# Development of machine learning: derived mhealth composite biomarkers for trial@home clinical trials

Zhuparris, A.

# PART III

## ESTIMATION OF SYMPTOM SEVERITY

CHAPTER 4

# Smartphone and wearable sensors for the estimation of facioscapulohumeral muscular dystrophy disease severity: cross-sectional study

Ahnjili Zhuparris,[1] MSc; Ghobad Maleki,[1] BSc, MD; Ingrid Koopmans,[1] MSc; Robert J Doll,[1] PhD; Nicoline Voet,[2] PhD; Wessel Kraaij,[3] PhD, Prof Dr; Adam Cohen,[1] MD, PhD, Prof Dr; Emilie van Brummelen,[1] PhD; Joris H De Maeyer,[4] PhD; Geert Jan Groeneveld,[1] MD, PhD, Prof Dr

*JMIR Form Res. 2023;7:e41178. doi:10.2196/41178*

1 Centre for Human Drug Research (CHDR), Leiden, NL

2 Department of Rehabilitation, Rehabilitation Center Klimmendaal, Nijmegen, NL

3 Leiden Institute of Advanced Computer Science, Leiden University, Leiden, NL

4 Facio Therapies, Leiden, NL

## Abstract

**Background:** Facioscapulohumeral muscular dystrophy (FSHD) is a progressive neuromuscular disease. Its slow and variable progression makes the development of new treatments highly dependent on validated biomarkers that can quantify disease progression and response to drug interventions. **Objective:** We aimed to build a tool that estimates FSHD clinical severity based on behavioral features captured using smartphone and remote sensor data. The adoption of remote monitoring tools, such as smartphones and wearables, would provide a novel opportunity for continuous, passive, and objective monitoring of FSHD symptom severity outside the clinic. **Methods:** In total, 38 genetically confirmed patients with FSHD were enrolled. The FSHD Clinical Score and the Timed Up and Go (TUG) test were used to assess FSHD symptom severity at days 0 and 42. Remote sensor data were collected using an Android smartphone, Withings Steel HR+, Body+, and BPM Connect+ for 6 continuous weeks. We created 2 single-task regression models that estimated the FSHD Clinical Score and TUG separately. Further, we built 1 multitask regression model that estimated the 2 clinical assessments simultaneously. Further, we assessed how an increasingly incremental time window affected the model performance. To do so, we trained the models on an incrementally increasing time window (from day 1 until day 14) and evaluated the predictions of the clinical severity on the remaining 4 weeks of data. **Results:** The single-task regression models achieved an R2 of 0.57 and 0.59 and a root-mean-square error (RMSE) of 2.09 and 1.66 when estimating FSHD Clinical Score and TUG, respectively. Time spent at a health-related location (such as a gym or hospital) and call duration were features that were predictive of both clinical assessments. The multitask model achieved an R2 of 0.66 and 0.81 and an RMSE of 1.97 and 1.61 for the FSHD Clinical Score and TUG, respectively, and therefore outperformed the single-task models in estimating clinical severity. The 3 most important features selected by the multitask model were light sleep duration, total steps per day, and mean steps per minute. Using an increasing time window (starting from day 1 to day 14) for the FSHD Clinical Score, TUG, and multitask estimation yielded an average R2 of 0.65, 0.79, and 0.76 and an average RMSE of 3.37, 2.05, and 4.37, respectively. **Conclusions:** We demonstrated that smartphone and remote sensor data could be used to estimate FSHD clinical severity and therefore complement the assessment of FSHD outside the clinic. In addition, our results illustrated that training the models on the first week of data allows for consistent and stable prediction of FSHD symptom severity. Longitudinal follow-up studies should be conducted to further validate the reliability and validity of the multitask model as a tool to monitor disease progression over a longer period.

## Introduction

Facioscapulohumeral muscular dystrophy (FSHD) is a progressive neuromuscular disease characterized by the wasting of muscles in the face, upper body, and legs.[1] The onset and progression vary greatly between individuals.[2] Early symptoms include difficulties in smiling, whistling, and shutting of the eyelids during sleep. These symptoms are followed by impaired upper-arm movements and walking. A total of 20% of individuals with FSHD eventually become wheelchair bound.[2] Less visible FSHD symptoms include fatigue and chronic pain.[3] In addition to the physical burden, individuals with FSHD also experience emotional, social, and socioeconomic burdens.[4,5] As a result, patients report increased deterioration in quality of life as the disease progresses.[6]

Currently, there are no therapies or interventions that prevent the wasting of muscles in patients with FSHD.[7] Muscle-strengthening drugs have been shown to have limited effect on the disease progression.[8] As a result, patients with FSHD largely rely on symptomatic treatments (eg, analgesics, exercise, and cognitive therapy). The development of novel treatment options to delay or halt FSHD disease progression is currently under investigation.[9,10] However, measuring the effect of such new treatments is complicated, as disease progression is slow and no objective surrogate end points, predictive for clinical benefit, have been established.

Two common clinical assessments for assessing FSHD symptom severity are the FSHD Clinical Score and Timed Up and Go (TUG) test. The FSHD Clinical Score is composed of an evaluation of the extent of the muscle weakness among 6 regions of the body.[11] The TUG is a test used to assess functional mobility.[12] The test requires a participant to rise from a chair, walk 3 m forward, turn around, and return to the chair. These clinician-rated assessments provide a snapshot of the disease status and are primarily focused on muscular strength and function that are inherently subjective. Identifying novel objective biomarkers for monitoring disease progression could additionally provide clinically relevant insights and aid drug development. Novel digital end points for neuromuscular disease drug development have already demonstrated to be sensitive to differentiating patients from healthy volunteers and are strongly correlated with clinician assessments.[13-15] The widespread adoption of smartphones and wearables could provide new opportunities for objective and continuous monitoring of FSHD disease progression outside the laboratory.

This study was designed to identify smartphone-based and remote sensor–based features that could be used to assess FSHD disease severity. These features may enable the passive remote monitoring of disease progression and might potentially facilitate early detection of treatment effects on FSHD symptoms and the patient's quality of life. We hypothesized that the behavioral features captured by these remote monitoring devices would capture the daily physical and social burden that patients with FSHD experience. Although other neuromuscular disease studies with similar protocols have used machine learning to construct their digital end points, until now, different monitoring periods were arbitrarily selected by various researchers.[16,17] Here, we investigated how different time windows affect the model's performance to estimate one's symptom severity over time.[18,19] As these features can vary considerably over time, we assessed the stability and test-retest reliability of the first week of data to estimate FSHD disease severity for the remainder of the trial. In this paper, we describe the development of a novel tool based on smartphone and remote sensor data to provide remote estimation of FSHD disease severity.

## Methods

### OVERVIEW

This study is an extension of a previous longitudinal clinical study that investigated the feasibility of monitoring and characterizing patients with FSHD and healthy controls in terms of biometric, physical, and social activities using data sourced from smartphones and other remote monitoring devices. Therefore, additional information regarding the data collection and data quality has been previously published.[15]

## PATIENTS

This was a noninterventional, cross-sectional study involving patients with FSHD. The study was performed between April and October 2019 in the Centre for Human Drug Research (CHDR) research unit in Leiden, the Netherlands. Table 1 provides an overview of the demographic distribution of the patients with FSHD enrolled in this study.

