Personalized Longitudinal Assessment of Multiple Sclerosis Using Smartphones

Oliver Y. Chén, *Member, IEEE*, Florian Lipsmeier, Huy Phan, *Member, IEEE*, Frank Dondelinger, Andrew Creagh, Christian Gossens, Michael Lindemann, and Maarten de Vos, *Member, IEEE*

***Abstract*—Personalized longitudinal disease assessment is central to quickly diagnosing, appropriately managing, and optimally adapting the therapeutic strategy of multiple sclerosis (MS). It is also important for identifying idiosyncratic subject-specific disease profiles. Here, we design a novel longitudinal model to map individual disease trajectories in an automated way using smartphone sensor data that may contain missing values. First, we collect digital measurements related to gait and balance, and upper extremity functions using sensor-based assessments administered on a smartphone. Next, we treat missing data via imputation. We then discover potential markers of MS by employing a generalized estimation equation. Subsequently, parameters learned from multiple training datasets are ensembled to form a simple, unified longitudinal predictive model to forecast MS over time in previously unseen people with MS. To mitigate potential underestimation for individuals with severe disease scores, the final model incorporates additional subject-specific fine-tuning using data from the first day. The results show that the proposed model is promising to achieve personalized longitudinal MS assessment; they also suggest that features related to gait and balance as well as upper extremity function, remotely collected from sensor-based assessments, may be useful digital markers for predicting MS over time.**

***Index Terms*—Ensemble learning, digital health technology, generalized estimation equation, longitudinal prediction, missing data imputation, multiple sclerosis, smartphone sensors, subject-specific fine-tuning.**

Manuscript received 20 September 2022; revised 17 February 2023 and 22 March 2023; accepted 20 April 2023. Date of publication 3 May 2023; date of current version 3 July 2023. This work was supported in part by F. Hoffmann-La Roche Ltd., the NIHR Oxford Biomedical Research Centre, and in part by the Research Support Allowance from the University of Bristol. *(Corresponding author: Oliver Y. Chén.)*

Oliver Y. Chén is with the University of Bristol, BS8 2LR Bristol, U.K. (e-mail: olivery.chen@bristol.ac.uk).

Florian Lipsmeier, Christian Gossens, and Michael Lindemann are with the Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd., 4070 Basel, Switzerland (e-mail: florian.lipsmeier@roche.com; christian.gossens@roche.com; michael.lindemann@roche.com).

Huy Phan was with the Queen Mary University of London, E1 4NS London, U.K., and also with Alan Turing Institute, NW1 2DB London, U.K. He is now with Amazon Alexa, Cambridge, MA 02142 USA (e-mail: huypq@amazon.co.uk).

Frank Dondelinger was with the Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd., 4070 Basel, Switzerland. He is now with the Novartis Institutes for BioMedical Research, Novartis AG, 4033 Basel, Switzerland (e-mail: frank.dondelinger@roche.com).

Andrew Creagh is with the Institute of Biomedical Engineering, University of Oxford, OX3 7DQ Oxford, U.K., and also with the Big Data Institute, University of Oxford, OX3 7LF Oxford, U.K. (e-mail: andrew.creagh@eng.ox.ac.uk).

Maarten de Vos is with the Departments of Engineering and Medicine, KU Leuven, 3000 Leuven, Belgium (e-mail: maarten.devos@kuleuven.be).

Digital Object Identifier 10.1109/JBHI.2023.3272117

I. Introduction

MULTIPLE sclerosis (MS) is a chronic autoimmune, inflammatory, and demyelinating disease of the central nervous system [1]. It affects approximately 2.3 million people worldwide [2], [3]. The disease has a pooled incidence rate of 2.1 per 100,000 persons/year across 75 reporting countries, with the mean age of diagnosis being 32 years [4].

Progression of MS commonly leads to accumulation of impairment in one or several functional domains, including upper extremity function, gait and balance, cognition, and vision [5]. Impairment to gait and balance as well as to upper extremity function can affect the quality of life and the ability to perform activities of daily living [6], [7], [8], [9], [10]. Previous reports suggest that up to 75–90% of people with MS (PwMS) experience impaired gait, and up to 60–76% of PwMS show signs of impaired upper extremity function [5], [11], [12], [13].

Regular assessment of functional ability can help to guide early treatment decisions and improve treatment outcomes. At present, individuals at risk of developing MS are examined based on a combination of in-clinic assessments, including checking for MS-related signs and symptoms, neuroimaging studies, and laboratory testing [14]. The Expanded Disability Status Scale (EDSS), rated by clinicians, measures overall disease impairment (i.e., disability in functional systems and ambulation) [15]. The Multiple Sclerosis Impact Scale (MSIS-29), self-assessed by patients, captures the physical and psychological impact of the disease [16], [17], [18]. The EDSS score ranges from 0 to 10 (representing normal to death due to MS). The MSIS-29 total score (physical score plus psychological score) ranges from 25 to 145 (representing best to worse physical and psychological functions).

To frequently assess, appropriately manage, and optimally adapt the therapeutic strategy of MS, it is critical to build suitable longitudinal models. To do so, however, one must confront a few challenges.

