Tracking Fatigue and Health State in Multiple Sclerosis Patients Using Connnected Wellness Devices

CATHERINE TONG, Department of Computer Science, University of Oxford
MATTHEW CRANER, Nuffield Department of Clinical Neurosciences, University of Oxford
MATTHIEU VEGREVILLE, Withings, Issy-les-Moulineaux, France
NICHOLAS D. LANE, Department of Computer Science, University of Oxford

Multiple Sclerosis requires long-term disease management, but tracking patients through the use of clinical surveys is hindered by high costs and patient burden. In this work, we investigate the feasibility of using data from ubiquitous sensing to predict MS patients' fatigue and health status, as measured by the *Fatigue Severity Scale (FSS)* and *EQ-5D* index. We collected data from 198 MS patients who are given connected wellness devices for over 6 months. We examine how accurately can the collected data predict reported FSS and EQ-5D scores per patient using an ensemble of regressors. In predicting for both FSS and EQ-5D, we are able to achieve errors aligning with the instrument' standard measurement error (SEM), as well as strong and significant correlations between predicted and ground truth values. We also show a simple adaptation method that greatly reduces prediction errors through the use of just 1 user-supplied ground truth datapoint. For FSS (SEM 0.7), the universal model predicts weekly scores with MAE 1.00, while an adapted model predicts with MAE 0.58. For EQ-5D (SEM 0.093), the universal model predicts weekly scores with MAE 0.097, while an adapted model predicts with MAE 0.065. Our study represents the first sets of results showing that fatigue and health state of MS patients can be measured using data from connected wellness devices and a small number of background features, with promising prediction performance with errors within the accepted range of error in the widely used clinically-validated questionnaires. Future extensions and potential applications of our results can positively impact MS patient disease management and support clinical research.

CCS Concepts: • **Human-centered computing → Empirical studies in ubiquitous and mobile computing**; • **Computing methodologies → Supervised learning by regression**; • **Applied computing → Health informatics**.

Additional Key Words and Phrases: Health data analysis, Machine Learning, Multimodal prediction, Connected devices

ACM Reference Format:
Catherine Tong, Matthew Craner, Matthieu Vegreville, and Nicholas D. Lane. 2019. Tracking Fatigue and Health State in Multiple Sclerosis Patients Using Connnected Wellness Devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 106 (September 2019), 19 pages. https://doi.org/10.1145/3351264

1 INTRODUCTION AND BACKGROUND

Over 2.5 million people worldwide are affected by Multiple Sclerosis (MS), in many countries it is the most common cause of neurological disability in young adults [19]. Multiple Sclerosis is associated with significant health-related and economic burden to the quality of life [7]. As there is currently no cure for MS, the complexity of MS demands that patients be active in their management of symptoms and receive support for it [5].

Authors' addresses: Catherine Tong, eu.tong@cs.ox.ac.uk, Department of Computer Science, University of Oxford; Matthew Craner, Nuffield Department of Clinical Neurosciences, University of Oxford; Matthieu Vegreville, Withings, Issy-les-Moulineaux, France; Nicholas D. Lane, Department of Computer Science, University of Oxford.

2474-9567/2019/9-ART106 $15.00
https://doi.org/10.1145/3351264

Managing MS mainly involves keeping track of clinical outcomes. Overwhelming *fatigue* is one of the most commonly reported symptom of MS [6]. MS fatigue is very different from normal forms of fatigue and has a profound impact on patients' quality of life [29]. Faitgue is monitored by patients for self-management of the symptom, and is also monitored through survey instruments in clinical trials to demonstrate the efficacy of MS drugs [8]. A large body of research also exists to address correlations between fatigue and other physiological factors and behavioral factors (e.g exercise [29]). However, since data collection is costly and slow, in longitudinal studies of MS fatigue, FSS may only be collected annually for 2-3 years.

For medical and economic researchers, another important measurement to collect is the *Health-Related Quality of Life (HRQoL)*. Policymakers make use of HRQoL to evaluate the effectiveness of health interventions in order to allocate resources effectively [3]. For MS, this means allocating more resources to treatment that leads to a delay in patients' progression to permanent disability, over treatment which only temporarily avoids disability [7].

Longitudinal tracking of these metrics are consequential to the resource allocation decision making, however frequent data collection involves high costs so researchers typically rely on economic modelling techniques [7]. Motivated by this, we propose a study looking into tracking fatigue and health state unobtrusively in the daily lives of MS patients through the use of connected wellness devices, in addition to a small number of day questions and background features. We are interested in applying machine learning techniques in delivering accurate predictions of patients' fatigue and health state over time in a low user burden fashion, which could support patient's self-management and further understanding of more dynamic patterns of these health indicators.

We conducted a 6-month study of 198 MS patients which collected their behavioral, physiological and self-reported health information. Each patient was given 3 connected wellness devices (sleep tracker, smart scale, smart watch) to use in their natural environments, these collect daily behavioral and physiological data of the patient. Each patient completed a background questionnaire; In addition, through a companion app, patients may fill in daily evaluations of fatigue and sleep quality, as well as weekly questionnaires containing the FSS and EQ-5D. In this work, we focus on the task of predicting FSS and EQ-5D scores reported each week by each patient. We propose using an ensemble of predictors per data source to handle issues arising from missing sensor modalities and perform a series of experiments to demonstrate the suitability of our approach. We also perform an additional task where we predict the mean FSS and EQ-5D scores spanning all weeks reported by a patient.

The main contributions of this work are as follows:

- First study predicting MS patients' reported fatigue severity scale (FSS) using data from connected devices, background information and daily questions at weekly intervals. We show that FSS (instrument error SEM 0.7) can be predicted with MAE 1.00 (r=0.58) with an unadapted ensemble model, which requires zero FSS survey response from the target patient at test time. Our adapted model (calibrated using 1 FSS survey response from the target patient) can predict FSS with MAE 0.58 (r=0.81).
- First study predicting MS patients' reported EQ-5D health state using data from connected devices, background information and daily questions at weekly intervals. We show that EQ-5D (SEM 0.093) can be predicted with MAE 0.097 (r=0.43) with an unadapted ensemble model, which requires zero EQ-5D survey response from the target patient at test time. Our adapted model (calibrated using 1 ED-5D survey response from the target patient) can predict EQ-5D with MAE 0.065 (r=0.77). These results are promising as they are within scales of the instrument error and also correlate significantly and strongly with ground truth.
- We also predict the per-participant averaged scores for FSS and EQ-5D over the 6-month study period, which represents a more stable view. We show that the mean FSS can be predicted with MAE 0.93 (r=48), and mean EQ-5D can be predicted with MAE 0.091 (r=39). Again, these are *universal* models able to be used by any patients without them providing any FSS or EQ-5D survey response.
- Discovery of important contextual background features of multiple sclerosis patients (e.g. age, disability level) which are seen by our models as important in predicting FSS and EQ-5D scores.

- Investigation of the sensitivity in varying different design choices of our prediction framework, such as time window length and considered data sources.
- Discussion and summary of insights from current study in related directions that we believe will be valuable for evolving the design of predictors of MS-related outcomes in the future.

2 RELATED WORK

This is increasing interest in using sensing technology to aid Multiple Sclerosis disease management. Numerous pilot studies have been carried out to explore the feasibility of deploying passive sensing in MS patient's natural environments so that longitudinal assessment of their symptoms could be performed. However, the focus of many such studies tend to be in exploring the feasibility of data collection using passive sensing and analyzing the correlations. [12] explored using real-time depth sensors at home to identify gait problems and falls in 21 MS patients. [21] used activity sensors to determine the physical activity level of 11 MS patients, and correlated their daily activity fluctuations to disability changes in MS patients. Certain technologies have already been developed for MS monitoring and management which are complementary to traditional in-clinic approaches. *MS Mosaic* is a smartphone app that allows MS patients to track their symptoms over time by manually reporting their symptoms (e.g. Fatigue) and also integrating health and fitness data available on smartphones to monitor symptom triggers. *Floodlight* is another smartphone app, which tracks changes in MS over time through 'active tests' of patients performing daily simple tasks on the app. [2] have reported the results from the same dataset studied in our work, but with basic analytic results about correlations between expected fatigue changes and behavioral measurements. The bulk of prior studies demonstrate the monitoring of user behaviors like steps, sleep hours etc. that are indirect measurements of symptoms and without any way to translate into a clinical understanding currently. As far as we know, no technology has been developed for tracking fatigue and health state as measured in terms of the FSS and EQ-5D-5L.