In total, 38 patients with genetically confirmed FSHD from the Netherlands and Belgium were included in the study. Eligible patients were 16 years or older, had genetically confirmed FSHD, and had an FSHD Clinical Score greater than zero. Patients had to be Android smartphone owners and willing to use either their own smartphone or an Android smartphone provided by CHDR for the duration of the study period. Patients with internal medical devices such as a pacemaker or deep brain stimulator were excluded from the study, as these could interfere with the Withings scale measurements.[20] Participants could not be pregnant or have a severe coexisting illness.

## ETHICS APPROVAL

This study was approved by the Ethics Committee of BEBO, Assen, the Netherlands (NL69288.056.19) and was registered on ClinicalTrials.gov (NCT04999735). Before any study-related activities, written informed consent was obtained from the patients. Participants received monetary compensation for their time and effort during the trial.

To preserve the privacy of the patients, we deidentified the data and limited the amount of personally identified information collected from the smartphone and the connected devices. The location coordinates of the GPS or the cellular networks were collected as relative coordinates (GPS coordinates with respect to another predetermined location). For the calls and SMS text messaging, only metadata are stored (ie, no actual phone calls or text is being processed and stored). The call and SMS text messaging logs only store a partial phone number, making it impossible to identify the original phone numbers. As for the Withings devices, we created a unique email address (containing patient identifiers) for each patient to couple the Withings device with CHDR MORE, thus eliminating the need for using the patients' personal email.

## INVESTIGATIONAL TECHNOLOGIES

Smartphone and remote sensor data were collected on the CHDR MORE platform. This customizable platform enables the collection, ingestion, and management of data sourced from monitoring devices. The CHDR MORE app was installed on the smartphone of each participant and allows for the unobtrusive collection of smartphone sensor data (sourced from the smartphone's accelerometer, gyroscope, magnetometer, GPS, light sensor, and microphone) as well as phone usage logs (eg, app usage, battery level, calls, and SMS text messages).

The smartphone sensor data provide insights into a participant's environment, such as location type and travel patterns (GPS), if human voices are present in the environment (microphone), and their physical activity (accelerometer and gyroscope). The phone usage logs give an indication of social activity (through social media and communication apps, calls, and SMS text messages) and smartphone usage (app usage). The app also collected Withings health data.

In this study, 3 Withings devices were used: Withings Steel HR smartwatch (monitors heart rate, sleep states, and a number of steps), Withings Body+ scale (monitors weight and body composition) and Withings BPM Connect (monitors heart rate, systolic blood pressure, and diastolic blood pressure). Together the Withings features reflect the daily physical activities of each of the participants.

This is the first study that aimed to monitor and estimate FSHD symptom severity using smartphone and wearable data. As this was an exploratory longitudinal study, specifically aimed to identify smartphone- and wearable-based features that were predictive of FSHD symptom severity, we did not identify any literature with a similar protocol. To identify these novel features, we decided to collect data from all available sensors and features from the CHDR MORE platform. As the symptoms of FSHD can

affect a patient's travel abilities,[21] physical activity, sleep,[11,22] and social lives,[23] we deemed these features relevant for estimating FSHD symptom severity.

## DATA COLLECTION

Participants were monitored for 6 continuous weeks. On days 1 and 42, the clinical evaluations (FSHD Clinical Score and TUG) were performed. On day 1, the CHDR MORE and Withings Health Mate apps were installed on their smartphones. Participants were asked to use their smartphones as normal. Participants were asked to continuously wear their Withings Steel HR smartwatch and weigh themselves and take their blood pressure weekly.

## DATA PREPROCESSING

Before modeling of the data, all sensor data were preprocessed and converted into features using Python (version 3.6.0) and the PySpark (version 3.0.1) library. The raw data were checked for missing values and outliers. Missing values were defined as the absence of data for a specific feature for each day, except for 2 types of measurements: the weekly measurements (eg, weight and blood pressure) and the data related to aperiodic activities (eg, phone calls or SMS text messages). Missing data were not imputed. Outliers were detected by manual visual inspection rather than automated statistical techniques, as our objective was to identify potential outliers that were a result of potential measurement errors rather than participants' behaviors. Measurement errors were deemed not relevant to our analysis, whereas outliers in behavior could still provide insights into a participant's symptom severity; therefore, sensitivity analysis was not conducted. Outliers would be subsequently excluded at the discretion of the authors (eg, removing overlapping sleep stages).

## FEATURE EXTRACTION

All raw data were collected from the smartphone and Withings devices. The features were then aggregated per day, as the symptom severity exhibited on a given day is the focus of FSHD clinical evaluation. As there are no FSHD assessments that assess FSHD symptoms over a longer period, we did not explore other aggregation methods. Discrete features (eg, step count) were summed per day per participant. Continuous features (eg, heart rate) were averaged per day per participant. Table 2 provides an overview of how the features were aggregated based on the data type. Table 3 summarizes which features were extracted from the smartphone and Withings sensors. In addition, Table 3 shows the features that were provided from the MORE platform but were not included for the analysis either due to outliers, missing data, or because they were not of clinical interest.

## FEATURE SELECTION

Before modeling, both expert-based manual and automated feature selections were performed. First, features were visually inspected by all authors. Excluded features were based on the number of available data points (eg, 9 participants did not have body composition data) and clinical relevance (eg, time spent on parenting apps was deemed clinically irrelevant). Next, two automated feature selection strategies were compared: (1) stepwise regression and (2) variance inflation factor (VIF). The stepwise regression strategy was an iterative process to select predictive variables that met a significance criterion ($P<.05$). Both forward and backward stepwise regression strategies were used. The VIF was calculated for all pairwise combinations of features to identify collinear features. Pairs of features having a VIF value greater than 10 were identified, and one of the features was subsequently removed for each of the pairs.[24] For comparison, we also fitted the model without any automated feature selection strategies. For each regression model, we applied each of the feature selection strategies.

## STATISTICAL ANALYSIS

Python (version 3.6.0) was used for the data analysis and modeling in conjunction with the Pandas,[25] NumPy,[26] Matplotlib,[27] and Sklearn

packages.[28] Three regression models were created: 2 single-task regression models, 1 for each clinical assessment and 1 for each multitask regression model, simultaneously estimating both clinical assessments. For the multitask regression model, a dummy variable was included to denote either the FSHD Clinical Score or TUG.

For all models, linear regression, random forest regressor, and gradient boost regressor were used. A grid search was performed to optimize the hyperparameters for each model. For the Elastic Net linear regression model, we optimized the hyperparameters for the α (range 0-200) and L1 ratio (range 0.0-1.0). For the random forest and gradient boost regressors, we optimized the hyperparameters for the number of estimators (range 0-200), maximum depth (range 1-20), maximum features (range: auto, square root, log2), and maximum leaf nodes (range 2-20). In addition, we optimized the learning rate (range 0.0-1.0) for the gradient boost regressor.