First, while regular assessment of functional ability is recommended to support optimal adaptation of the therapeutic strategy [19], clinical measures of functional ability are at present, unfortunately, only infrequently administered. This makes it difficult to align disease measurements timely or accurately with therapeutic strategy recommendations. This difficulty is further

complicated by the fluctuating nature of MS symptoms [19], [20], [21]. Finally, the data collected from such (infrequent) evaluations are not suitable for longitudinal model development. Recently, smartphone sensor-based assessments are beginning to allow remote and frequent assessment at home without supervision. Research has shown that the remote, frequent (daily), and objective assessment of MS-related functional impairment is feasible with smartphone sensor-based assessments [22], [23]. One can typically take these assessments at home without supervision and with minimal burden on the patient [24]. The relatively low cost and high penetration rate of smartphone devices, as compared to the availability of hospital devices and physicians, make such assessments available to a large population [25], [26]. Taken together, more frequent assessments of functional ability using sensor-based smartphone technology are likely to provide notable value in tracking MS-related impairment.

Although recent advancements in sensor technology enable more frequent sensor-based disease assessments, developing a suitable longitudinal model still faces two major challenges. First, longitudinal sensor data often contain missing values. If not treated, one cannot make full use of the data during model development, as any missing values would make other data from that day unusable, discarding potentially useful information. Second, although PwMS exhibit common patterns at the population level, personal disease prediction on new PwMS using results discovered from others may not capture the subject-specific information of a new PwMS.

A beginning to address these challenges can perhaps be made by considering a longitudinal method that balances extrapolation and personalization. Here, we propose a personalized longitudinal framework to automatically assess MS over time. Using smartphone sensor data from a feasibility study [22], we demonstrate that the framework has the potential to (1) extract MS-specific digital clinical markers and (2) assess individual MS trajectories longitudinally.

II. Methods

The structure of this section is organized as follows. In Section II-A, we introduce the data used in this paper. In Section II-B, we discuss how we deal with missing data in longitudinal studies. In Sections II-C and II-D, we introduce the model and how to make inferences about the parameters. Section II-E presents model ensembling to make out-of-sample predictions. In Section II-F, we introduce personalized fine-tuning to improve longitudinal disease assessment. Finally, in Section II-G, we investigate the optimal number of imputed datasets. The data organization, imputation, model development, and parameter ensemble are summarized in Fig. 1.

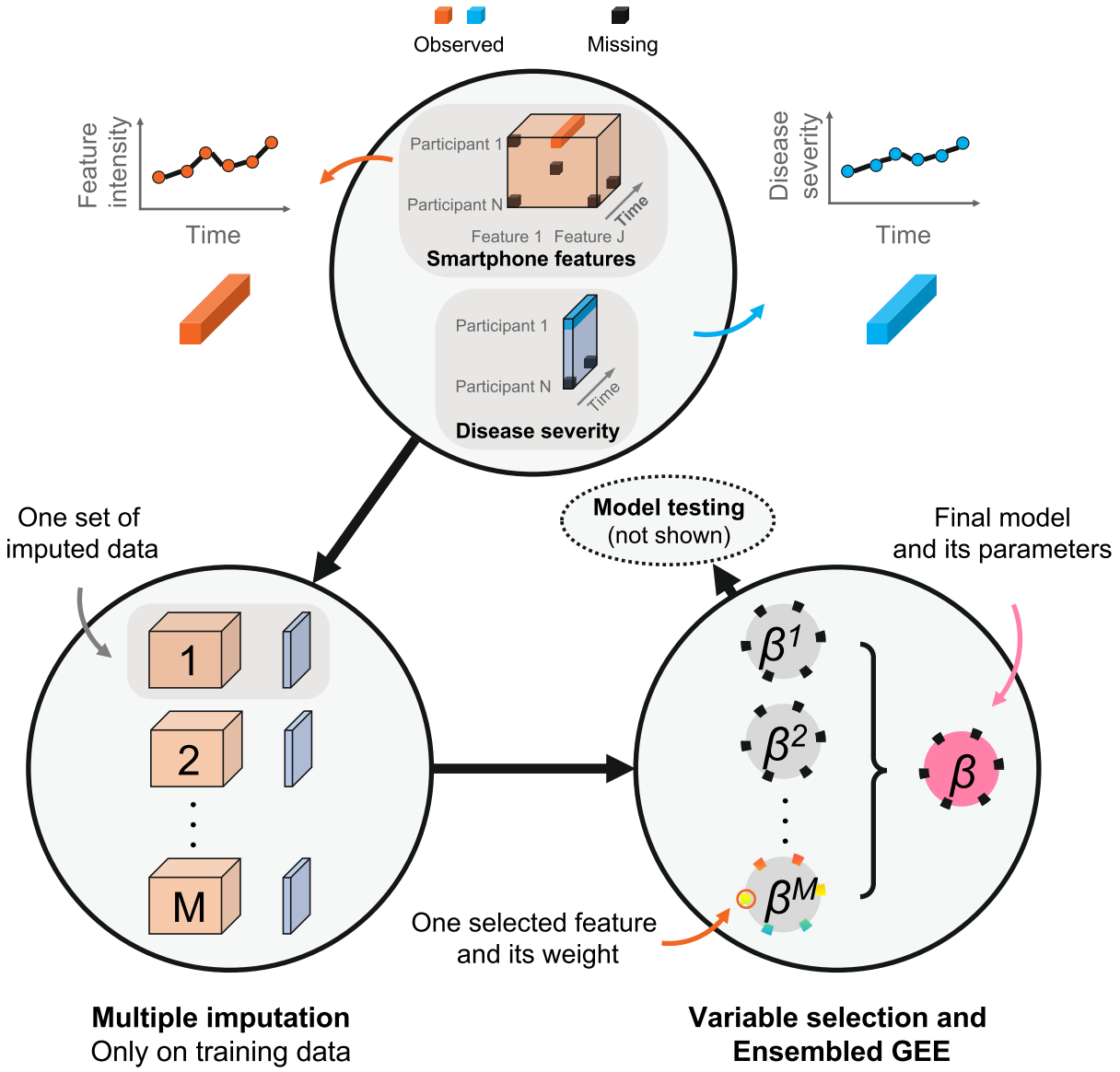
A. Patient Data Collected Remotely at Home by Smartphones