Closely related to our work are which have utilised passive sensing in monitoring symptoms for general health and other diseases, many of which have been carried out to reproduce predictions for clinically-validated self-report survey instruments [16]. More recently, Wang et al develops a prediction system that tracks schizophrenia symptoms based on a standard instrument using passive sensing from mobile phones, they were able to accurately predict reported schizophrenia scores using Gradient Boosted Regression Trees (GBRT) [26]. Wang et al proposed a set symptom features from wearables and smart phones to track college student's level of depression [27], they used generalized linear mixed model (GLMM) to predict self-reported depression scores and found correlations between their proposed symptom features and depression scores. Other than disease monitoring, a number of studies have also been carried looking into monitoring of more general wellness indicator, since this could be applied to the general non-clinical population, the sample dataset sizes studied could be much bigger, therefore also allowing more advanced techniques to be applied (e.g. deep neural networks). Veličković et al analyzes multimodal time-series data and predicts the ability to achieve weight objective for users of smart connected devices using deep long-short-term memory architectures [25].

3 OUR STUDY AND DATA

This section introduces the study carried out and the dataset.

3.1 Multiple Sclerosis Study

We conducted a 6-month study with 198 patients diagnosed with Multiple Sclerosis from November 2016 to May 2017. The participants were based in the United States and were recruited online for a study to understand their MS condition through use of connected health devices, which may be kept by the participants after the study. Data from each participant is collected through the following means:

(1) Entries to a screener survey
(2) Measurements collected through use of connected wellness devices
(3) Entries to *week* questionnaires

In addition, entries to *day* questions are optionally provided by participants - we also use this in our framework to learn patterns but, as we shall describe in later sections, this is not a necessary piece of information to collect aside from the device data measurements for the considered predictive modelling tasks, and we do not assume the general user to provide this information.

To simulate conditions corresponding to natural usage, participants are free to deploy the devices and fill in information as frequently or infrequently as they prefer. All data recorded during the study was analyzed anonymously. We now describe the collected data in more detail.

Background questions. Before the commencement of the study, all participants completed a screener survey to give background information about their health. Information from these background questions is treated as static, forming the only non-time-varying component in our analyzed dataset. The survey collected self-reported socio-demographic variables and clinical variables. All of the collected variables are categorical with the exception of age, height and number of years with MS.

- *Socio-demographic variables.* This collects data on demographics, the patients' understanding of MS, use of technology to monitor health, as well as the patient's mode of financing treatment.
- *Clinical variables.* This collects data on the patient's MS, symptoms experienced, other co-existing diseases, and their treatment method. Variables from symptoms and co-existing diseases are dichotomized into yes or no, so only a binary indication of whether symptoms/disease are present is available.

Device measurements. Each participant received three products from the *Withings* range: a sleep tracker, a weighing scale and a smart watch [1]. Measurements from these devices recorded within the study period (total of 189 days) were extracted for analysis. Device usage varies, with the participants contributing a mean of 122.8 days with at least one valid device measurement; three participants produced no valid device data at all during the study period.

The wellness devices provide measurements relating to sleep, activity and other physiological information, which are well-placed to capture the time-varying quality of MS. All measurements are in daily resolutions, and were inferred from the *Withings* applications through use of the devices. An example measurement is the step count of a day, as opposed to a accelerometer time series throughout the day. In this example, the smart watch's commercial-grade machine learning algorithms have been involved in producing this measurement from raw signals; These algorithms are supervised learning methods trained from hundreds of controlled subject data collections, which are then released and used on in-the-wild data for millions of *Withings* users.

- *Withings Aura (sleep tracker).* Aura consists of two devices, a sleep sensory pad underneath the mattress, and a bedside device with environmental sensors which also act as an alarm clock, speaker and lamp. We consider sleep measurements recorded by Aura.
- *Withings Activité Steel (smart watch).* The Activité Steel is a wearable device which tracks activities (gait, running, swimming and sleeping). We only collect data on gait and sleeping as proxies to participants' activities and sleep respectively. We keep sleep measurements recorded by Aura and the watch separate, in accordance with the procedure taken in similar situations [28].
- *WIthings Body Cardio (smart scale).* This scale provides readings on weight, composition and heart health. The scale also provides pulse wave velocity (PWV), the rate at which blood pressure pulse propagates through the circulatory system, which is an important clinical parameter for evaluating cardiovascular risk [15]. The scale can be set up for separate users in case the participant's family also use it, but only data from the participant's profile was collected.

- *Health Mate App.* This is a commercial app for managing *Withings* devices currently used by millions of users. A user profile syncs data from the devices and allows the participants to review their data and set goals [13]. In our study, the app also sends notifications to the participants and act as a portal of access to the questionnaires.

Week questionnaires. Each participant receives a notification from the app every week prompting him/her to complete a *week* questionnaire. The questionnaire is formed of two standard questionnaire instruments: FSS and EQ-5D-5L. Both are clinically validated instrument widely used in MS study and management. In our dataset, 1693 weekly surveys were completed, but 24 participants did not complete any weekly surveys at all. We describe the questionnaires in detail:

- *Fatigue Severity Scale (FSS).* The FSS is a 9-item questionnaire. The items relate to fatigue severity, its effects on the subject's activities and lifestyle, as well as its emotional burden. The responses are rated on a 7-point Likert scale, and the overall FSS score is taken as the mean of the 9 responses. The overall score ranges from 1 to 7, where a higher score corresponds to higher levels of fatigue. A number of studies have been carried out validating the FSS instrument, the standard error of measurement (SEM) of FSS was suggested to be 0.7 points, meaning a change of less than 0.7 points may be due to measurement error of the instrument [11].
- *EQ-5D-5L index value.* EQ-5D covers 5 dimensions of health: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression [24]. The EQ-5D-5L index is a single index value computed from the EQ-5D-5L system, which describes a person's health state relative to others in the country. The conversion from the 5 dimensions to a single value is done using publicly available country-specific value sets, and in this case, US-specific [23]. The 5Q-5D-5L index value ranges from 0 to 1, with 0 meaning death and 1 complete health [17]. The SEM of FSS was suggested to be 0.093 [14].

A complete response to every question on the questionnaire (from both instruments) is required to qualify as a valid response. We use the FFS and EQ-5D-5L index scores as our ground truth. We use the notation FSS-W and EQ-W to refer to the FSS score and EQ-5D-5L index value obtained from each weekly survey respectively; and we use FSS-M and EQ-M to refer to the mean per-participant score of FSS-Ws and EQ-Ws obtained from all weekly surveys completed by each participant. In total, 1693 weekly surveys were completed, however 24 participants did not complete any weekly surveys at all.

Daily questions. The daily questions are attempted by 73 participants. The low participation rate (fewer than half the subjects) may be because we framed the daily questions as an optional part of the study. At the beginning of the study, every participant received a welcome message from the app which asked if he/she would like to receive daily notifications about these items for the rest of the study. A daily notification setting might have seemed excessive to the participants so only 73 subjects accepted.

These daily questions are informal prompts (i.e. not clinically validated instruments like the week questionnaires) asking patients to rate their fatigue and sleep quality. The subject may indicate this on a 5-point Likert scale, from No to Extreme Fatigue, and Very Good to Very Bad Sleep. Each question is prompted independently through the app, so a patient can answer only one or both of these questions per day, we thus treat data from each question as a separate source. In total, 3316 daily questions are reported by 73 participants.

Table 1. Our considered dataset of 151 patients and 1401 labelled examples. In (b), within each source s, $N_{p(s)}$ is the number of patients, $N_{w(s)}$ the number of examples and $F_{(s)}$ the number of features. Background features are a special case as they are always present so $N_{p,w(background)} = N_{p,w}$.