Each model was validated using a group 5 outer-fold and 5 inner-fold nested cross-validation. By using group cross-validation, for each fold, we ensure that the participants in the validation are not also present in the training fold. While the data for all participants were used for the modeling, the cross-validation procedure was used for out-of-sample testing; hence, for each fold of the cross-validation procedure, only a subsample of participants' data were used. Further, the random forest and gradient boost regressor models only consider a subsample of participants and features per decision tree node. The elastic-net linear regression penalization would also reduce the potential features considered in the model. The cross-validation and models together would improve the generalizability and robustness of the models and therefore reduce the probability of spurious correlations.

We applied each of the feature selection strategies to each of the regression models and compared the results of each model. The model that provided the highest R2 (variance explained) and the lowest root-mean-square error (RMSE) was selected as the best-performing model. The R2 and the RMSE explain the variance and the error between the true clinical scores and the predicted scores of the regression models, respectively.

To assess how varying time window affects the model's estimation of symptom severity, we used an incrementally increasing time window to train the regression models, starting with day 1 and adding the following days until the first 2 weeks of data were included in the training set. To train, optimize, and assess each model's generalizability, we applied a 5-fold nested cross-validation model. To validate the performance of these models, we used the remaining 4 weeks of data as an external validation data set. To assess the stability of the trained models to yield consistent estimations of symptom severity, we trained the FSHD Clinical Score, TUG, and multitask models on the first week of data. We estimated the symptom severity for the subsequent weeks. We selected the first week, as each patient would have each day of the week represented in their data set.

In sum, we investigated 3 final models, 2 single-task models, and 1 multitask model. For each model, we considered 3 types of regression models (the linear regression, the random forest regressor, and the gradient boost regressor). For each model, we considered 3 feature selection strategies (no automated feature selection, stepwise regression, and VIF); hence, in total, we compared 27 models. Given that we are mainly interested in the comparison of the predictions of single-task and multitask models and the influence of the time windows on the predictions, we reported only the results of these models.

## Results

No patients dropped out of the study. One patient was wheelchair-bound and therefore unable to perform the TUG. The FSHD Clinical Scores ranged between 1 and 13, with a median score of 5. The TUG times ranged between 5.5 seconds and 15.8 seconds, with a median time of 7.7 seconds. Before modeling, several features were manually excluded. Nine patients had

no body composition (eg, fat and muscle mass) data. As a result, the Withings body composition data (except weight) were excluded from the final analysis. We excluded SMS text message–related features as not all the patients used SMS text messaging (less than 30% of patients), and the SMS text message features were not deemed clinically relevant. Further, we excluded smartphone apps from the analysis that were used by less than 5% of the patients. We did not exclude any outliers as none of the data points were viewed as potential measurement errors. In a previous publication, we provided an overview of the proportion of observations that were missing per feature.[15]

The FSHD Clinical Score for 24 participants did not change over the 6 weeks. The scores of the remaining 14 participants changed by +1 or −1 point. The average difference between the day 1 and day 42 TUG scores was 0.38 seconds (95% CI 0.12-0.63). After reviewing the stability of the TUG and FSHD scores, we decided to use the averaged clinical assessment scores as the outcomes for all models. Subsequently, each feature was also averaged over the 6 weeks. These averaged features were used as inputs for the regression models.

Using all 6 weeks of data, we built a single-task model that used the CHDR MORE features to estimate the FSHD Clinical Score for each participant. Comparing the estimated scores and the true FSHD Clinical Score yielded an R2 of 0.57 and an RMSE of 2.09. This was achieved using VIF-selected features and Elastic Net–penalized linear regression. A total of 11 features were predictive of the FSHD Clinical Score, as seen in Figure 1. The features were related to app usage, blood pressure, location visits, and calling behaviors. Figure 2 (top) shows the estimated FSHD Clinical Score in relation to the actual FSHD Clinical Score.

Similarly, the comparison of the TUG single-task model estimated TUG and the actual TUG yielded an R2 of 0.59 and an RMSE of 1.66 (seconds) for each participant. This was achieved with forwarding selection stepwise regression and Elastic Net–penalized linear regression. In total, 13 feature were predictive of the TUG score (Figure 1). The feature categories related to age, app usage, calling behaviors, sleep, physical activity, and location visits were predictive of TUG. Figure 2 (bottom) illustrates the relationship between the predicted and actual TUG times.

The multitask model achieved an R2 of 0.74 and an RMSE of 1.89 for the FSHD Clinical Score and TUG prediction together. The same model achieved an R2 of 0.66 and an RMSE of 1.97 for the FSHD Clinical Score and an R2 of 0.81 and an RMSE of 1.61 for the TUG separately. The gradient boost regressor selected 50 predictive features. The relative feature importance is presented in Figure 3. The 5 most important features were light sleep duration, total steps per day, mean steps per minute, the number of times the social and communication apps were opened, and the number of incoming calls. Figure 4 illustrates the relationship between the predicted clinical scores and the actual clinical scores.

For each clinical score, we evaluated the effect of different monitoring periods on the estimation of symptom severity. The best performing FSHD Clinical Score single-task model, TUG single-task model, and multitask model yielded the highest R2 on day 3 (0.70), week 2 (0.86), and day 1 (0.86), and the lowest RMSE on day 3 (2.8), week 2 (1.9), and day 6 (3.4), respectively. As seen in Figure 5, although our analysis has identified windows that yielded the highest R2 and RMSE, we found that the mean (SD) of the R2 and RMSE for the FSHD Clinical Score single-task model, TUG single-task model, and multitask model was 0.65 (0.03) and 3.37 (0.19), 0.79 (0.05) and 2.05 (0.09), and 0.76 (0.08) and 4.37 (0.20), respectively. When evaluating the stability, the models trained on a week's worth of data were used to estimate the symptom severity for subsequent days. We found that the FSHD Clinical Score, TUG, and multitask models achieved median R2 (median RMSE) of 0.51 (3.66), 0.42 (2.44), and 0.72 (2.61), respectively (as seen in Figure 6).

## Discussion

### PRINCIPAL FINDINGS

We developed and compared 2 regression models to monitor and estimate FSHD symptom severity outside the clinic with remote sensor data

to estimate the FSHD Clinical Score and TUG for each participant. For the first type of model, both clinical assessment scores were separately estimated using 2 single-task regression models. For the second type of model, both clinical assessment scores were simultaneously estimated using a multitask regression model.

The 2 single-task models selected features that were uniquely predictive of each of the clinical scores. In addition, the models' selected features were found to be predictive for both scores (time spent at health locations and total call duration). Other studies have found that (a modified version of) the TUG significantly correlated to the FSHD Clinical Score,[12,29] indicating that these clinical scores share mutual information. Simultaneously estimating multiple tasks with shared features can improve the model performance.[30-32] This supports the notion that a multitask approach would improve the estimation of FSHD symptom severity.

