>42</sup>.

**Correlation analysis.** Correlation analysis was utilised to determine the association or dependence between sets of random variables, such as the dependence between features, or to assess a features' clinical utility by measuring the association to an established clinical metric. This study investigated the (linear) Pearson's  $r$  correlation and the (non-linear) Spearman's Rho  $\rho$  correlation between features, between features and PROs, and between clinical assessments to determine levels of association. The strengths of the correlations were classified as good-to-excellent ( $r > 0.75$ ), moderate-to-good ( $r = 0.50–0.75$ ), fair ( $r = 0.25–0.49$ ) or no correlation ( $r < 0.25$ )<sup>43</sup>.

**Feature reliability.** Intra-rater (i.e., test-retest) reliability was determined using intra-class correlation coefficient (ICC) values<sup>44</sup>, which were used to assess the degree of similarity between repeated features over the course of the study for each patient. In this work, the  $ICC(3, k)$  was calculated<sup>45</sup>—which considers the two-way random average measures with  $k$  repeated measurements—for the 14-day session across subjects, where the raters  $k$  are the study days. Reliability was categorised as either poor ( $ICC < 0.5$ ), moderate ( $ICC = 0.5–0.75$ ), good ( $ICC = 0.75–0.9$ ), or excellent ( $ICC > 0.9$ )<sup>46</sup>.

**Correcting for multiple hypothesis testing.** Multiple hypothesis testing was performed due to the large volume of features by





**Fig. 7 Self-supervised learning pipeline.** Continuous (passive) actigraphy was recorded from patients' Apple smartwatch over the study duration. Deep convolutional neural networks (DCNN) were pre-trained on 700,000 person days in the publicly available UK Biobank using self-supervised learning—and fine-tuned with the Capture-24 dataset—to estimate participant's daily activity patterns in the wearABLE-PRO study. Physical activity (PA) metrics of daily-life, for example, the time spent walking, the frequency of exercise, or the length and quality of sleep were investigated as markers to characterise symptoms of disease in people with RA compared to HC.

controlling the false discovery rate (FDR) at level  $\alpha$  using the linear step-up procedure introduced by Benjamini and Hochberg (BH)<sup>47,48</sup>.

### Machine-learning estimation of RA status and severity

This work explored how state-of-the-art machine learning (ML) models characterise the impact of RA during the daily life of participants in the 14-day wearABLE-PRO study. Multivariate modelling aimed to explore the ability of active, passive, and PRO measures to (1) distinguish RA participants from healthy controls (HC), and (2) to estimate RA disease severity: between RA participants with moderate symptoms (RA mod) and severe symptoms (RA sev) as binary classification tasks. Expansions of this analysis subsequently investigated how the in-clinic RAPID-3 assessment, a continuous measure of RA severity, could be estimated from the combination of PRO and sensor-based outcomes.

**Overview of models.** This analysis compared both linear and non-linear ML models to transform PRO and sensor-based outcomes to capture RA status and severity. Regularised linear regression (LR) models, with combinations of  $\ell_1$  and  $\ell_2$  priors, such as LR-lasso ( $\ell_1$ ), LR-ridge ( $\ell_2$ ), and LR-elastic-net ( $\ell_1 + \ell_2$ ) were compared to yield predictive, yet sparse model solutions<sup>49</sup>. Further regularisation extensions were also investigated using the sparse-group lasso (SG-lasso)—an extension of the lasso that promotes both group sparsity and within group parameter-wise ( $\ell_2$ ) sparsity, through a group lasso penalty and the lasso penalty—which aims to yield a sparse set of groups and also a sparse set of covariates in each selected group<sup>50,51</sup>.

Linear regression regularised models were also compared to decision tree (DT) based non-linear models, for instance the off-the-shelf Random Forest (RF)<sup>52</sup> and Extreme Gradient Boosted Trees (XGB)<sup>53</sup>. Both LR- and DT-based models can intrinsically perform regression or classification depending on the task required. In the LR case, classification is denoted as logistic regression (though a logit-link function). NOTE: in this analysis LR can refer to both linear regression for continuous outputs or logistic regression for classification outputs. In the DT case, the mean prediction of the individual trees creates a continuous output for regression. For further details on the models employed in this study, we refer the reader to the Supplementary Methods.

**Model evaluation.** To determine the generalisability of our models, a stratified subject-wise k-fold cross-validation (CV) was employed. This consisted of randomly partitioning the dataset into  $k=5$  folds, which was stratified with equal class proportions where possible. Participant data remained independent between training, validation, and testing splits. One set was denoted the training set (in-sample), and the remaining 20% of the dataset was then denoted testing set (out-of-sample) on which predictions were made.