(a) Patient demographics and reported scores.

Item	Distribution
Age	From 22 to 75 years, mean 44 years
Gender	141 females, 10 males
Type of MS	146 RRMS, 3 SPMS, 1 PPMS, 1 PRMS
Years with MS	From 0 to 44 years, mean 10 years
Disability level	48 normal, 52 mild disability, 34 moderate disability, 34 gait disability, 29 early cane
FSS score	mean 5.14, standard deviation 1.41
EQ-5D score	mean 0.745, standard deviation 0.147

(b) Data distribution from each source.

Source	$N_{p(s)}$	$N_{w(s)}$	$F_{(s)}$
Aura	118	914	55
Scale	94	609	35
Watch	87	501	45
Fatigue q.	62	556	5
Sleep q.	62	510	5
Background q.	151	1401	77

3.2 Dataset

We describe the steps taken in creating a dataset for training a machine learning model to predict FSS and EQ-5D-5L. Our goal is to construct a labelled dataset $\mathcal{D} = (\mathcal{X}, \mathcal{Y})$, with N_w labelled examples. $\mathcal{X}$ are the aggregated features from different data sources and $\mathcal{Y}$ are the score labels. We preprocess the data, then identify time windows for aggregation into features, and collect ground truth.

Preprocessing. We set sensible criteria on to remove obvious outliers (e.g. unrealistic heights, sleep duration of over 20 hours) and replace with the population mean.

Ground Truth. To obtain ground truth, we consider all completed week questionnaires and obtain the corresponding overall FSS score (FSS-W) and EQ-5D-5L index (EQ-W) for each questionnaire.

Feature Set Construction.

To obtain features from time-varying data sources (devices and questions), we consider for each week questionnaire a time window of d days up to the day that it was filled out.

Our inclusion criteria is as follows: We first disregard 4 subjects who have fewer than 90 days of device measurements and 24 subjects who did not complete any week questionnaire. We then disregard time sequences which do not have entries from at least 1 time-varying source for at least 1 day within the time window. Finally, we only consider the remaining valid time sequences and their associated subjects. Setting $d = 10$, this results in $N_w = 1401$ labelled sequences from $N_p = 151$ subjects. Table 1 describes the demographics and ground truth provided by these subjects.

We compute 5 features from each sequence: maximum, minimum, mean, standard deviation and count. The maxima, minima, means and standard deviations describe the patient's extremes, average, variation of behaviour and physique during the time window. Finally the count captures the level of usage of a device or a function by the patient, for example the count of sleep duration recorded by the watch is the number of times the subject wore the watch to sleep.

Apart from the time-varying sources, we also have static information from each subject through their responses to the background questions. We include these features into consideration for every labelled example in our dataset. An overview of the feature set construction procedure is shown in Figure 1.

Source-Specific View. Since the patients rarely use all devices during one time window, there are often many missing features in one labelled example due to absent sources. To avoid this, we take a source-specific view and denote $\mathcal{D} = \{\mathcal{D}_{(s)}\}, s \in \mathcal{S}$, where $\mathcal{S}$ is a set of considered sources. Unless stated otherwise, $\mathcal{S}$ are 6 data

sources: Aura, Scale, Watch, Fatigue question, Sleep question and Background questions. We use the subscript $_{(s)}$ to denote items that are source-specific. In this view, the dataset $\mathcal{D}_{(s)}$ only contains the labelled examples in $\mathcal{D}$ where features from s must be present, specifically its feature set is $\mathcal{X}_{(s)} \in \mathcal{R}^{N_w(s) \times F_{(s)}}$, where $N_w(s) \leq N_w$ is the number of examples with features from source s and $F_{(s)}$ is the total number of features from source s prior to any feature selection. Table 1 lists the source-specific numbers.

4 OUR LEARNING FRAMEWORK

This section introduces our approach in modelling patients' data to predict their reported fatigue and health state.

4.1 Problem Statement

We consider a regression task to predict a patient's FSS-W and EQ-W scores. With a feature set computed from a patient's unobtrusive data collected from wellness devices and self-report data, the goal is to assign to it a corresponding ground truth survey score. In particular, we predict the patient's FSS-W and EQ-W scores; we treat these as two independently tasks which adheres to the same learning framework.

We assume a leave-one-subject-out cross-validation setting, so the model is trained with $\mathcal{D}^{train} = (\mathcal{X}^{train}, \mathcal{Y}^{train})$, the training dataset of examples belonging to all but the test subject, and tested on $\mathcal{D}^{test} = (\mathcal{X}^{test}, \mathcal{Y}^{test})$, which consists only of examples belonging to the test subject. In other words, at test time, the trained model would require the following from a test user to make a prediction:

- measurements from connected devices
- a small number of background features
- entries to daily questions (only if these are provided by the test user)
- zero FSS or EQ-5D survey response (as the model is trained only with examples from train-set users)

In case of model adaptation, one FSS or EQ-5D survey response from the test user will be required.

The results from a leave-one-subject-out setting shows the prediction performance of predicting a new patient whose data has never been seen by the trained model. We adopt this setting because it is more aligned to the potential use case where patients beginning to monitor their health passively can do so immediately as they start generating data on their wellness devices.

4.2 Our Framework

In each fold of leave-one-subject-out cross validation, we perform the following:

(1) Perform feature selection on each source-specific dataset $\mathcal{D}_{(s)}^{train}$;
(2) Train a regression model $\mathcal{M}_{(s)}$ on each $\mathcal{D}_{(s)}^{train}$, using only the selected features;
(3) Test each model $\mathcal{M}_{(s)}$ on the corresponding $\mathcal{X}_{(s)}^{test}$ to give prediction $\hat{\mathcal{Y}}_{(s)}$;
(4) Aggregate predicted scores $\hat{\mathcal{Y}}_{(s)}$ from all sources to output $\hat{\mathcal{Y}}$;
(5) (If Adapt) Transform predictions using reported score of the test subject's first labelled example $\mathcal{Y}_0^{test}$.

We now discuss our approach in more detail.

Ensemble Approach to Missing Data. As features from multiple modalities are present to a varying degree in our dataset, it is desirable to learn accurate models which can leverage information from multiple modalities while considering as many datapoints as possible. We adopt an ensemble learning structure formed of predictors which focus on different spectrums of the datasset with high data concentration. In effect this is an ensemble of base predictors trained specifically with features coming from the same modality or source. As illustrated in Figure 1, a major advantage of using an ensemble is that datapoints with different missing modalities can be considered.