Indeed, the multitask modeling of both the FSHD Clinical Score and the TUG outperformed the single-task models. Additionally, the multitask model identified features not selected as important by the single-task models (eg, sleep and the resting heart rate). The clinical assessments and their respective single-task models only captured a limited range of disease symptoms, which misses the opportunity to model other aspects of the disease (eg, sleep impairments [33,34] and arrhythmic abnormalities [35]). The multitask model, however, identified features representative of a broader range of FSHD symptoms. As shown in the SHAP (SHAPley Additive exPlanations) plot (Figure 3), participants with a higher mean step per minute, light sleep duration, soft activity duration, and total steps (indicated by the red feature value) had lower SHAP values. This indicates that participants with more physical activity and better sleep quality had a lower FSHD Clinical Score and TUG. Although the multitask model outperformed the single-task models, the multitask model required approximately twice as many features as the single-task models. Using fewer features could be considered beneficial as it reduces the number of sensors needed. Additionally, it eases the interpretation of the results. Therefore, there is a tradeoff between the performance of estimation of disease severity and the complexity of the data set and model. However, given that the multitask model showed an important improvement over the single-task models, we recommend using the multitask model for future estimation of the FSHD Clinical Score and TUG.

It is critical to determine how much data are needed to obtain reliable inferences without burdening the patients and the clinicians. Insufficient data can lead to inaccurate extrapolations, whereas excessive data can lead to wasted time and resources. This study investigated how long a patient needs to be monitored to estimate symptom severity reliably. Our results demonstrated that behaviors exhibited that based on our sample, the optimal time window (based on the highest R2 and lowest RMSE) varied for each task. The multitask model yielded the overall highest R2 based on a training data set of the first day. Although we identified that 5 days of data seem sufficient for training the multitask model, a longer or shorter time window would still provide consistent estimation of the symptom severity. However, our results also demonstrate that selecting any time window between days 1 and 14 would produce relatively stable results. Our results also demonstrated that training the multitask model on the first week of data allowed for constant and reliable estimations of symptom severity for the subsequent weeks. This further supports the notion that multitask should be used to estimate the clinical scores for longitudinal studies.

The agreement between the clinical scores and the remotely monitored features did not achieve 100% adherence. This may be due to the sensors being unable to capture specific aspects of the clinical score. For example, features captured by the remote monitoring system may not provide sufficient proxies for arm, scapular, and abdominal weaknesses (which the FSHD Clinical Score specifically addresses). Adding additional sensors and features could potentially allow for more complete modeling of FSHD. For example, an additional accelerometer could try to capture arm swings [36] or detect the (limited) shoulder range of motion.[37] Another explanation for the imperfect model fit is that the clinical scores have limited accuracy in capturing disease severity. There can be variation within

a specific clinical score, as patients with the same scores may exhibit different FSHD symptoms. For example, patients with scores between 2 and 4 may have impairments related to facial muscles and upper limbs, whereas others may be unable to walk on their heels.[11]

The clinical scores provide snapshots of muscular strength and function, whereas the remote monitoring approach provides a more continuous measure of (FSHD-related) social and physical activity. Additionally, the clinical scores were assessed at the clinic, whereas the sampling of the remotely monitored features occurred at home, and in daily practice. Altogether, these 2 clinical scores may not be the optimal clinical assessment strategies for fully assessing FSHD symptom severity. These are only 2 of several FSHD-related assessments that can be used in a clinical trial. The remotely monitored features may show different correlations with other FSHD-related assessments such as the Clinical Severity Scale for FSHD [38,39] and the Pittsburgh Sleep Quality Index.[39,40] Although the remotely monitored features may not correlate strongly with the 2 clinical scores, they still provide relevant insights into FSHD-related symptoms. Our multitask model could prove to be a promising tool for monitoring the FHSD severity based on patients' everyday activities outside the clinic.

Although the models cannot replace the TUG or FSHD Clinical Scores for estimating the disease severity, these models can potentially be used as a (complimentary) tool in clinical studies. When validated in longitudinal studies, given the continuous sampling of data from multiple sensors, this FSHD tool could potentially be used to track the symptom severity for long periods of time without patients having to visit a clinic. Previous studies have demonstrated that this approach of using smartphone-based models to quantify medication responses can be advantageous.[37,38] When implemented in a clinical trial, the FSHD tool might be evaluated as a tool to monitor drug effectiveness by tracking drug-induced changes in FSHD symptom severity.[41] Additionally, it might enable the identification of improvements in specific aspects of the disease severity (e.g., muscle function or sleep quality). Therefore, remote monitoring might aid clinicians' assessments of a patient's status during a clinical trial based on the review of the patient's in-clinic assessments and out-of-clinic daily activity.

We present an FSHD tool that estimates the FSHD Clinical Score and TUG using smartphone and remote sensor data. The conclusions drawn from this study are preliminary in view of the relatively small sample size and cross-sectional study nature. Given the short observation period, we did not expect changes in the patients' FSHD scores. As a result, we could not validate the use of the model to estimate changes in the FSHD severity over time. A trial where the FSHD clinical score is expected to change could help validate the FSHD tool's capacity to detect changes in FSHD symptom severity. Additionally, the FSHD tool could be improved by including more patients with FSHD and adding other remote sensors. All in all, the remote monitoring approach presented here could be a promising tool for monitoring FSHD severity outside the clinic environment.

## Conclusions

We presented a smartphone-based and remote sensor–based FSHD tool that can estimate a patient's FSHD symptom severity. This is the first study to demonstrate how to monitor patients with FSHD remotely and subsequently model their FSHD Clinical Score and TUG simultaneously. The tool holds potential for monitoring disease progression and drug intervention effects outside the clinic, pending a longitudinal follow-up study to validate the capacity of the FSHD tool to detect changes in the disease severity score over time due to disease progression or drug intervention.

## REFERENCES


1 Statland JM, McDermott MP, Heatwole C, Martens WB, Pandya S, van der Kooi E, et al. Reevaluating measures of disease progression in facioscapulohumeral muscular dystrophy. Neuromuscul Disord 2013;23(4):306-312 [doi:10.1016/j.nmd.2013.01.008] [Medline: 23406877]

2 Tawil R, Van Der Maarel SM. Facioscapulohumeral muscular dystrophy. Muscle Nerve 2006;34(1):1-15. [doi: 10.1002/mus.20522] [Medline: 16508966]

3 Statland JM, Tawil R. Risk of functional impairment in facioscapulohumeral muscular dystrophy. Muscle Nerve 2014;49(4):520-527. [doi: 10.1002/mus.23949] [Medline: 23873337]

4 Blokhuis AM, Deenen JCW, Voermans NC, van Engelen BGM, Kievit W, Groothuis JT. The socioeconomic burden of facioscapulohumeral muscular dystrophy. J Neurol 2021;268(12):4778-4788 [doi:10.1007/s00415-021-10591-w] [Medline: 34043041]

5 Hamel J, Johnson N, Tawil R, Martens WB, Dilek N, McDermott MP, et al. Patient-reported symptoms in facioscapulohumeral muscular dystrophy(PRISM-FSHD). Neurology 2019;93(12):e1180-e1192 [doi: 10.1212/WNL.0000000000008123] [Medline: 31409737]

6 Morís G, Wood L, FernáNdez-Torrón R, González Coraspe JA, Turner C, Hilton-Jones D, et al. Chronic pain has a strong impact on quality of life in facioscapulohumeral muscular dystrophy. Muscle Nerve 2018;57(3):380-387 [doi: 10.1002/mus.25991] [Medline: 29053898]