We used data from a 24-week, prospective study (clinicaltrials.gov identifier: NCT02952911) aimed to assess the feasibility of remotely monitoring PwMS with sensor-based assessments [22], [24]. Smartphone data were collected using a preconfigured smartphone (Samsung Galaxy S7) that prompted the participants to perform daily assessments of upper extremity function, gait and balance, and cognition [22], [24]. The local ethics committees approved the study. Written informed consent was obtained from all participants; see [22] for the exclusion criteria for PwMS and other study details.

Walking ability and upper extremity function are indispensable to support independence in activities of daily living, and their functions are often impaired in PwMS [27], [28]. Previous studies have shown that digital features related to walking ability and upper extremity functions are useful to assess PwMS [29], [30], [31]. Identifying and isolating the subset(s) of features that capture impairment in these functional domains and that are potentially predictive of MS disease severity, therefore, may help to improve one's understanding of the pathological aspects of MS that are related to gait and balance as well as to upper extremity function. Additionally, the selected features may inform clinical decisions that aim at preserving the functional abilities in these two domains. In parallel, the magnitude of impairment in these features may be associated with the severity of MS. For example, worsening walking ability is associated with MS disease progression and/or one or more relapses [32]. Similarly, worsening upper extremity function is frequently reported in PwMS and is increasingly present as the disease progresses [33]. Together, these lines of evidence suggest that gait and upper extremity function features may be useful markers to longitudinally assess the severity and progression of the disease.

The data in this study are objective measurements from a Two-Minute Walk Test (2MWT), assessing gait and dynamic balance, and a Draw a Shape (DaS) Test, appraising upper extremity function. Specifically, the 2MWT data were recorded using smartphones' accelerometers and the DaS data were recorded using smartphones' touchscreen sensors [22]. The DaS Test assesses fine finger or manual dexterity while the participants are instructed to hold the smartphone in the untested hand and draw on the smartphone touchscreen six pre-written alternating shapes of increasing complexity (linear, rectangular, circular, sinusoidal, and spiral) with the second finger of the tested hand as fast and as accurately as possible within a maximum time (30 seconds for each of the two attempts per shape) [24]. The 2MWT can be performed on even ground where participants could walk straight for as far as ≥ 200 meters without U-turns. The participants are instructed to walk as fast and as long as they can, but safely, for two minutes; they are allowed to wear regular footwear and, as needed, an assistive device and/or orthotic [24].

We consider two gait features from the 2MWT and 51 hand function features from the DaS test. The smartphone data were collected for 24 weeks. During this period, most participants performed five to seven active tests per week; even in the last week of the study, participants completed all active tests four days per week on average [24]. The study also included the clinician-rated Expanded Disability Status Scale (EDSS) scores and the self-reported MS Impact Scale (29-item scale) (MSIS-29) scores. The longitudinal MSIS-29 study was conducted on 80 PwMS and the longitudinal EDSS predictions were conducted on 65 PwMS.



Top: Data organization. The smartphone data are organized in an orange cubic, where multi-dimensional feature data are collected from each participant over time. Specifically, the orange bar indicates one feature measured over time for one subject. Disease outcomes are similarly prepared, where the blue bar shows the disease scores measured over time. The black boxes refer to missing data. Lower left: Data imputation. Missing data are imputed via multiple imputation (MI). To accommodate the uncertainties in the imputed data, we compute multiple (M) sets of complete data. Lower right: Model development. A generalized estimation equation is fit on each of the fully imputed datasets to obtain parameter estimation. The M sets of estimated parameters are then pooled to generate a unified model, which will be used for the model test. The (pooled) parameters are applied to features from the test sample to estimate outcomes. To prevent data imputation from influencing the test results, only complete observations will be used during the test (namely, MI will not be used during the test). The estimated results are then compared with the observed outcomes to verify the performance as well as reproducibility of the model.

Fig. 1. A schematic representation of the longitudinal model.

The number of self-reported MSIS-29 scores ranges from 1 (two PwMS) to 23 for each individual with a mean of 8.5. The frequency of the MSIS-29 scores is every two weeks [22], [24]. The number of EDSS scores is 1 (24 PwMS), 2 (30 PwMS), or 3 (11 PwMS) for each individual, with a mean of 1.8. For PwMS with three EDSS scores, they were measured by clinicians at scheduled clinic visits (baseline, Week 12, and Week 24) [22], [24]. The testing frequency for both 2MWT and DaS tests is daily [22]. To better match the smartphone data with the disease scores, we sampled the smartphone data every fortnight; the

resulting smartphone features matched the frequency of the MSIS-29 scores, and it of the EDSS scores at baseline, Week 12, and/or Week 24.

Although the EDSS scores are clinically more useful than the self-reported MSIS-29 scores, for the present study, we mainly focused on demonstrating the model's longitudinal perspective using MSIS-29 scores. This is because MSIS-29 scores were collected every two weeks, and, as such, there were more MSIS-29 scores available for longitudinal model development. In parallel, by painting a finer longitudinal disease curve every fortnight using the predicted MSIS-29 scores, one can better evaluate the proposed longitudinal model. To complement this, we also evaluated the model for predicting EDSS scores. Although there were only a small number of EDSS scores per subject for training and test, our analyses suggested that the proposed model seems promising in predicting - despite sparse - EDSS scores.