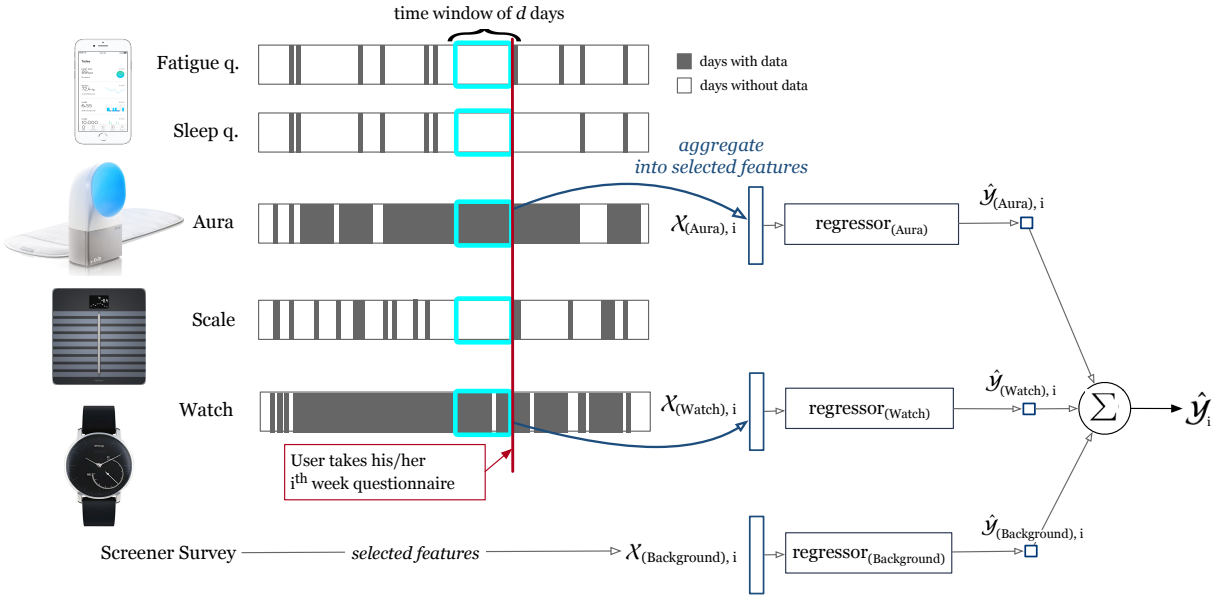

Fig. 1. Schematics showing our learning pipeline for one labelled example. The left part of the figure illustrates the feature construction procedure from 5 time-varying sources (aura, scale, watch, fatigue q., sleep q.) and 1 static source (screener survey). Each horizontal bar represents the sequence of measurements given by each time-varying source from the beginning till the end of the study. We consider a time window of d days up to the date that the subject completes a week questionnaire, and if there is enough data within the window, it is aggregated to produced selected features. The right part of the figure illustrates the ensemble model which then proceeds to perform regression per source and finally combines the predictions to give one predicted score. In the above example, only features from aura, watch and background questions are present. Images of Withings devices sourced from [1].

Base Regression Model. The specific base regressor model adopted in our study is the AdaBoost Regressors [4]. Adaboost Regressors itself is an ensemble method which fits predictors sequentially on a given dataset by setting weights to the predictors and the datapoints such that the subsequent predictors focus more on difficult cases. As Adaboost is also a tree-based method, it produces a value of relative importance for each input feature, this is known as the importance vector to be used for feature selection.

Feature Selection. We perform feature selection in every fold of leave-one-subject-out cross validation. In each fold, the train dataset of every source, $\mathcal{D}_{(s)}^{train}$, is fed into an Adaboost which generates an importance vector. We obtain the selected features for each source by performing k iterations where we retain features with an importance above the mean importance of the remaining features. The higher k is set, the fewer the number of resulting features. Since we have the highest number of features to begin with for background questions, we set k to be higher for background questions; this gives better performance as we shall show.

Aggregating Source-specific Predictions. In combining the predictions given by each modality, we put to use the resulting feature importance vector generated in the feature selection phase. We compute the mean feature importance per source and use it as a weighting factor. We normalize the mean importance with *softmax*

function, the final prediction of a labelled example $\mathcal{X}_i$ is given by:

$$\hat{\mathcal{Y}}_i = \sum_{s \in \Gamma_i} w_s \hat{\mathcal{Y}}_{(s)i} \quad \text{where} \quad w_s = \frac{\exp(\hat{\mathcal{Y}}_{(s)i} \bar{f}_s)}{\sum_{s \in \Gamma_i} \exp(\hat{\mathcal{Y}}_{(s)i} \bar{f}_s)} \tag{1}$$

where $\bar{f}_s$ is the averaged importance of features from source s, and Γ_i is the set of sources that are present in $\mathcal{X}_i^{test}$.

Adaptation to the individual. We consider adaptation techniques to improve generalization of the model through use of small amounts of data specific to an individual.

We propose a simple adaptation strategy involving a translation using each subject's first recorded residual error. This is motivated by the observation that many predicted scores were consistently off by some similar value for datapoints belonging to the same subject. The simple transformation introduces a bias using the subject's first completed weekly questionnaire, and then applies an adjustment to the bias onto all subsequent predictions given to that subject. In effect, we apply a scalar translation of the prediction according to the residual error of the first week. For predictions in the subsequent weeks $i \geq 1$,

$$\hat{\mathcal{Y}}_i' = \hat{\mathcal{Y}}_i + b_0 \quad \text{where} \quad b_0 = \mathcal{Y}_0^{test} - \hat{\mathcal{Y}}_0 \tag{2}$$

note, we order datapoints belonging to the same subject chronologically in $\mathcal{Y}$ such that $\mathcal{Y}_0^{test}$ refers to the test subject's first reported label.

We also consider an alternate adaptation strategy based on Gaussian Mixture Models (GMM) [18]. The method was originally proposed for speaker classification, and we adopted the formulation for Gaussian Mixture Regression (GMR) models instead [22] (the base regressor in our ensemble model becomes GMR). The core idea of the adaptation is to update the parameters of a general GMR using Maximum A Posterior according to some small amounts of labelled examples from the test user. As shown in later experiments, this adaptation strategy gives worse performance than the residual error method.

5 TRACKING MS HEALTH DYNAMICS: EXPERIMENTS AND ANALYSIS

We perform a series of experiments to examine our model's prediction of patients' FSS-W and EQ-W scores.

5.1 Experimental Setup

Model Settings. In the rest of this section, we refer to the ensemble model introduced in Section 4.2 as 'our model', and its residual error adapted version as 'our model (res. error adapted)'. Unless stated otherwise, we adopt the unadapted model with the following settings by default. For the purpose of fully exploiting our dataset, we use data from all sources by default $\mathcal{S}$ = {Aura, Scale, Watch, Fatigue q., Sleep Q., Background q.}. As a result, our ensemble model consists of 6 base regressors, each as an adaboost regressor. We set the considered time window length to $d = 10$ days. Our feature selection process performs $k = 1$ iteration for all sources except for background questions, for which we use $k_{backgroundq.} = 3$ in predicting FSS-W and $k_{backgroundq.} = 2$ for predicting EQ-W.

Evaluation. As baseline, we use a hypothesized model which always predicts the global mean (mean of $\mathcal{Y}^{train}$ in each fold). We evaluate the prediction performance with leave-one-subject-out cross validation, and report using mean absolute error (MAE) and Pearson correlation coefficient (r). MAE assesses the absolute difference between predicted and ground truth scores across the entire population. For FSS this is with reference to a score that ranges from 1 to 7, and for EQ-5D from 0 to 1. Pearson correlation coefficient r gives the correlations between predicted and ground truth scores across the entire population, where p-value indicates the statistical significance of the coefficient. Throughout our text, we use * to denote significant correlations with $p < 0.0001$.

Table 2. Results of our model predicting FSS-W and EQ-W.

Model	FSS-W		EQ-W	
	MAE	r	MAE	r
Our model	1.00	0.58*	0.097	0.43*
Our model (Res. error adapted)	0.58	0.81*	0.065	0.77*
Baseline	1.12	-	0.110	-

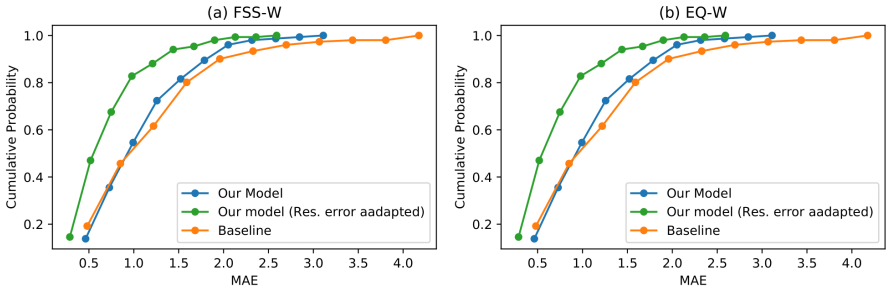

Fig. 2. CDFs of per-patient MAE obtained in predicting FSS-W and EQ-W.

5.2 Prediction Performance

Table ?? shows the MAE and Pearson correlation for our models in predicting the weekly scores FSS-W and EQ-W. In both tasks, both the unadapted and adapted models outperform baseline and achieve significant and strong correlation with ground truth. The adapted models achieve the best results with MAE 0.58 for FSS-W and MAE 0.065 for EQ-W respectively, which are both below the inherent error of the instruments (SEM 0.7 for FSS and SEM 0.093 for EQ-5D). The result shows our existing pipeline can accurately predict patients' FSS-W and EQ-W scores. Figure 2 shows a comparison of the cumulative distribution of the per-person mean absolute error in greater detail, where we see that the residual error adapted model performs better than the unadapted mode, but both outperform the baseline.