7 Tawil R, Mah JK, Baker S, Wagner KR, Ryan MM, Sydney Workshop Participants. Clinical practice considerations in facioscapulohumeral muscular dystrophy Sydney, Australia, 21 September 2015. Neuromuscul Disord 2015;26(7):462-471. [doi: 10.1016/j.nmd.2016.03.007] [Medline: 27185458]

8 Ramos VFML, Thaisetthawatkul P. A case of fascioscapulohumeral muscular dystrophy misdiagnosed as Becker's muscular dystrophy for 20 years. Age Ageing 2012;41(2):273-274. [doi: 10.1093/ageing/afr095] [Medline: 21795275]

9 Zhou L, Parmanto B. Reaching people with disabilities in underserved areas through digital interventions: systematic review. J Med Internet Res 2019;21(10):e12981 [doi: 10.2196/12981] [Medline: 31654569]

10 Churová V, Vyškovský R, Maršálová K, Kudláček D, Schwarz D. Anomaly detection algorithm for real-world data and evidence in clinical research: implementation, evaluation, and validation study. JMIR Med Inform 2021;9(5):e27172 [doi: 10.2196/27172] [Medline: 33851576]

11 Lamperti C, Fabbri G, Vercelli L, D'Amico R, Frusciante R, Bonifazi E, et al. A standardized clinical evaluation of patients affected by facioscapulohumeral muscular dystrophy: the FSHD clinical score. Muscle Nerve 2010;42(2):213-217. [doi: 10.1002/mus.21671] [Medline: 20544930]

12 Chan V, Hatch M, Kurillo G, Han J, Cadavid D. Development of an optimized timed up and go (OTUG) for measurement of changes in mobility impairment in facioscapulohumeral muscular dystrophy (FSHD) clinical trials (2228). Neurology 2020;94(15):2228.

13 Boukhvalova AK, Fan O, Weideman AM, Harris T, Kowalczyk E, Pham L, et al. Smartphone level test measures disability in several neurological domains for patients with multiple sclerosis. Front Neurol 2019;10:358 [doi: 10.3389/fneur.2019.00358] [Medline: 31191424]

14 Servais L, Camino E, Clement A, McDonald CM, Lukawy J, Lowes LP, et al. First regulatory qualification of a novel digital endpoint in Duchenne muscular dystrophy: a multi-stakeholder perspective on the impact for patients and for drug development in neuromuscular diseases. Digit Biomark 2021;5(2):183-190 [doi: 10.1159/000517411]. Medline: 34723071]

15 Maleki G, Zhuparris A, Koopmans I, Doll RJ, Voet N, Cohen A, et al. Objective monitoring of facioscapulohumeral dystrophy during clinical trials using a smartphone app and wearables: observational study. JMIR Form Res 2022;6(9):e31775 [doi: 10.2196/31775] [Medline: 36098990]

16 Jauhiainen M, Puustinen J, Mehrang S, Ruokolainen J, Holm A, Vehkaoja A, et al. Identification of motor symptoms related to Parkinson disease using motion-tracking sensors at home (KÄVELI): protocol for an observational case-control study. JMIR Res Protoc 2019;8(3):e12808 [doi: 10.2196/12808] [Medline: 30916665]

17. Jeannet PY, Aminian K, Bloetzer C, Najafi B, Paraschiv-Ionescu A. Continuous monitoring and quantification of multiple parameters of daily physical activity in ambulatory Duchenne muscular dystrophy patients. Eur J Paediatr Neurol 2011;15(1):40-47. [doi: 10.1016/j.ejpn.2010.07.002] [Medline: 20719551]

18 Zhong T, Zhuang Z, Dong X, Wong KH, Wong WT, Wang J, et al. Predicting antituberculosis drug-induced liver injury using an interpretable machine learning method: model development and validation study. JMIR Med Inform 2021;9(7):e29226 [doi: 10.2196/29226] [Medline: 34283036]

19 Chae SH, Kim Y, Lee K, Park H. Development and clinical evaluation of a web-based upper limb home rehabilitation system using a smartwatch and machine learning model for chronic stroke survivors: prospective comparative study. JMIR Mhealth Uhealth 2020;8(7):e17216 [doi: 10.2196/17216] [Medline: 32480361] 20. Body+ - Frequently asked questions about safety. Withings. URL: https://support.withings.com/hc/en-us/articles/218438708-Body-Frequently-asked-questions-about-safety [accessed 2021-02-24]

21 Faux-Nightingale A, Kulshrestha R, Emery N, Pandyan A, Willis T, Philp F. Upper limb rehabilitation in facioscapulohumeral muscular dystrophy: a patients' perspective. Arch Rehabil Res Clin Transl 2021;3(4):100157 [doi: 10.1016/j.arrct.2021.100157] [Medline: 34977539]

22 Mul K, Lassche S, Voermans NC, Padberg GW, Horlings CG, van Engelen BG. What's in a name? The clinical features of facioscapulohumeral muscular dystrophy. Pract Neurol 2016;16(3):201-207. [doi: 10.1136/practneurol-2015-001353] [Medline: 26862222]

23. Johnson NE, Quinn C, Eastwood E, Tawil R, Heatwole CR. Patient-identified disease burden in facioscapulohumeral muscular dystrophy. Muscle Nerve 2012;46(6):951-953 [doi: 10.1002/mus.23529] [Medline: 23225386]

24 Hair JFJ, Anderson RE, Tatham RL, Black WC. F. In: Multivariate Data Analysis, 3rd Ed. New York: Macmillan; 1995.

25 McKinney W. Data structures for statistical computing in python. 2010 Presented at: Proceedings of the 9th Python in Science Conference; June 28-July 3, 2010; Austin, Texas. [doi: 10.25080/majora-92bf1922-00a]

26 Oliphant TE. A Guide to NumPy. USA: Trelgol Publishing; 2006.

27 Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng 2007;9(3):90-95. [doi: 10.1109/mcse.2007.55]

28 Buitinck L, Louppe G, Blondel M, Pedregosa F, Muller AC, Grisel O, et al. API design for machine learning software: experiences from the scikit-learn project. ArXiv. Preprint posted online September 1, 2013 2013:108-122.

29 Huisinga J, Bruetsch A, Mccalley A, Currence M, Herbelin L, Jawdat O, et al. An instrumented timed up and go in facioscapulohumeral muscular dystrophy. Muscle Nerve 2018;57(3):503-506 [doi: 10.1002/mus.25955] [Medline: 28877559]

30 Yu G, Liu Y, Shen D. Graph-guided joint prediction of class label and clinical scores for the Alzheimer's disease. Brain Struct Funct 2016;221(7):3787-3801 [doi: 10.1007/s00429-015-1132-6] [Medline: 26476928]

31. Li Y, Tian X, Liu T, Tao D. Multi-task model and feature joint learning. 2015 Presented at: Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence; 25-31 July, 2015; Buenos Aires, Argentina.