**Feature-wise and prediction-wise aggregation.** In this work, we experimented with feature-wise and prediction-wise aggregation. In feature-wise aggregation, features were computed either as: daily feature values over the 14-day study period; the average daily feature value over a 7-day period (weekly); the average daily feature value over a 14-day period (fortnightly). Predictions could then be evaluated for each day (denoted *observation-wise*) or aggregated over all days through majority voting each individual prediction per subject (denoted *subject-wise*). For example, daily and weekly averaged features result in daily, or weekly predictions (i.e., *observation-wise*), which were summarised into *subject-wise* outcomes by majority voting over the repeated predictions.

**Evaluation metrics.** Multi-class classification metrics were reported as the *observation-wise* median and interquartile (IQR) range over one CV, as well as the *subject-wise* outcome for that CV,

using: auROC, area under the receiver operating characteristic curve;  $k$ , Cohen's kappa statistic<sup>54,55</sup>;  $F_1$ , F1-score. The coefficient of determination,  $r^2$ , the mean absolute error (MAE), and root-mean squared error (RMSE) were used to evaluate modelling the (continuous) in-clinic RAPID-3 scores<sup>56</sup>.

### DATA AVAILABILITY

Anonymised individual participant data that support the findings of this study are available from the corresponding author, upon reasonable request and subject to GSK's approval.

### CODE AVAILABILITY

Apple Watch sensor processing was performed using a bespoke version of the biobankAccelerometerAnalysis toolkit, found at: <https://github.com/OxWearables/biobankAccelerometerAnalysis>. Deep networks were built using Python v3.7 through a PyTorch v1.7 framework. Our self-supervised learning activity prediction code and trained models are publicly available at: <https://github.com/OxWearables/ssl-wearables>, including pre-trained models on 100K participants in the UK Biobank. Some guided test exercises and health metrics calculated are proprietary to Apple ResearchKit (<http://researchkit.org/>) and Apple HealthKit (<https://developer.apple.com/documentation/healthkit>) which we refer the reader for more details. Statistical and machine learning analysis was developed using scikit-learn v1.1.1. Further analysis code can be made available from the corresponding author upon reasonable request.

Received: 16 November 2022; Accepted: 18 January 2024;  
Published online: 12 February 2024

### REFERENCES

- Grassi, W., De Angelis, R., Lamanna, G. & Cervini, C. The clinical features of rheumatoid arthritis. *Eur. J. Radiol.* **27**, S18–S24 (1998).
- Banderas, B., Skup, M., Shields, A. L., Mazar, I. & Ganguli, A. Development of the rheumatoid arthritis symptom questionnaire (rasq): a patient reported outcome scale for measuring symptoms of rheumatoid arthritis. *Curr. Med. Res. Opin.* **33**, 1643–1651 (2017).
- Lubeck, D. P. Patient-reported outcomes and their role in the assessment of rheumatoid arthritis. *Pharmacoeconomics* **22**, 27–38 (2004).
- Campbell, R., Ju, A., King, M. T. & Rutherford, C. Perceived benefits and limitations of using patient-reported outcome measures in clinical practice with individual patients: a systematic review of qualitative studies. *Quality Life Res.* 1–24 (2021).
- Gossec, L., Dougados, M. & Dixon, W. Patient-reported outcomes as end points in clinical trials in rheumatoid arthritis. *RMD Open* **1**, e000019 (2015).
- Flurey, C. A., Morris, M., Richards, P., Hughes, R. & Hewlett, S. It's like a juggling act: rheumatoid arthritis patient perspectives on daily life and flare while on current treatment regimes. *Rheumatology* **53**, 696–703 (2014).
- Piga, M., Cangemi, I., Mathieu, A. & Cauli, A. Telemedicine for patients with rheumatic diseases: systematic review and proposal for research agenda. *In Seminars in Arthritis and Rheumatism*, Vol. 47, 121–128 (Elsevier, 2017).
- Taylor, K. I., Staunton, H., Lipsmeier, F., Nobbs, D. & Lindemann, M. Outcome measures based on digital health technology sensor data: data-and patient-centric approaches. *NPJ Digital Med.* **3**, 1–8 (2020).
- Yun, H. et al. Assessing rheumatoid arthritis disease activity with patient-reported outcomes measurement information system measures using digital technology. *Arthritis Care Res.* **72**, 553–560 (2020).
- Munos, B. et al. Mobile health: the power of wearables, sensors, and apps to transform clinical trials. *Ann. New York Acad. Sci.* **1375**, 3–18 (2016).
- Crouthamel, M. et al. Using a researchkit smartphone app to collect rheumatoid arthritis symptoms from real-world participants: feasibility study. *JMIR mHealth uHealth* **6**, e9656 (2018).
- Hamy, V. et al. Developing smartphone-based objective assessments of physical function in rheumatoid arthritis patients: the PARADE study. *Digital Biomarkers* **4**, 26–44 (2020).
- Prioreschi, A., Hodkinson, B., Avidon, I., Tikly, M. & McVeigh, J. A. The clinical utility of accelerometry in patients with rheumatoid arthritis. *Rheumatology* **52**, 1721–1727 (2013).
- Gossec, L. et al. Detection of flares by decrease in physical activity, collected using wearable activity trackers in rheumatoid arthritis or axial spondyloarthritis: an application of machine learning analyses in rheumatology. *Arthritis Care Res.* **71**, 1336–1343 (2019).