B. Treating Missing Data in Longitudinal Studies

Missing data are common in data collection and pose challenges for longitudinal disease prediction [34], [35]. When there are missing inputs or outcomes in the training data, model development becomes difficult (if including missing values) or inefficient (if excluding days with missing values). A practical longitudinal disease predictive model, therefore, needs to first treat the problem of missing data.

In this paper, we used the multiple imputation (MI) [36] to handle missing data. We adopted it primarily because it was relatively easy to control for model efficiency (see (4)) by adjusting the number of imputations performed [37], [38] (see below). Additionally, it provided a principled way to estimate the uncertainty associated with the imputation [39], [40]. Finally, for practitioners, MI may be easier to implement as it is available in standard software. One can refer to [41], [42], [43], [44] for additional treatments of missing data. Note that while the use of MI for randomized clinical trials warrants some caution [45], the design of this study did not include any randomization into treatment groups, and hence the application of MI was straightforward; we, however, do assume that data are missing at random (MAR).

Here, we briefly outline the steps of MI. First, a column that contains missing values is regarded as a target column (and treated as a response variable), and the remaining columns of the data set are used as predictors. The missing values in the targeted column are then filled using the predicted mean matching (PMM) [44], [46]. For predictors that are incomplete in themselves, the most recently generated values are used so that the predictors are complete before making imputations for the target column. Second, the first step is repeated for every column that contains missing data. Finally, the first two steps are repeated multiple (M) times to obtain M sets of complete data. These steps can be done using the *R* package *mice* [47].

C. Model Development Using the GEE

The longitudinal predictive model ensembles multiple generalized estimation equations (GEEs) followed by a subjective-specific fine-tuning (see Fig. 1). The choice of choosing the GEE is twofold. First, it can take both correlated and uncorrelated repeated longitudinal measurements (both within and between PwMS). Second, even if one mis-specified the correlation structure, the parameter estimates would still be consistent [48]. The subject-specific fine-tuning is added to treat potential overfitting during training and underestimating for new PwMS with severe scores during the test.

Before introducing the subject-specific aspect of the model (see Section II-F), let's first outline the general longitudinal predictive model. Consider N PwMS, where the i^{th} subject has p_w gait features and p_d hand function features, measured at time $t = (1, 2, \ldots, T_i)$ (the subscripts w and d indicate walking (gait) and dexterity (hand) functions, respectively). The feature matrix for the i^{th} subject is thus $\boldsymbol{x}_i^{\mathsf{T}} = (\boldsymbol{x}_{iw}^{\mathsf{T}}, \boldsymbol{x}_{id}^{\mathsf{T}})$, $\boldsymbol{x}_{ij}^{\mathsf{T}} = (\boldsymbol{x}_{ij1}, \ldots, \boldsymbol{x}_{ijp_j})$, for $j \in \{w, d\}$, and $\boldsymbol{x}_{ijk}$ contains longitudinal measurements $\boldsymbol{x}_{ijk} = (x_{ijk}(1), \ldots, x_{ijk}(T_i))^{\mathsf{T}}$, for $1 \leq k \leq p_j$, where T denotes the transpose operation. Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT_i})^{\mathsf{T}}$ be the longitudinal outcome for the i^{th} subject and y_{it} be the outcome for the i^{th} subject at time t.

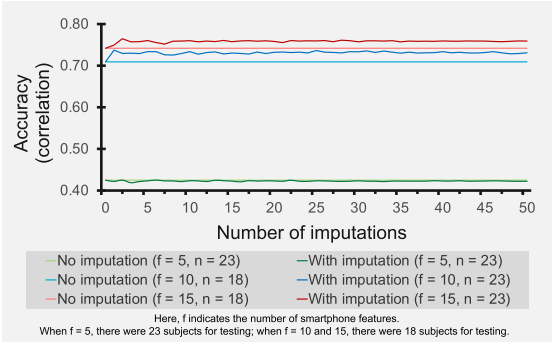
Consider $p_w = 2$ gait features from a Two-Minute Walk (TMW) test, and $p_d = 51$ hand function features from a Draw a Shape (DaS) test. The model also includes covariates such as sex, age, race, and site (from where data were collected). In the following, unless explicitly stated, features from 57 randomly selected PwMS were used for developing the MSIS-29 model, which was then evaluated on 23 new PwMS. Features from 45 randomly selected PwMS were used for developing the EDSS model, which was then evaluated on 20 new PwMS. To avoid a chance split during training and testing, we will perform bootstrap experiments in Section III-A.

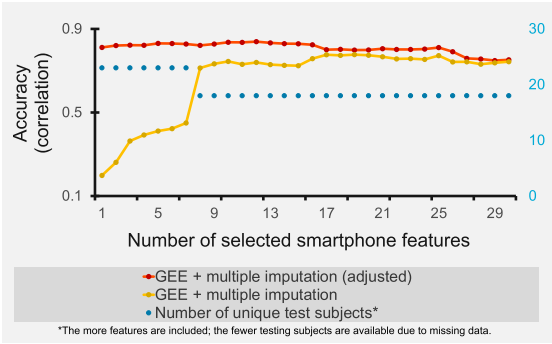
We considered feature analysis in two steps. The first step selected features significantly associated with the outcome (see [49]). The second step performed parameter estimation using the GEE for selected features (see Fig. 4).