32 Zhang D, Shen D, Alzheimer's Disease Neuroimaging Initiative. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. Neuroimage 2012;59(2):895-907 [doi:10.1016/j.neuroimage.2011.09.069] [Medline: 21992749]

33 Runte M, Spiesshoefer J, Heidbreder A, Dreher M, Young P, Brix T, et al. Sleep-related breathing disorders in facioscapulohumeral dystrophy. Sleep Breath 2019;23(3):899-906. [doi: 10.1007/s11325-019-01843-1] [Medline: 31025273]

34 Leclair-Visonneau L, Magot A, Tremblay A, Bruneau X, Pereon Y. Anxiety is responsible for altered sleep quality in facio-scapulo-humeral muscular dystrophy (FSHD). Neuromuscul Disord 2013;23(9-10):823-824. [doi:10.1016/j.nmd.2013.06.642]

35 Emre Cagliyan C, Gelincik's H, Celic AI, Filiz Koc A. Impaired autonomic and repolarization abnormalities are observed in patients with facioscapulohumeral dystrophy despite normal myocardial functions. J Neurol Neurosurg 2018;05(01):127-131. [doi: 10.19104/jnn.2018.42]

36 LeMoyne R, Tomycz N, Mastroianni T, McCandless C, Cozza M, Peduto D. Implementation of a smartphone wireless accelerometer platform for establishing deep brain stimulation treatment efficacy of essential tremor with machine learning. Annu Int Conf IEEE Eng Med

Biol Soc 2015;2015:6772-6775. [doi: 10.1109/EMBC.2015.7319948] [Medline: 26737848]
37 Boissy P, Diop-Fallou S, Lebel K, Bernier M, Balg F, Tousignant-Laflamme Y. Trueness and minimal detectable change of smartphone inclinometer measurements of shoulder range of motion. Telemed J E Health 2017;23(6):503-506. [doi: 10.1089/tmj.2016.0205] [Medline: 27911652]
38 Ricci E, Galluzzi G, Deidda G, Cacurri S, Colantoni L, Merico B, et al. Progress in the molecular diagnosis of facioscapulohumeral muscular dystrophy and correlation between the number of KpnI repeats at the 4q35 locus and clinicalphenotype. Ann Neurol 1999;45(6):751-757. [doi: 10.1002/1531-8249(199906)45:6<751::aid-ana9>3.0.co;2-m] [Medline: 10360767]
39 Della Marca G, Frusciante R, Vollono C, Dittoni S, Galluzzi G, Buccarella C, et al. Sleep quality in facioscapulohumeral muscular dystrophy. J Neurol Sci 2007;263(1-2):49-53. [doi: 10.1016/j.jns.2007.05.028] [Medline: 17597162]
40 Buysse DJ, Reynolds CF, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh sleep quality index: a new instrument for psychiatric practice and research. Psychiatry Res 1989;28(2):193-213. [doi: 10.1016/0165-1781(89)90047-4] [Medline: 2748771]
41 Zhang H, Guo G, Song C, Xu C, Cheung K, Alexis J, et al. PDLens: smartphone knows drug effectiveness among Parkinson's via daily-life activity fusion. 2020 Presented at: MobiCom '20: The 26th Annual International Conference on Mobile Computing and Networking; 21-25 September, 2020; London, United Kingdom. [doi: 10.1145/3372224.3380889]

**TABLE 1** An overview of characteristics of the FSHD participants (N=38).

| Demographics | Values |
|---|---|
| GENDER, N | |
| Female | 23 |
| Male | 15 |
| RACE, N | |
| African American | - |
| Mixed | 1 |
| White | 37 |
| Age (years), mean (SD) (minimum, maximum) | 44 (14.5) (18, 64) |
| Weight (kg), median (SD) (minimum, maximum) | 79 (16) (52, 130) |
| BMI (kg/m2), median (SD) (minimum, maximum) | 25 (4) (20, 44) |
| FSHD Clinical Score, median (SD) (minimum, maximum) | 5 (3) (1, 13) |
| Timed Up and Go test (seconds), median (SD) (minimum, maximum) | 7.7 (2.4) (5.5, 15.8) |

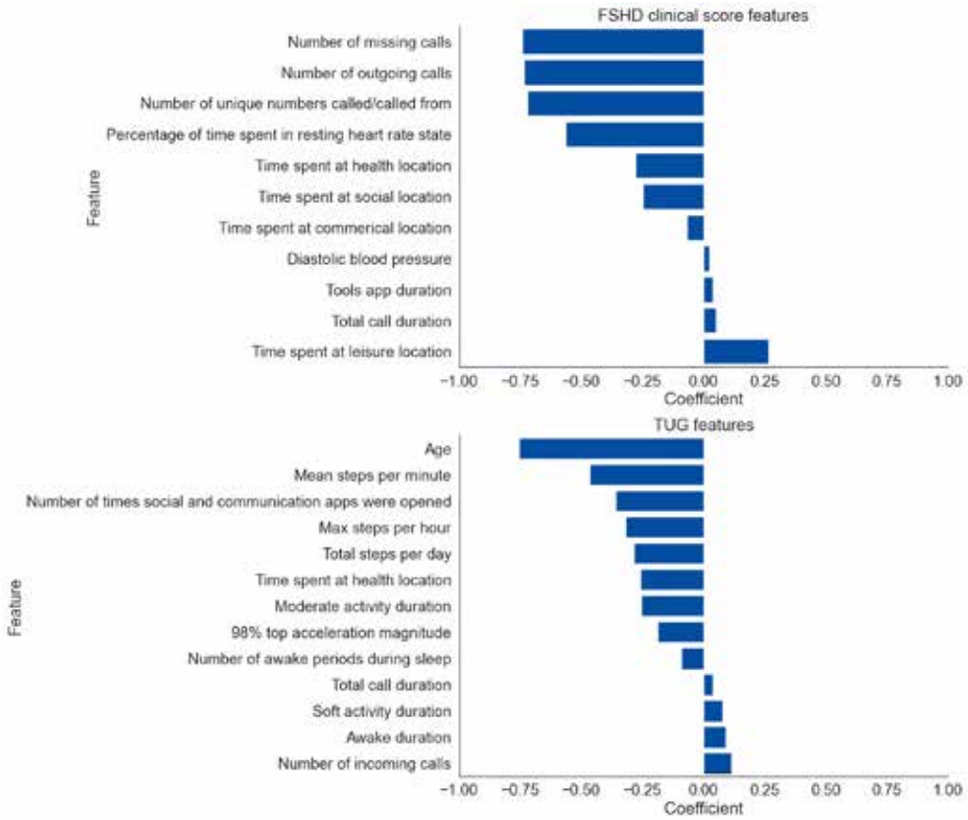
**TABLE 2** A simplified summation of how the features were aggregated based on the data type.

| Data Type | Time Unit | Example Feature | Aggregation Format | Example Aggregation |
|---|---|---|---|---|
| COUNT | Per day<br>Per hour | Steps | Sum<br>Mean<br>Max | Total Steps<br>Max Steps Per Hour<br>Mean Steps Per Hour |
| CONTINUOUS DATA WITHIN A RANGE | Per day | Heart Rate | Min (5%)<br>Median (50%)<br>Max (95%) | Lowest 5% Heart Rate<br>Median Heart Rate<br>Maximum 95% Hr |
| DURATION | Per day | App Usage | Total Duration<br>Mean Duration | Total Duration of Social Apps Opened<br>Mean Duration of Social App Use Per Interaction |
| GPS COORDINATES | Per day | Location | Sum<br>Max<br>Mean | Total Distance Travelled<br>Mean And Max Distance From Home |

**TABLE 3** An overview of the features provided from the MORE platform and the features that were subsequently aggregated per day (with the exception of the body measurements as that was measured once a week).