Within-person time series view. Figure 3 shows the performance of the models when viewed on a week-to-week level within a patient. The 8 patients presented have achieved the best, 75th percentile, 25th percentile and worst result using our unadapted model for each task, where the n^{th} percentile means the patient's MAE is better than n% of the studied population. We use these patients to illustrate cases where our model performs better and worse.

Patient (a-iv) , where the FSS-W model performs worst, might be a case of unreliable ground truth as the patient fills out the questionnaire the same way every week and reported the same score of 1.0 for 16 weeks. While our unadapted model also predicted a more or less stable weekly result for this patient, our residual-error adaptation method was able to correct the offset and reduced MAE from 3.1 to 0.08. In our dataset we observe 5 other patients who reported the same FSS-W value every week, and their MAEs were reduced by an average of 43% after residual-error adaptation was applied.

We further observe that both models perform worse for patients who reported lower scores (Patients (a-iii), (a-iv) and b(iv)), which is related to the unbalanced distribution of scores in our dataset. Our dataset is skewed towards higher scores (90% of the reported scores are ≥ 4 for FSS and ≥ 0.5 for EQ-5D), thus the models are undertrained for lower scores and result in an overestimation. Therefore, it is especially in these low-true-score

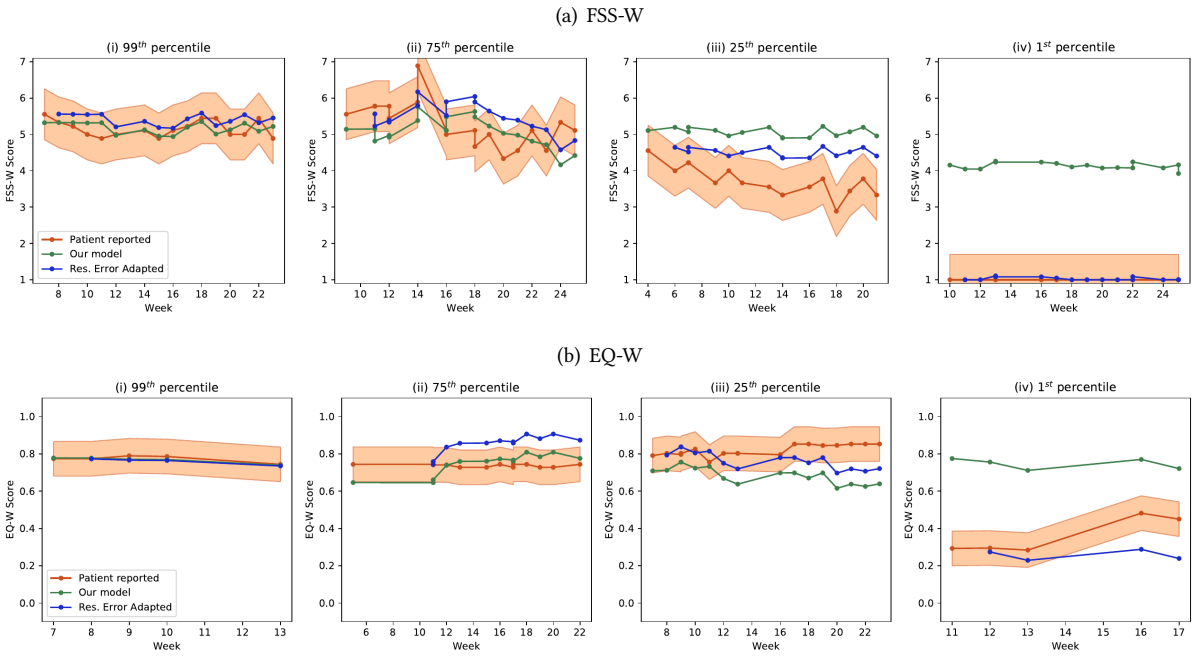

Fig. 3. Time series view of predicted vs reported FSS-W and EQ-W scores. An uncertain region is coloured in orange to denote the standard error of the instrument (SEM). In each row, from left to right are the patients with the best to the worst predicted MAEs. The horizontal axis denotes the number of weeks since the patient first started using any of the provided devices.

cases that the residual adaptation method work best, since the residual error of the first week would account for an estimation of the offset between high-score and low-score patients.

Matching polarity in week-to-week changes. Through visual inspection of Figure 3 we observe that the predictions made by our models are also able to reflect the week-to-week fluctuations of the weekly score; for example, even in the most challenging cases the model was able to detect changes in polarity of the trend of user scores from week to week.

For the prediction results to be useful to patients on a week-to-week basis, we investigate whether the polarity of the changes in consecutive predicted scores reflects the truth. For example, if our model predicts that a patient has FSS-W 7.0 in week 0 and FSS-W 6.0 in week 1, the predictions can capture the trend of his/her fatigue if the patient also reported a higher score in week 0 than week 1. To investigate this, we consider significant gaps ($|\mathcal{Y}_{i+1} - \mathcal{Y}_i| \geq \text{SEM}$) in consecutive scores, and report the polarity accuracy as the percentage of times that the polarity of the predicted gap matches that of the true gap. Figure 4 shows that this accuracy increases as the predicted gap increases, meaning the model is increasingly confident as it predicts larger gaps. As illustrated by the vertical lines in Figure 4, we also note that the averaged predicted and reported gaps are very similar, for FSS-W they are within 0.9% and for EQ-W 5.0%. This results gives us confidence to track these scores every week.

Available Sources. Figure 5 shows the prediction errors for the datapoints when features from all sources (i.e. all daily questions and all devices), only daily questions or only devices are available. For both FSS-W and EQ-W, we find that datapoints considering all sources can be predicted with the lowest MAEs, suggesting that missing

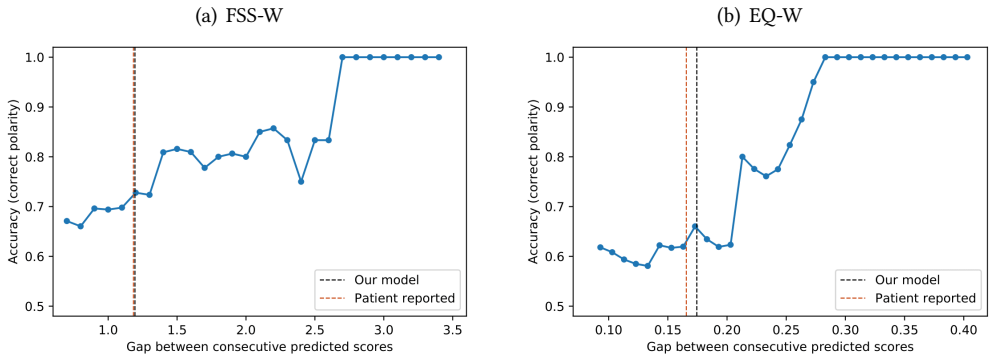

Fig. 4. Comparing the polarity of gaps found in our predicted and ground truth scores. The horizontal axis is g, the magnitude of a gaps between consecutive predicted scores, and the vertical axis is the percentage of times that a gap with magnitude $\geq g$ correspondes to the correct polarity. The orange vertical line is the mean magnitude of significant gaps in the ground truth scores, while the black line is that in the predictions.

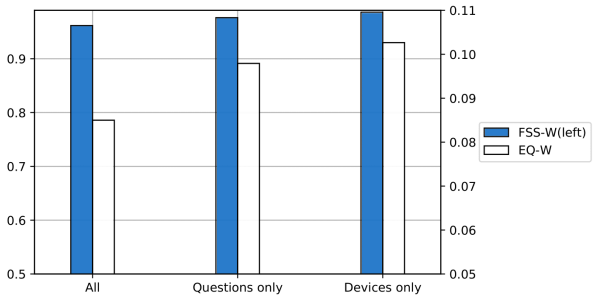

Fig. 5. Mean absolute error for weeks when features from all sources, only questions or only devices are available.

modalities does play an role and the model benefits from having features from different sources. However, the difference in MAEs between 'all' and 'devices' is small in both cases (2.6% for FSS-W and 2.1% for EQ-W), this suggests the benefits brought by features from daily questions is limited and the models can perform well even when they are absent.