During feature selection, we control the total number of selected features to discover a small number of top digital features.[1] For demonstration purposes, six features (the best four smartphone features plus age and gender) were selected. As a complementary analysis, we explored different numbers of features in the final model (see Fig. 2). Our results suggested that, overall, the model performance was consistent across different numbers of features.

More specifically, given a fixed number of test subjects, the model performance without fine-tuning (*i.e.*, GEE plus multiple imputation) seemed to improve as more features were selected. Further added features saw stagnated improvement. The model

[1]Practically, one could select features based on the p-values (*e.g.*, after a pair-wise correlation test, the features with p-values less than 0.05 are retained), the distance metric between the features and the outcome (*e.g.*, the features whose correlations with the outcome that are greater than 0.7 are retained), or, relatedly, constrain the number of selected features (*e.g.*, by retaining several top features most correlated with the outcome). Specifically, consider a feature $\boldsymbol{x}$. Denote $\rho_{(\boldsymbol{x},\boldsymbol{y})}$ as the result of a statistical test between the feature $\boldsymbol{x}$ and the outcome $\boldsymbol{y}$. For disease severity outcomes, the value of ρ can be the correlation between the feature and the outcome; it can also be the corresponding p-values. Let ϵ be a pre-specified threshold for ρ. One can set ϵ to control for the p-values (from a correlation test), a distance metric (if ρ is the correlation coefficient), or the total number of selected features.





Top: Model performance using different numbers of imputations. Model performance improved with imputations compared to without. Bottom: Model performances for the mean approach, raw generalized estimation equations (GEE) approach, and adjusted GEE approach across different numbers of selected features. Due to missing values in features, the size of the available independent test sample (indicated by the height of the blue dots) decreased when more features were used. This is because we only considered observed features for testing (namely we did not perform MI for the test sample to prevent it from potentially biasing prediction performance). In general, the prediction accuracy using MI and GEE (indicated by the yellow line) was relatively high and stable across various numbers of features selected. The fine-tuning further improved prediction performance (see the orange line).

Fig. 2. Longitudinal MSIS-29 prediction using different sets of imputations and various numbers of features.

performance with fine-tuning was relatively stable using small numbers of top features; the performance slightly worsened when the number of features further increased - this is likely because the longitudinal association between the further added features and the disease outcome was not strong. Although we treat missing data in the training set via imputation, to avoid bringing bias into the test data by imputation, we did not perform MI on the test set and evaluated the model performance only on available observations. Naturally, if one included a larger number of features, there would be fewer numbers of PwMS left with complete features. Formally, the (general) marginal model that studies the relationship between the features and outcomes is:

$$\mathbb{E}\left(y_{it}|\boldsymbol{x}_{it},\boldsymbol{z}_i\right)=g^{-1}(\mu+\boldsymbol{x}_{it}^{\mathsf{T}}\boldsymbol{S}\boldsymbol{\alpha}+\boldsymbol{z}_i^{\mathsf{T}}\boldsymbol{\gamma}) \tag{1}$$

where $g(\cdot)$ is a link function, μ is the intercept, $\boldsymbol{x}_{it}^{\mathsf{T}}=[\boldsymbol{x}_{iwt}^{\mathsf{T}},\boldsymbol{x}_{idt}^{\mathsf{T}}]$ and $\boldsymbol{x}_{ijt}^{\mathsf{T}}=(x_{ij1}(t),x_{ij2}(t),\ldots,x_{ijp_j}(t))$, for $j\in\{w,d\}$. Here, $\boldsymbol{S}=\text{blockdiag}\{\boldsymbol{I}_w,\boldsymbol{I}_d\}$, and $\boldsymbol{I}_j=\text{diag}\{i_{j1},i_{j2},\ldots,i_{jp_j}\}$, wherein $i_{jk}=1$ if $\rho_{(\boldsymbol{x}_{jk},\boldsymbol{y},N-2)}<\epsilon$ and 0 otherwise, where $\boldsymbol{x}_{jk}=(\boldsymbol{x}_{1jk}^{\mathsf{T}},\boldsymbol{x}_{2jk}^{\mathsf{T}},\ldots,\boldsymbol{x}_{Njk}^{\mathsf{T}})^{\mathsf{T}}$ denotes a particular feature across all PwMS over time; $\boldsymbol{z}_i$ is a vector containing all covariates for the i^{th} subject, and $\boldsymbol{\gamma}$ is its coefficient. Finally, write $\boldsymbol{\beta}=(\boldsymbol{\alpha},\boldsymbol{\gamma})$.

Denote the selected features as $\tilde{\boldsymbol{X}}=\boldsymbol{X}\boldsymbol{S}$ and the outcome disease scores as $\boldsymbol{Y}$. Via MI, we obtained M sets of complete data $(\boldsymbol{Y}^{(1)},\tilde{\boldsymbol{X}}^{(1)},\boldsymbol{Z}),\ldots,(\boldsymbol{Y}^{(M)},\tilde{\boldsymbol{X}}^{(M)},\boldsymbol{Z})$, where the superscript (m) denotes the m^{th} set of imputed data.