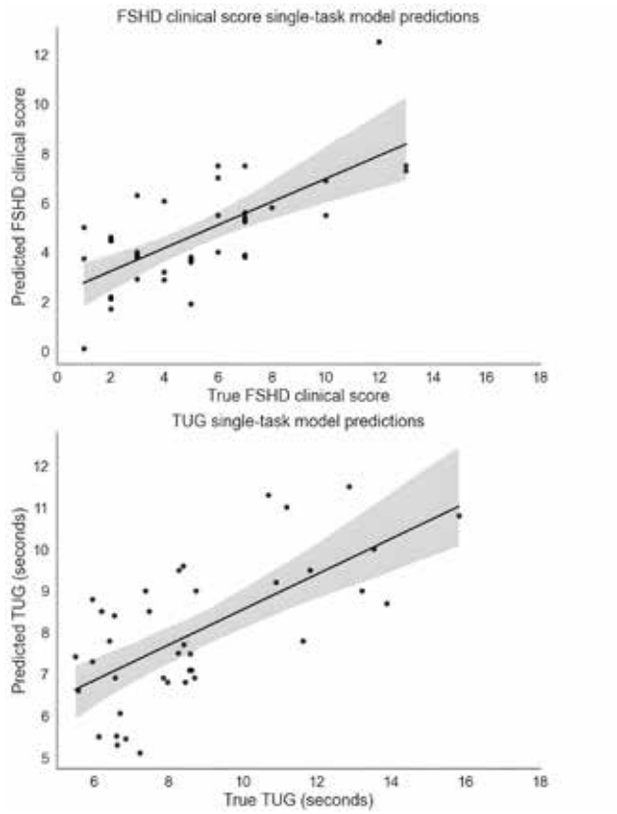
| Category | MORE Features | Derived Features (Per day) | Excluded Features |
|---|---|---|---|
| DEMOGRAPHICS | Age<br>Gender | Age<br>Gender | |
| ACCELERATION | Acceleration magnitude<br>Gyroscope<br>Magnetometer | 98% Acceleration magnitude | Mean Acceleration Magnitude |
| ACTIVITY | Steps<br>Heart Rate<br>Physical activity duration<br>Calories | **STEPS:** Total steps, max steps per hour, mean steps per hour **HEART RATE:** 5%, 50% & 95% beats per minute (BPMs), standard deviation of BPMs, % time spent in resting state **PHYSICAL ACTIVITY:** soft, moderate and intense activity duration | Calories<br>Distance Travelled<br>Distance Per Step |
| APPS | APP CATEGORIES: Health & Fitness, Recreational, Communication & Social, Tools, Shopping | Duration<br>Times Open | House & Home<br>Libraries & Demo<br>Reading<br>Travel |
| BODY | Diastolic Blood Pressure<br>Systolic Blood Pressure<br>Heart Pulse (Bpm)<br>Weight | Diastolic blood pressure<br>Systolic blood pressure<br>Heart pulse (bpm)<br>Weight | Height (M)<br>Fat mass (kg)<br>Fat ratio (%)<br>Hydration<br>Muscle Mass |
| LOCATION | LOCATION CATEGORIES: Commercial, Health, Home, Leisure, Public, Social, Travel | Total duration at place<br>Total distance travelled<br>Total no of unique places visited<br>Max distance from home<br>Time spent commuting | |
| SOCIAL | Calls<br>Voice | Number of calls<br>Number of unique numbers<br>Number of incoming, outgoing and missing calls<br>Number of calls from known & unknown numbers<br>Total duration of calls<br>Average duration of calls<br>% Time human voice is detected | Text messages (SMS) |
| SLEEP | | Number of sleep sessions<br>Total sleep duration<br>Number of sleep phases (awake, light sleep and deep sleep)<br>Duration of sleep phases (awake, light and deep sleep)<br>Time between sleep sessions<br>Time to fall asleep | |

**FIGURE 1** Linear regression coefficients for the features selected by the single-task FSHD Clinical Score and TUG models. Features with a coefficient of zero are not shown.




*FSHD: facioscapulohumeral muscular dystrophy; TUG: Timed Up and Go.*

**FIGURE 2** True FSHD Clinical Scores and TUG times against the predicted scores using the respective FSHD Clinical Score and TUG regression models. The lines represent a regression line with a 95% CI band.




*FSHD: facioscapulohumeral muscular dystrophy; TUG: Timed Up and Go.*

**FIGURE 3** SHAP (SHAPley Additive exPlanations) variable importance plot showing the feature importance of the top 20 most important features, in which the features are ranked in descending order. Each scatter point represents one prediction. The color of the scatter point reflects the value of the real data. If the actual value of the data point was high, then the color was red. If the value was low, then the color was blue. The SHAP value, as illustrated on the x-axis, shows the direction and magnitude of each feature's contribution toward predicting the facioscapulohumeral muscular dystrophy symptom severity.