15. Pourahmadi, M. R. 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. Anatom.* **230**, 484–495 (2017).
16. Pratap, A. et al. Evaluating the utility of smartphone-based sensor assessments in persons with multiple sclerosis in the real-world using an app (elevateMS): observational, prospective pilot digital health study. *JMIR mHealth uHealth* **8**, e22108 (2020).
17. Webster, D. E. et al. Clinical validation of digital biomarkers and machine learning models for remote measurement of psoriasis and psoriatic arthritis. *medRxiv* (2022).
18. Omberg, L. et al. Remote smartphone monitoring of Parkinson's disease and individual response to therapy. *Nat. Biotechnol.* **40**, 480–487 (2022).
19. Creagh, A. P. et al. Smartphone- and smartwatch-based remote characterisation of ambulation in multiple sclerosis during the two-minute walk test. *IEEE J. Biomed. Health Inf.* **25**, 838–849 (2021).
20. Creagh, A. et al. Smartphone-based remote assessment of upper extremity function for multiple sclerosis using the draw a shape test. *Physiol. Measur.* **41**, 054002 (2020).
21. Lipsmeier, F. et al. Reliability and validity of the Roche PD mobile application for remote monitoring of early parkinson's disease. *Sci. Rep.* **12**, 1–15 (2022).
22. Lipsmeier, F. et al. A remote digital monitoring platform to assess cognitive and motor symptoms in huntington disease: cross-sectional validation study. *J. Med. Internet Res.* **24**, e32997 (2022).
23. El Miedany, Y. et al. Toward electronic health recording: evaluation of electronic patient-reported outcome measures system for remote monitoring of early rheumatoid arthritis. *J. Rheumatol.* **43**, 2106–2112 (2016).
24. Coravos, A., Khozin, S. & Mandl, K. D. Developing and adopting safe and effective digital biomarkers to improve patient outcomes. *NPJ Digital Med.* **2**, 1–5 (2019).
25. Pincus, T., Yazici, Y. & Bergman, M. J. Rapid3, an index to assess and monitor patients with rheumatoid arthritis, without formal joint counts: similar results to das28 and cdaï in clinical trials and clinical care. *Rheum. Dis. Clin.* **35**, 773–778 (2009).
26. Hamy, V. et al. Patient-centric assessment of rheumatoid arthritis using a smart-watch and bespoke mobile app in a clinical setting. *Sci. Rep.* **13**, 18311 (2023).
27. Yuan, H. et al. Self-supervised learning for human activity recognition using 700,000 person-days of wearable data. *arXiv preprint arXiv:2206.02909* (2022).
28. Willetts, M., Hollowell, S., Aslett, L., Holmes, C. & Doherty, A. Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96,220 uk biobank participants. *Scientific reports* **8**, 1–10 (2018).
29. Ainsworth, B. E. et al. 2011 compendium of physical activities: a second update of codes and met values. *Med. Sci. Sports Exerc.* **43**, 1575–1581 (2011).
30. Walsmsley, R. et al. Reallocating time from device-measured sleep, sedentary behaviour or light physical activity to moderate-to-vigorous physical activity is associated with lower cardiovascular disease risk. *MedRxiv* (2020).
31. Andreu-Perez, J. et al. Developing fine-grained actigraphies for rheumatoid arthritis patients from a single accelerometer using machine learning. *Sensors* **17**, 2113 (2017).
32. Sokka, T. et al. Physical inactivity in patients with rheumatoid arthritis: data from twenty-one countries in a cross-sectional, international study. *Arthritis Care & Research: Official Journal of the American College of Rheumatology* **59**, 42–50 (2008).
33. Keogh, A. et al. A thorough examination of morning activity patterns in adults with arthritis and healthy controls using actigraphy data. *Digital Biomarkers* **4**, 78–88 (2020).
34. Mathiowetz, V., Weber, K., Kashman, N. & Volland, G. Adult norms for the nine hole peg test of finger dexterity. *The Occupational Therapy Journal of Research* **5**, 24–38 (1985).
35. Bohannon, R. W. Sit-to-stand test for measuring performance of lower extremity muscles. *Perceptual and motor skills* **80**, 163–166 (1995).
36. of Health, U. D. et al. Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance. *Health and Quality of Life Outcomes* **4**, 79 (2006).
37. Mercieca-Bebber, R., King, M. T., Calvert, M. J., Stockler, M. R. & Friedlander, M. The importance of patient-reported outcomes in clinical trials and strategies for future optimization. *Patient Related Outcome Measures* **9**, 353 (2018).
38. Mann, H. B. & Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics* 50–60 (1947).
39. Hollander, M., Wolfe, D. A. & Chicken, E. *Nonparametric statistical methods*, Vol. 751 (John Wiley & Sons, 2013).
40. Gibbons, J. D. & Chakraborti, S. *Nonparametric Statistical Inference: Revised and Expanded* (CRC press, 2014).
41. Kruskal, W. H. & Wallis, W. A. Use of ranks in one-criterion variance analysis. *J. Am. Stat. Assoc.* **47**, 583–621 (1952).
42. Brown, M. B. & Forsythe, A. B. Robust tests for the equality of variances. *J. Am. Stat. Assoc.* **69**, 364–367 (1974).
43. Portney, L. G. & Watkins, M. P. *Foundations of clinical research: applications to practice*, vol. 892 (Pearson/Prentice Hall Upper Saddle River, NJ, 2009).
44. Weir, J. P. Quantifying test-retest reliability using the intraclass correlation coefficient and the sem. *J. Strength Condit. Res.* **19**, 231–240 (2005).
45. Shrout, P. E. & Fleiss, J. L. Intraclass correlations: uses in assessing rater reliability. *Psychol. Bull.* **86**, 420 (1979).
46. Koo, T. K. & Li, M. Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* **15**, 155–163 (2016).
47. Shaffer, J. P. Multiple hypothesis testing. *Ann. Rev. Psychol.* **46**, 561–584 (1995).
48. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.: Ser. B (Methodological)* **57**, 289–300 (1995).
49. Hastie, T., Tibshirani, R. & Friedman, J. *The elements of statistical learning: data mining, inference, and prediction* (Springer Science & Business Media, 2009).
50. Friedman, J., Hastie, T. & Tibshirani, R. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736* (2010).
51. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. A sparse-group lasso. *J. Comput. Graph. Stat.* **22**, 231–245 (2013).
52. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
53. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794 (2016).
54. He, H. & Garcia, E. A. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **21**, 1263–1284 (2009).
55. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **20**, 37–46 (1960).
56. Rao, C. R. *Linear statistical inference and its applications*, vol. 2 (Wiley New York, 1973).