The problem in (1) can be solved using the GEE. Specifically, the estimation of $\boldsymbol{\beta}$ can be done by solving the following score function:

$$\boldsymbol{D}_i^{\mathsf{T}}\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu}_i)=0$$

where $\boldsymbol{D}_i=\dfrac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}^{\mathsf{T}}}$, $\boldsymbol{\mu}_i$ is the expectation of $\boldsymbol{y}_i$, and $\boldsymbol{V}_i$ is the estimate of the variance-covariance matrix of $\boldsymbol{y}_i$. Specifically,

$$\boldsymbol{V}_i=\Delta_i^{1/2}\mathcal{R}_i(\alpha)\Delta_i^{1/2}$$

where $\Delta_i=\text{diag}\{\text{var}(y_{i1}),\text{var}(y_{i2}),\ldots,\text{var}(y_{iT_i})\}$, $\mathcal{R}_i(\alpha)$ is a $T_i\times T_i$ "working" correlation matrix [48], and α represents a vector of parameters associated with a specific model for the correlation of $\boldsymbol{y}_i$. When $\mathcal{R}_i(\alpha)$ is the true correlation matrix for $\boldsymbol{y}_i$, $\boldsymbol{V}_i=\text{cov}(\boldsymbol{y}_i)$. When the variance-covariance structure is mis-specified, the parameter estimates would still be consistent [48].

D. Inference

A GEE model was fit on selected features from each imputed data set. Let $\hat{\boldsymbol{\beta}}^m=(\hat{\beta}_1^m,\hat{\beta}_2^m,\ldots,\hat{\beta}_p^m)$ denote the estimated parameter corresponding to the m^{th} set of imputed data, where p (*e.g.*, $p=6$) is the number of selected smartphone features (*e.g.*, $q=4$) plus the number of covariates ($p-q=2$), and $1\le m\le M$. Then, the $(1-\alpha)\times 100\%$ confidence interval (CI) for the estimated $\hat{\beta}_j$ corresponding to the j^{th} feature, where $1\le j\le p$, is:

$$\left[\bar{\hat{\beta}}_j-t_{\left(n-1,\frac{\alpha}{2}\right)}\frac{s_j}{\sqrt{M}},\quad \bar{\hat{\beta}}_j+t_{\left(n-1,\frac{\alpha}{2}\right)}\frac{s_j}{\sqrt{M}}\right]$$

where $\bar{\hat{\beta}}_j=\sum_{m=1}^M\hat{\beta}_j^m/M$ and $s_j=\sqrt{\sum_{m=1}^M(\hat{\beta}_j^m-\bar{\hat{\beta}}_j)^2/M}$.

E. Model Ensemble and Out-of-Sample Prediction

We obtained the ensemble parameter $\bar{\boldsymbol{\beta}}$ from M GEE models:

$$\bar{\boldsymbol{\beta}}=\frac{1}{M}\sum_{m=1}^M\hat{\boldsymbol{\beta}}^m.$$

TABLE I
MODEL COMPARISON

	Mean prediction						Longitudinal prediction					
	EDSS			MSIS-29			EDSS			MSIS-29		
	Mixed effect	GLM	GEE	Mixed effect	GLM	GEE	Mixed effect	GLM	GEE	Mixed effect	GLM	GEE
r	0.353 (0.352)	0.293 (0.308)	0.319 (0.336)	0.500 (0.502)	0.644 (0.643)	0.682 (0.676)	0.232 (0.234)	0.412 (0.426)	0.482 (0.497)	0.627 (0.627)	0.655 (0.656)	0.691 (0.692)
r (adjusted)	*	*	*	0.837 (0.819)	0.878 (0.878)	0.880 (0.879)	*	*	*	0.792 (0.772)	0.800 (0.799)	0.805 (0.805)
MSE	1.070 (1.296)	1.203 (1.196)	1.194 (1.188)	25.670 (27.888)	22.610 (22.755)	22.334 (22.609)	1.167 (1.439)	1.155 (1.147)	1.102 (1.098)	21.550 (23.466)	19.740 (19.841)	19.503 (19.677)
MSE (adjusted)	*	*	*	15.610 (16.692)	13.221 (13.283)	13.048 (13.177)	*	*	*	14.847 (15.633)	14.888 (14.858)	14.694 (14.638)

*Many people with MS (PwMS) only have one to two EDSS scores. It is, therefore, impractical to make further adjustments using data from day one.

Left: Mean outcome prediction. The GEE model is compared with two baseline models, a mixed-effect model and a GLM, to predict both averaged EDSS and MSIS-29 scores. The prediction accuracy is evaluated using the correlation (r) and the mean squared errors (MSE). Right: Longitudinal outcome prediction. The same analysis is performed for longitudinal disease prediction, where the longitudinal disease outcomes are individual EDSS and MSIS-29 scores. Here, black-coloured values and grey-coloured values (within the parentheses) were results from ensemble GEE models using two averaging approaches (averaging sets of parameters *vs.* averaging sets of predictions). More specifically, the former obtained M imputed data sets, each (via a GEE) produced M sets of parameters; it then averaged the M sets of parameters to produce one unified set of parameters before verifying them on the test set to produce one set of predicted outcomes. The latter first fitted M sets of parameters on the test data, obtain M sets of predictions, and then average the M sets of predictions to obtain one set of final predictions. Our analyses suggest that these two averaging approaches yielded very similar prediction results.

Afterwards, longitudinal prediction on disease outcomes for a new subject k at time t in the test set is made using:

$$\hat{y}_{kt} = g^{-1}\left(\hat{\mu} + \tilde{\boldsymbol{X}}_{kt}\bar{\boldsymbol{\alpha}} + \boldsymbol{z}_k\bar{\boldsymbol{\gamma}}\right) \tag{2}$$

where $\tilde{\boldsymbol{X}}_{kt} = \boldsymbol{X}_{kt}\hat{\boldsymbol{S}}$ represents the selected features from subject k at time t and the selection depends only on the training set (as $\hat{\boldsymbol{S}}$ is estimated from the training data); $\bar{\boldsymbol{\alpha}}$ is of the same dimension with the original number of smartphone features with q non-zero entries in locations corresponding to the q selected features in $\bar{\boldsymbol{\beta}}$ and zeros in the remaining entries; similarly, $\bar{\boldsymbol{\gamma}}$ is of the same dimension of total covariates with $p-q$ non-zero entries corresponding to the $p-q$ selected covariates in $\bar{\boldsymbol{\beta}}$ and zeros in the remaining entries.