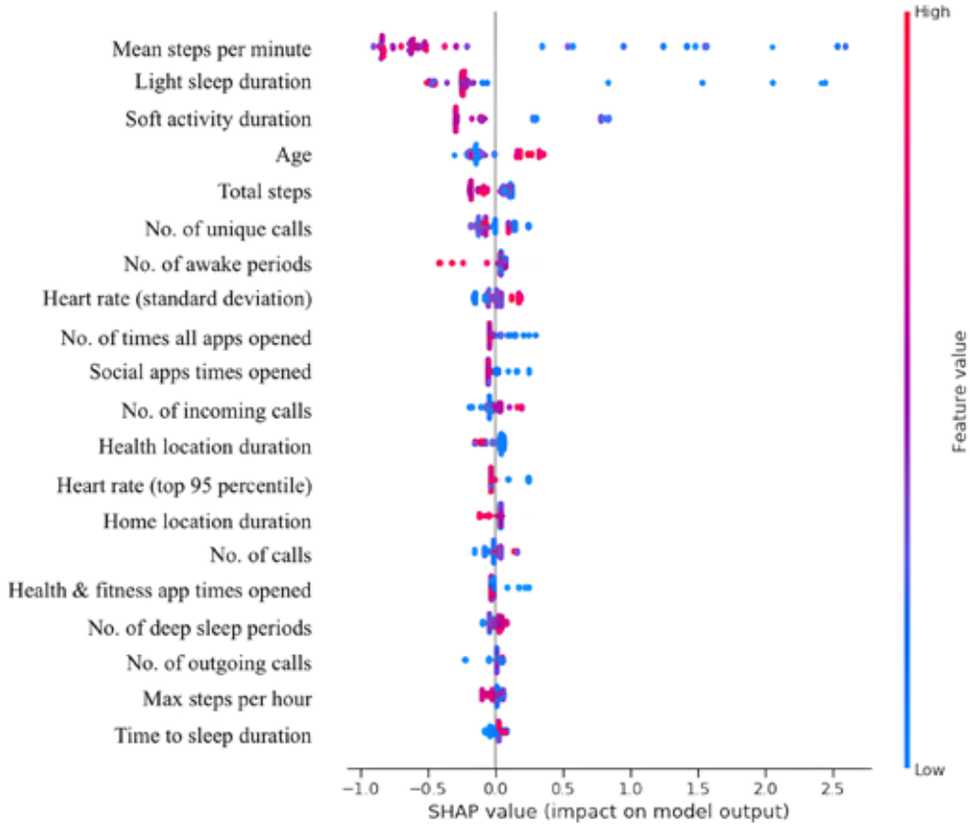
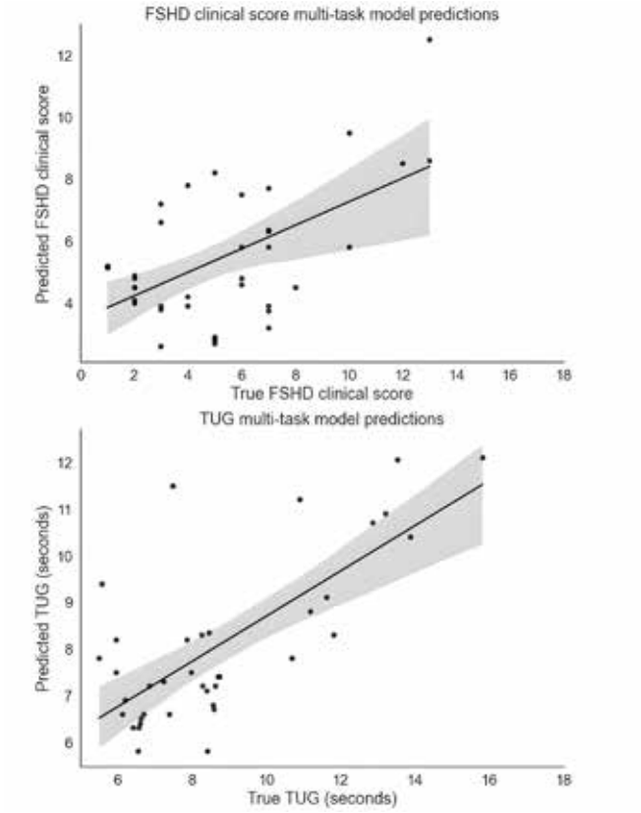
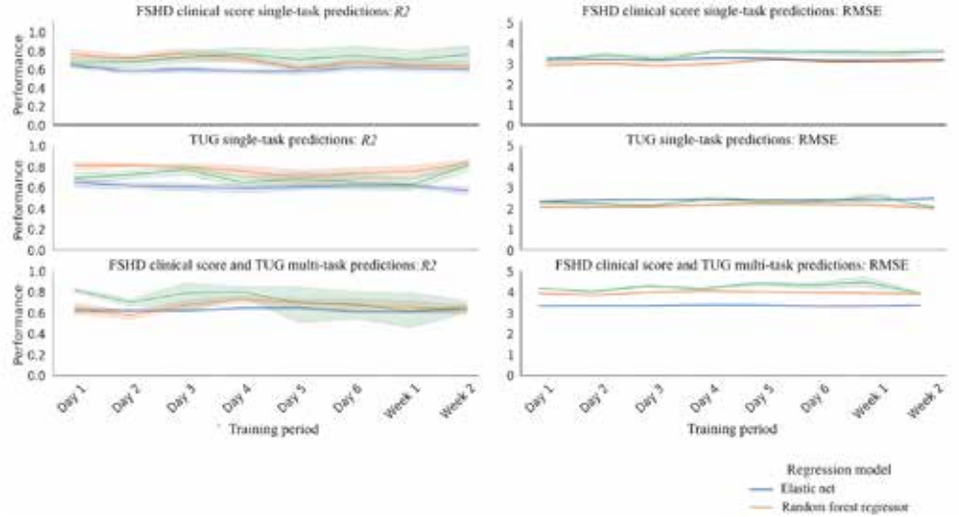


**FIGURE 4** Scatterplot of the estimated FSHD Clinical Scores and TUG times in relation to the actual FSHD Clinical Scores and TUG using the multi-task learning regression model. The lines represent the regression lines with a 95% CI band.



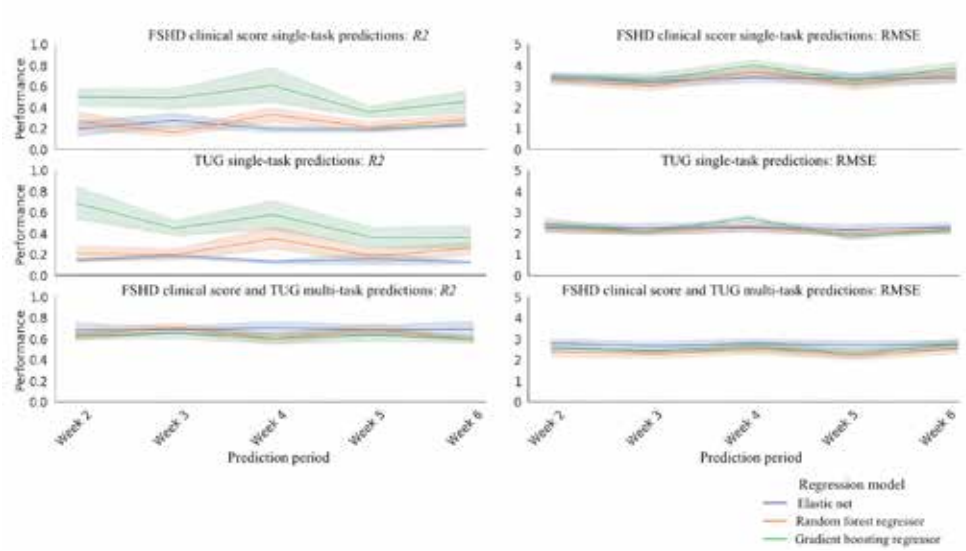

*FSHD: facioscapulohumeral muscular dystrophy; TUG: Timed Up and Go.*

**FIGURE 5** Evaluating the performance of the single-task FSHD Clinical Score, TUG, and the multitask FSHD Clinical Score and TUG regression models trained on an incrementally increasing time window. The colored lines represent the 3 types of regression models trained on the data (Elastic Net, Random Forest Regressor, and Gradient Boosting Regressor). For each model and each incremental time window, the top and bottom plots show the R2 and RMSE, respectively. The lines represent the median performance, and the bands represent the 95% CI.




*FSHD: facioscapulohumeral muscular dystrophy; RMSE: root mean square error; TUG: Timed Up and Go.*

**FIGURE 6** Evaluating the performance of the single-task FSHD Clinical Score, TUG, and the multitask FSHD Clinical Score and TUG regression models trained on the first week of data to estimate symptom severity for the subsequent weeks. The colored lines represent the 3 types of regression models trained on the data (Elastic Net, Random Forest Regressor, and Gradient Boosting Regressor). For each model and each week, the top and bottom plots show the R2 and RMSE respectively. The lines represent the median performance, and the bands represent the 95% CI.




*FSHD: facioscapulohumeral muscular dystrophy; RMSE: root mean square error; TUG: Timed Up and Go.*