5.3 Sensitivity Analysis

Importance of background features. As described in Section 3.1, 77 background features were collected for each subject. In Table 3, we vary the number of background features taken as input to our model, through adjusting the number of feature selection iteration steps k, which we found to be optimal when set at $k = 3$ for FSS-W and $k = 2$ for EQ-W. The selected background features are listed in Table ??.

Having background features is crucial for predicting FSS-W and EQ-W in MS patients, as we observe that MAEs increases by an average of 12% when they are absent. However, it is interesting to note that only a small number of background features (4 to 6 features) is needed for good performance. These selected background features (e.g. disability level) can be understood as providing the most important contextual information about each patient; This is necessary for our models to effectively account for their individual differences, and as a

Table 3. Importance of background features. Selected features here refers to the features selected in more than half of the folds in the leave-one-subject-out cross validation.

Include		FSS-W			EQ-W		
background features	k	No. of selected features	MAE	r	No. of selected features	MAE	r
✓	1	23	1.02	0.53*	22	0.098	0.47*
✓	2	8	1.02	0.58*	6	**0.097**	**0.43***
✓	3	4	**1.00**	**0.58***	2	0.098	0.40*
	-	-	1.12	0.29*	-	0.110	0.10*

Table 4. Comparing considered devices, $^{\#}p < 0.05$.

Considered	FSS-W		EQ-W	
sources	MAE	r	MAE	r
Background q.	1.22	-0.43#	0.175	0.31*
Aura, background q.	1.04	0.45*	0.097	0.48*
Scale, background q.	1.05	0.47*	0.098	0.47*
Watch, background q.	1.05	0.50*	0.097	0.51*
Aura, Scale, Watch, background q.	1.02	0.45*	0.098	0.43*
Aura, Scale, Watch	1.16	0.16*	0.110	0.07#

Table 5. Comparing grouping formulations.

		FSS-W		EQ-W	
Framework	Group by	MAE	r	MAE	r
Ensemble	**Data source**	1.00	0.58*	0.097	0.43*
Ensemble	Modality	1.05	0.47*	0.098	0.38*
Concatenation	-	1.06	0.41*	0.100	0.48*

result leads to better predictive performance and generalization. We believe that the importance of background features relates to MS being a highly individual condition which varies greatly from person to person.

Effect of devices. In Table 4 we provide the comparison in performance of our model using data from different device sources. For FSS-W, models considering features from one device are on average 3% worse than a model considering features from all 3 devices, which indicates predictions benefit from having features from different devices. For EQ-W, models considering features from one or all devices perform on a similar level, which is a promising observation as it indicates that features from just one single device can be enough to predict EQ-W, as long as some background features are available too.

We also extend our discussion of the importance of background features in the context of device data. We observe that MAEs increase by an average of 13% when device data is used without background features, which is consistent with our previous sets of results presented in Table 3. Importantly, we find that device data is of greatest value to the modelling of FSS-W and EQ-W when used in conjunction with background data - using device data allows for computation of week-to-week variations resulting in predictions such as those presented in Figure 3, while using background features provides the context information to adjust for differences amongst individual patients.

Ensemble approach vs concatenation. Table 5 shows that an ensemble approach outperforms a simple 'concatenation' framework which considers features from all data sources in a single concatenated vector containing data from all sources (filling missing values with means), which is then trained using an Adaboost

Table 6. Comparing time windows.

	FSS-W		EQ-W	
d	MAE	r	MAE	r
5	1.05	0.50*	0.100	0.38*
7	1.04	0.49*	0.099	0.39*
10	1.00	0.58*	0.097*	0.43*
14	1.01	0.53*	0.099*	0.35*
21	1.00	0.55*	0.098*	0.41*

Table 7. Comparing adaptation strategies.

Adaptation strategy	No. of weeks to adapt	FSS-W		EQ-W	
		MAE	r	MAE	r
-	0	1.00	0.58*	0.097	0.43*
Residual error	1	0.65	0.82*	0.065	0.77*
GMM-MAP	1	0.90	0.58*	0.091*	0.52*
GMM-MAP	6	0.78	0.53*	0.080*	0.53*

Table 8. Results of our model predicting FSS-M and EQ-M.

	FSS-M		EQ-M	
Model	MAE	r	MAE	r
Our model	0.93	0.48*	0.091	0.39*
Baseline	1.03	-	0.100	-

Regressor. Within the ensemble framework, we also explore differences in performance between grouping the data by its source (devices and day questions) versus further dividing into modalities. Diving into modalities here means grouping measurements which are meaningful together or always appear together, for example, grouping the gait and sleep measurements provided by the smart watch as two separate modalities because gait and sleep measurements relate to two very distinct behaviors. We observe that the ensemble model works better with the data source grouping.

Effect of window length. Table 6 shows the effect of varying the length d of the time window considered for computing the time-varying features. We observe that for FSS-W, the model performs better when the time window consists of more than 7 days, while for EQ-W the optimal lies at 10 days. Since both survey instruments ask the patients to reflect upon their health status in the past week, we expect that the time window for extracting readings should be of similar scales, this aligns with our observation that the optimal length is 10 days.

Effect of adaptation. We compare the adaptation strategies of residual error translation against a Maximum A Posterior approach used with Gaussian Mixture Models (GMM-MAP). For GMM-MAP, we adapt the parameters of a model that has been fitted using $\mathcal{D}^{train}$ by considering the first n labelled examples of the test patient.

Table 7 highlights that using the simple residual error adaptation strategy is a pragmatic approach which works effectively in personalizing the model to individual patients. To ensure this adaptation is effective, we also consider the changes in MAE per person before and after applying the translation. Out of 151 tested subjects, 3 and 4 patients experienced a significant increase in MAE (above the instrument SEM) post-adaptation for FSS-W and EQ-W respectively. The reason behind such worsening performance is that these patients' first recorded week is an outlier to the rest of his/her scores, differing by a mean of 44%. It is possible that the adaptation could benefit from a recalibration, or even future consideration of giving prompts to inform the patient that their current week would be used as calibration.

5.4 Other Analysis

In this subsection, we consider another application of our framework: predicting the overall mean FSS and EQ-5D score per person. We present the model performance results and also perform an investigation into the selected features in all considered tasks.

Table 9. Selected measurements per source.

Source	Selected Measurements	FSS-W	FSS-M	EQ-W	EQ-M
Aura	awake duration, night heart/ respiratory rate	✓	✓	✓	✓
	sleep duration, time to wake	✓		✓	✓
	time to sleep	✓	✓	✓	
	light sleep duration	✓	✓		✓
	bed-in time, REM/ deep sleep duration		✓	✓	✓
	no. of times awaking			✓	✓
Scale	weight, standing heart rate, pulse wave velocity	✓	✓	✓	✓
	hydration/ muscle/ fat mass	✓		✓	
Watch	sleep duration, light sleep duration, bed-in times	✓	✓	✓	✓
	step count, walking speed	✓	✓	✓	✓
	deep sleep duration		✓	✓	✓
	time to sleep, awake duration	✓	✓		✓
Fatigue q.	fatigue level score	✓	✓	✓	✓
Sleep q.	sleep quality score	✓	✓	✓	✓
Background q.	age	✓	✓	✓	✓
	'whether the subject has fatigue symptoms'	✓	✓		
	height	✓		✓	
	disability level	✓		✓	✓
	'whether the subject has depression'			✓	
	'whether the subject has pain symptoms', years with MS			✓	✓

Predicting FSS-M and EQ-M. We straightforward apply the previous training pipeline, except now the total number of examples in the datatset is $N = 151$, and we have labels as the mean of all reported FSS-W and EQ-W scores per subject, which we refer to as FSS-M and EQ-M respectively. In this task, the relevant time series is no longer a 10-day window but the entire 6-month time series for each participant. We extract only the days with data entries and aggregate the time-varying measurements as before. Table 8 shows that our models outperform baselines by 8% on average.