Besides averaging M sets of parameters to yield a unified predictive model, one could alternatively average multiple sets of predictions from each of the M models. These two types of averaging techniques, however, yielded very similar outcomes in this study (see Table I), and we proceed with the predictive model described in (2).

The data organization, imputation, model development, and parameter ensemble are summarized in Fig. 1. As a remark, we did not perform MI for the test data to ensure that (1) the evaluation of model performance was only done on the observed (test) data; and (2) the ability of the model to conduct personalized disease assessment on new PwMS was independent of the imputation mechanism.

F. Personalized Longitudinal Disease Assessment and Its Improvement

Longitudinal disease assessment needs to address two important problems: extrapolation and personalization. For extrapolation, it needs to identify markers present in a broad patient population so that they can be extended to new samples. For personalization, it needs to account for idiosyncratic information that is unique to each patient and is, therefore, potentially not encoded in the population-level parameters.

In methods development, these two sets of problems translate to the following tasks.

Task 1: Can we discover features whose patterns are associated with disease outcomes longitudinally whereby such association is generally present in all PwMS in the training sample? If so, they may be useful to make forecasts for new individuals (see Fig. 3).

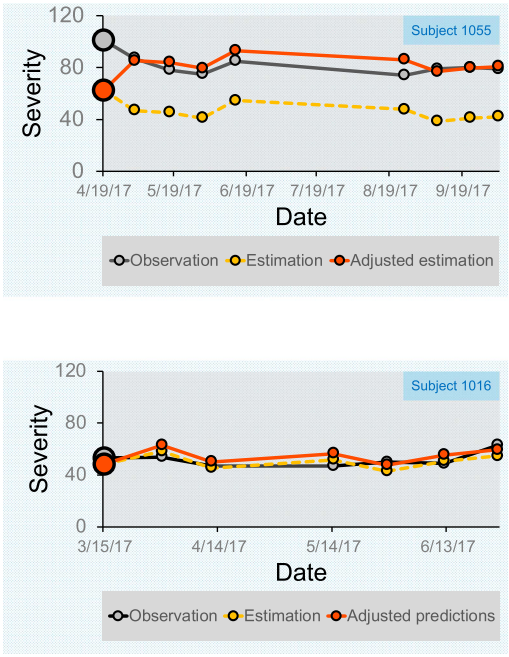
Task 2: Can we further adjust the longitudinal predictions to reflect individual differences? Said differently, the trained parameters from the GEE model reflect the relationship between the features and the outcomes specific to the training sample. If we were to use these parameters to make forecasts about outcomes in new PwMS, they may not fully account for the relationship between the features and outcomes in new individuals. For example, a population-level model in practice may underestimate disease severity in new PwMS during out-of-sample testing (see Figs. 3 and 5). On the other hand, if one considers subject-specific parameters (such as that in the mixed effect models) or incorporates prior information about a subject (such as that in Bayesian learning) for the training model, such parameters or priors still only reflected individual information with regards to the training sample. Naturally, one would want to make an individualized adjustment to the population-level model (whose parameters were developed on the training sample) to incorporate subject-specific information that is contained in the test sample. But constructing an (independent) subject-specific model for the test sample to include idiosyncratic information may over-fit the test data.

To balance extrapolation and personalization, we added the following subject-level fine-tuning into the longitudinal model:

$$\hat{y}_{kt}^{\text{adj}} = \begin{cases} \hat{y}_{kt} & t = 1 \\ \hat{y}_{kt} + (y_{k1} - \hat{y}_{k1}) & t \geq 2 \end{cases} \tag{3}$$

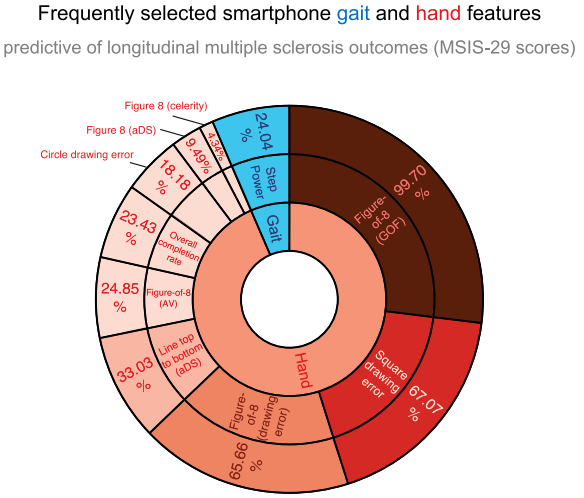
where $\hat{y}_{kt}$ is the estimated outcome for a new subject k at time t using (2), y_{k1} is the earliest ($t = 1$) observed outcome for subject k and $\hat{y}_{k1}$ is its estimation; $\hat{y}_{kt}^{\text{adj}}$ is the fine-tuned estimate.

The above correction means, after making disease estimations using the population-level parameters from the ensemble GEE



Two individuals, one with severe MSIS-29 scores (top panel) and another with minor scores (bottom panel), are shown. The grey circles represent observed longitudinal scores. The yellow circles are estimated scores using the ensemble generalized estimation equations (GEE) parameters. The orange circles are estimated scores further fine-tuned using data from the first day (see text for details).