## ACKNOWLEDGEMENTS

We are grateful to all the study participants and their families for their time and dedication to this study. The authors would also like to thank Priyanka Bobbili PhD, Julien Bendelac BSc, Jessica Landry MSc, Maral DerSarkissian PhD, Mihran Yenikomshian MBA, and Med Kouaici (MEng) from Analysis Group (MA, USA) for their support in app. design & development and data collection, and to Elinor Mody from Reliant Medical Group (MA, USA) for patient recruitment. The wearAble-PRO study was funded and sponsored by GSK Plc. The research described in this paper was funded by GSK Plc. This research also acknowledges support from the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC). Aiden Doherty is supported by the Wellcome Trust [223100/Z/21/Z].

## AUTHOR CONTRIBUTIONS

A.P.C. conceptualised the data analysis, designed methodology, software, and interpretation. V.H., H.Y., and G.M. contributed software applications for analysis. V.H., R.T., W.-H.C., R.W., L.G.-G. contributed to the design of the study and towards the data analysis and interpretation. C.L., C.Y., and M.S.D. were involved in the design of the study, data collection, and software for data acquisition. A.D., L.G.-G., and D.A.C. jointly supervised. A.P.C. wrote the manuscript; all other authors: review & editing.

## COMPETING INTERESTS

A.P.C., H.Y., G.M., A.D., D.A.C. are employees of the University of Oxford. A.P.C. is a GSK postdoctoral fellow and acknowledges the support of GSK. D.A.C. received research funding from GSK to conduct this work. In addition, A.D., H.Y., and G.M. acknowledge the support of Novo Nordisk plc. A.D. AD is supported by the Wellcome Trust [223100/Z/21/Z]. V.H., W.-H.C., R.T., R.W. and L.G.-G. are employees of GSK and own stock and or shares. C.L., C.Y., M.S.D. are employees of Analysis Group, which received research funding from GSK to conduct the study.

## ADDITIONAL INFORMATION

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41746-024-01013-y>.

**Correspondence** and requests for materials should be addressed to Andrew P. Creagh.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024