The motivation behind conducting this task is as follows: Firstly, this task aligns more closely to the current state of clinical practise where a patient visits a clinician and fills in questionnaires roughly every 6 months; showing that we can predict a score that represents the average response over the entire 6-month period in addition to the weekly variations is beneficial to bringing a more stable view to the patients' level of fatigue symptoms and health status. Moreover, performing this task allows us to compare the selected important features by a model trying to predict a weekly score versus an overall mean score, and draw insights about how a dynamical and stable view differ in relation to the data recorded from our study.

Selected Feature study. Table ?? lists the measurements from each data source which have been selected over the half the time in the leave-one-out setting for predicting the weekly (FSS-W, EQ-W) and overall (FSS-M, EQ-M) scores. These readings are seen by the model as important in predicting FSS and EQ-5D-5L.

Time-varying Features. One of the biggest difference between the features selected by the weekly and mean scores is that more standard deviation features are selected for FSS-M and EQ-M than for FSS-W and EQ-W. This

is expected as the overall view considers a longer sequences and the variations in each measurement should be more pronounced than the weekly case which only considers a time window of 10 days.

Another difference between the selected features by the weekly and overall scores lies in the body composition readings, such as fat and hydration mass. FSS-W and EQ-W frequently select features of these readings but FSS-M and EQ-M never do, this is also reflected in the absence of strong and significant correlation for these items in the correlation matrix. This hints such body composition metrics may be related to fatigue and health status in more dynamic and granular level.

Background Features. We see that for the EQ-W scores, important background features not only relates to the physical information of the patient but also his/her emotional wellbeing (depression); this agrees with expectation as the EQ-5D was designed to measure quality of life, which heavily relates to non-physiological metrics. One interesting observation is that height and 'whether the patient has depression' are not frequently selected features for predicting EQ-M but they are for EQ-W. Height differences may be more important for the weekly model due to the limited information coming from a 10-day time frame. Whether a patient has depression or not may only greatly impact their dynamic week-to-week scores but not an overall mean.

6 DISCUSSION

Our results imply several interesting observations about the significance of different features and data sources depending on the task at hand, as well as the comparative performance of different variations of a machine learning framework to predict FSS and EQ-5D scores from a mixture of wellness devices and self reports. We make the following conclusions useful for future work:

- Having static background features per patient is an indispensable input to the framework, however there is no need for a great number of such features. In general we observe that having 4 to 6 background features are sufficient. Future study could explore asking for more in-depth responses to such background information, e.g. quantifying the patient's disability level using the standard scale of Expanded Disability Status Scale (EDSS) [9].
- The residual-error adaptation strategy using each user's first week as calibration works well and is the most effective to account for patients who reported low-scores as they are a minority in our dataset and also in the MS population. However this strategy may fail if the first week is itself an outlier. It may be beneficial to indicate to the patient that his/her data from the first week of device usage would be used for calibration and to offer options for indicating the need for recalibration when the patient feels that his/her health has changed significantly.
- If the number and usage of data sources varies within a cohort of studied patients, a ensemble method works well to cope with missing data. However, in case of limited resources, data sources or dataset size, it is sensible to first consider a simpler and less comprehensive model which focuses on only a subset of devices. This would have a larger detrimental effect in predicting FSS-W than EQ-W as seen in our results, but it will still be a feasible option.
- Collecting data from self-reports in the form of daily questions (fatigue q. and night q.) does not contribute to a significant difference (>2%) in model performance for the tasks considered. This may be related to the inherent unreliability in self-reports and future frameworks should consider adding this data only if such information is readily available or the collection is of low user burden.

Altogether, our results confirm the feasibility that weekly reported scores of Fatigue and Health State can be accurately and regularly tracked in MS patients using data from connected wellness devices and a small number of background features, with or without additional daily self-reports that patients may choose to provide. One implication is that, now patients may be able to employ such a model to track themselves for long periods of time at weekly intervals, simply by using a general model that requires their calibration once. Given that the

disease does progress we might expect that the patients could re-calibrate every 6 months in order to improve personal adaptations.

To pave way for work in this direction, future studies should seek to address the limitations in data collection. It would also be interesting to have some incorporation of measurements of patients' emotional state, for instance, through use of smartphone data or Ecological Momentary Assessment (EMA) methods. Future work could also collect information relating to environmental effects (e.g. temperature, humidity, which had been suggested in MS literature to be possible symptom triggers). In order to better adapt to personal changes, low-user-burden ways of calibrating the models could be investigated, e.g. asking for a binary indication of whether this week feels more fatigued than last week.

6.1 Anticipated Applications

The predictive modelling performed in our study generates information on MS patient's fatigue and health states under a weekly and stable 6-month view. There is a wide range of potential scenarios where the information generated would be useful, we discuss the motivating scenarios for different stakeholders:

- *Patients: self-monitoring.* The predicted information can supplement patients' current records, so they can develop a regular and holistic view of their conditions. Without the information generated by our framework, currently patients will have an quantitative update to their status to record and share only when they complete a survey, but by using our framework and connected wellness devices on a week-to-week basis they will have a indication as to two key MS indicators, with zero effort on their part and not having to remember to do anything (like answer a series of questions). More this is a much more shareable metric than that would be available if the patient just tries to record a sense of how they feel. It could also help patients reinforce and put into a quantitative fashion their feelings, e.g. rather than feeling bad for not coping, they would be able to indicate a change in an independent scale.
- *Patients: identifying Triggers.* Currently, MS patients keep symptom diaries in order to identify symptom triggers specific to themselves, e.g. exercising on a sunny versus rainy day. The collected by the devices (e.g. walking speed) and the information generated from our framework (e.g. weekly fatigue level) an be organized and visualized through an app platform, together with other forms of data that the patient may give the app access to (e.g. local weather). This is a convenient way for the patient to holistically review different aspects of their conditions and gain understanding of their disease. A potential future direction is to further develop recommendation systems integrated within such platform, so as to suggest possible symptom triggers and encourage experimentation with suitable management strategies. Case in point, most MS patients are physically inactive because they fear the potential fatigue despite evidence that exercise can lessen fatigue in the long term[10]. A possible scenario is patients could use generated fatigue patterns to quantify the benefit of exercise and to identify good times for exercising.
- *Clinicians: evaluating interventions.* Different forms of interventions to improve multiple sclerosis outcomes include medication taking, diet and exercise, however, clinicians currently can only rely on verbal or qualitative feedback from patients to judge the effectiveness of a prescribed treatment. With our framework, patients can choose to share information generated about their conditions when discussing the intervention effectiveness with their clinicians. Since such information will be on a scale of clinically-validated instruments and comparable between different MS patients, clinicians can then judge on a case-by-case basis whether the intervention is effective or not, with additional support from weekly information generated by our model.
- *Researchers: granular view of disease progression.* Many studies have noted the scarcity of records about MS at more fine-grained timescales, and pointed out that sparse observations might lead to overlooking important dynamic fluctuations [20]. MS Researchers can explore the use of connected wellness devices to

obtain a granular view of the condition and analyze the week-to-week information generated about the patients. This could bring potential benefits in a number of areas in MS which has yet to be elucidated, e.g. the disease progression from RRMS to SPMS is poorly understood but this has important implications for treatment as many drugs effective when targeted at RRMS are useless with SPMS [19].

Finally, we anticipate that the information generated can be smoothly integrated into the current consumer and clinical health space while minimizing any additional burden for patients or deviation from clinical practice. Although patients do have to be equipped with smart devices and fill in some background information, but once this is set up and calibrated, the burden of measuring week-to-week fatigue and health state scores is be transferred to the devices and our framework. One strength of our framework is that patients may use any subset of connected wellness devices for predictions to be made. Patients may use connected wellness devices naturally and only fill in daily questions if they would also like to keep a digital record of their symptoms (MS patients are already currently advised to keep symptom diaries in their own ways so this is only a matter of changing recording platforms). In addition, our proposed framework also does not require continuous collection of ground truth labels from the patients (just 1 label is needed for adaptation) so this is placing minimal burden on the patients and will not require clinicians to administer extra symptom questionnaires.

7 CONCLUSION

As a lifelong debilitation disease that affects millions of people worldwide, MS needs to be better understood by researchers and better monitored by patients. Having large-scale longitudinal data of MS patient's condition furthers this goal, and in this study we carried out the first investigation into methods to obtain such data through ubiquitous sensing at ease. In particular, we focused our tasks on data relating to MS patient's fatigue and quality of life, through use of two widely used instruments by the MS community: Fatigue Severity Scale (FSS) and EQ-5D index. We conducted a study to collect behavioral, physiological device data and self-reported data from 198 MS patients, using connected wellness devices over 6 months. In our investigations, we proposed an ensemble regression model which can cope with the data's missing data, as well as adaptation techniques to further improve generalization performance. Our models are able to achieve good prediction performance for the tasks considered, namely predicting a per-participant mean reported score and a per-week per-participant score for each metric. We find that, using data from connected devices, daily reports and background features, with the unadapted model performance are in line with acceptable instrument errors, for FSS (SEM 0.7) we report MAE 1.00 and for EQ-5D (SEM 0.093) we report MAE 0.097. Adapted models improve these results further (FSS: MAE 0.58, EQ-5D: 0.065). These promising results in our dataset show the feasibility in continuous and unobtrusive tracking of fatigue and health state, and potential for future replications into larger-scale replication studies, which has positive implications for supporting MS patient disease management and clinical research.

REFERENCES

[1] 2019. Withings. https://www.withings.com. Accessed: 2019-07-20.

[2] Sourav Bhattacharya, Alberto Gil CP Ramos, Fahim Kawsar, Nicholas D Lane, Lynn M Gionta, Joanne Manidis, Greg Silvesti, and Mathieu Vegreville. 2018. Monitoring Daily Activities of Multiple Sclerosis Patients with Connected Health Devices. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. ACM, 666–669.

[3] E. Fogarty, C. Walsh, R. Adams, C. McGuigan, M. Barry, and N. Tubridy. 2013. Relating health-related Quality of Life to disability progression in multiple sclerosis, using the 5-level EQ-5D. *Mult. Scler.* 19, 9 (2013).

[4] Yoav Freund and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences* 55, 1 (1997), 119–139.

[5] M. M. Goldenberg. 2012. Multiple sclerosis review. *P T* 37, 3 (2012).

[6] M. Heine, I. van de Port, M.B. Rietberg, E.E.H. van Wegen, and G. Kwakkel. 2015. Exercise therapy for fatigue in multiple sclerosis. *Cochrane Database of Systematic Reviews* 9 (2015). https://doi.org/10.1002/14651858.CD009956.pub2

[7] G. Kobelt, J. Berg, P. Lindgren, S. Fredrikson, and B. Jonsson. 2006. Costs and quality of life of patients with multiple sclerosis in Europe. *J. Neurol. Neurosurg. Psychiatry* 77, 8 (2006).
[8] L. B. Krupp, P. K. Coyle, C. Doscher, A. Miller, A. H. Cross, L. Jandorf, J. Halper, B. Johnson, L. Morgante, and R. Grimson. 1995. Fatigue therapy in multiple sclerosis: results of a double-blind, randomized, parallel trial of amantadine, pemoline, and placebo. *Neurology* 45, 11 (1995).
[9] John F Kurtzke. 1983. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). *Neurology* 33, 11 (1983), 1444–1444.
[10] Amy E Latimer-Cheung, Lara A Pilutti, Audrey L Hicks, Kathleen A Martin Ginis, Alyssa M Fenuta, K Ann MacKibbon, and Robert W Motl. 2013. Effects of exercise training on fitness, mobility, fatigue, and health-related quality of life among adults with multiple sclerosis: a systematic review to inform guideline development. *Archives of physical medicine and rehabilitation* 94, 9 (2013), 1800–1828.
[11] Y. C. Learmonth, D. Dlugonski, L. A. Pilutti, B. M. Sandroff, R. Klaren, and R. W. Motl. 2013. Psychometric properties of the Fatigue Severity Scale and the Modified Fatigue Impact Scale. *J. Neurol. Sci.* 331, 1-2 (2013).
[12] P. Newland, J. M. Wagner, A. Salter, F. P. Thomas, M. Skubic, and M. Rantz. 2016. Exploring the feasibility and acceptability of sensor monitoring of gait and falls in the homes of persons with multiple sclerosis. *Gait Posture* 49 (2016).
[13] Nokia. 2018. *Nokia Health Mate app, Your Activity Tracker and LIfe Coach User Guide.* Nokia.
[14] Mari Palta, Han-Yang Chen, Robert M Kaplan, David Feeny, Dasha Cherepanov, and Dennis G Fryback. 2011. Standard error of measurement of 5 health utility indexes across the range of health for use in estimating reliability and responsiveness. *Medical Decision Making* 31, 2 (2011), 260–269.
[15] Tânia Pereira, Carlos Correia, and Joao Cardoso. 2015. Novel methods for pulse wave velocity measurement. *Journal of medical and biological engineering* 35, 5 (2015), 555–565.
[16] M. Rabbi, S. Ali, T. Choudhury, and E. Berke. 2011. Passive and In-situ Assessment of Mental and Physical Well-being using Mobile Sensors. *Proc ACM Int Conf Ubiquitous Comput* 2011 (2011).
[17] R. Rabin and F. de Charro. 2001. EQ-5D: a measure of health status from the EuroQol Group. *Ann. Med.* 33, 5 (2001).
[18] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn. 2000. Speaker verification using adapted Gaussian mixture models. *Digital signal processing* 10, 1-3 (2000), 19–41.
[19] L. A. Rolak. 2003. Multiple sclerosis: it's not the disease you thought it was. *Clin Med Res* 1, 1 (2003).
[20] K. M. Schreurs, D. T. de Ridder, and J. M. Bensing. 2002. Fatigue in multiple sclerosis: reciprocal relationships with physical disabilities and depression. *J Psychosom Res* 53, 3 (2002).
[21] Layal Shammas, Tom Zentek, Birte von Haaren, Stefan Schlesinger, Stefan Hey, and Asarnusch Rashid. 2014. Home-based system for physical activity monitoring in patients with multiple sclerosis (Pilot study). *Biomedical engineering online* 13, 1 (2014), 10.
[22] Hsi G. Sung. 2004. *Gaussian mixture regression and classification.* Ph.D. Dissertation. Rice University.
[23] B. van Hout, M. F. Janssen, Y. S. Feng, T. Kohlmann, J. Busschbach, D. Golicki, A. Lloyd, L. Scalone, P. Kind, and A. S. Pickard. 2012. Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. *Value Health* 15, 5 (2012).
[24] Mandy van Reenen and Bas Janssen. 2015. *EQ-5D-5L User Guide.* EQ-5D.
[25] Petar Veličković, Laurynas Karazija, Nicholas D. Lane, Sourav Bhattacharya, Edgar Liberis, Pietro Liò, Angela Chieh, Otmane Bellahsen, and Matthieu Vegreville. 2018. Cross-modal Recurrent Models for Weight Objective Prediction from Multimodal Time-series Data. In *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth '18).*
[26] Rui Wang, Weichen Wang, Min SH Aung, Dror Ben-Zeev, Rachel Brian, Andrew T Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Emily A Scherer, et al. 2017. Predicting symptom trajectories of schizophrenia using mobile sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 110.
[27] Rui Wang, Weichen Wang, Alex daSilva, Jeremy F Huckins, William M Kelley, Todd F Heatherton, and Andrew T Campbell. 2018. Tracking depression dynamics in college students using mobile phone and wearable sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 43.
[28] Weichen Wang, Gabriella M Harari, Rui Wang, Sandrine R Müller, Shayan Mirjafari, Kizito Masaba, and Andrew T Campbell. 2018. Sensing behavioral change over time: Using within-person variability features from mobile sensing to predict personality traits. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 141.
[29] K. Wynia, B. Middel, J. P. van Dijk, J. H. De Keyser, and S. A. Reijneveld. 2008. The impact of disabilities on quality of life in people with multiple sclerosis. *Mult. Scler.* 14, 7 (2008).