Fig. 3. Personalized longitudinal disease assessment using smartphones.



The distribution of selected gait and hand features. One thousand bootstrap experiments were performed to discover gait and hand features consistently selected for longitudinal MSIS-29 prediction. The selected gait and hand features were plotted on a pie chart: from inner circle to outer circle were feature category (gait *vs.* hand features), feature name, and the frequency each feature was selected from one thousand bootstrap experiments. Gait and hand features were coded in blue and red, respectively (see text for details on individual features). Note that, as six features were allowed during each bootstrap, the selected (six) features may vary across different bootstraps. But noticeably some features were more frequently selected.

Fig. 4. Key features in smartphone-based remote longitudinal outcome prediction for multiple sclerosis patients.

on new PwMS (dashed lines in Fig. 3), we make a subject-level correction (by adjusting a constant $(y_{k1} - \hat{y}_{k1})$ on estimates made from day two and beyond using disease information from day one (solid orange lines in Fig. 3).

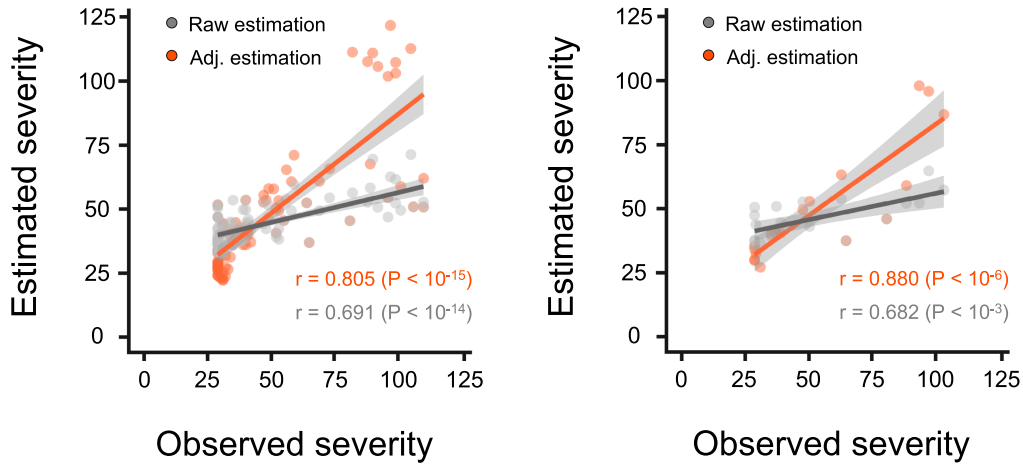
There are a few reasons for using one day's data for fine-tuning (instead of using multiple days) in this study. First, when there are only a small number of data points recorded, using data from multiple days for tuning would result in fewer days of data left for testing (and could potentially artificially boost test results). Second, our analyses demonstrated that fine-tuning using data from one day indeed led to improved personalized disease monitoring while avoiding over-fitting the test data. The reason that the fine-tuning avoided over-fitting was twofold. (a) It did not use any information from the outcomes obtained after day one. (b) It only adjusted a constant $(y_{k1} - \hat{y}_{k1})$ on the prediction. Had the estimated longitudinal trend been inaccurate using (2) (see the yellow line in Fig. 3), a constant adjustment in (3) would not improve the trend prediction. Finally, when longer test data become available for each individual in the future, one can potentially further improve the personalization by using a fine(r)-tuning based on more data or by extracting a subject-level prior for each new subject using data obtained during an initial period. But our message of considering and balancing extrapolation and fine-tuning from the current proof-of-concept study remains.

G. Optimal Number of Imputed Datasets

When employing MI, an interesting question arises. How many sets of imputed data are needed? Although the percentage of missing data in our MS data set is relatively small ($\gamma = 1.42\%$), missing just one entry of an important feature on a particular day would make other features collected on that day not used during model development.[2] More concretely, this question needs to be addressed in two aspects. First, how many sets of imputations would yield high prediction accuracy – this needs to be evaluated empirically. Second, theoretically, how can we determine that the so-called "high prediction accuracy" is sufficiently high?

To empirically examine the optimal number of sets of imputations, we considered different choices of M and calculated their

[2]Once the model has been developed (*i.e.*, after the parameters have been trained), having missing data during an out-of-sample test is less concerning. As smartphone data are relatively easy to collect, if there were missing data on one day, one can repeat the tests (*e.g.*, DaS and 2MWT) the next day (or next week) as the disease is not likely to advance in a few days.



Left: The estimated against observed MSIS-29 scores for previously unseen people with MS (PwMS). The grey dots correspond to the (raw) estimations using ensemble GEE in **Eq.** (2). The orange dots are the fine-tuned estimations using **Eq.** (3). Right: To examine whether the significant correlations observed in the left figure are due to repeated measurements per subject, the individual mean estimations are plotted against individual mean observations. The orange and grey lines represent the linear association between estimated and observed values.

Fig. 5. Estimating multiple sclerosis (MS) severity over time in new individuals.

out-of-sample prediction accuracies (the correlation between true and estimated disease severity scores) (see Fig. 2). Our results showed that the prediction accuracy with MI was generally higher than it with missing values. Additionally, once more than 15 imputations were made, the accuracy became relatively stable. Next, the number of imputations and efficiency have the following approximate relationship [39]:

$$\text{Efficiency} = \left(1 + \frac{\gamma}{M}\right)^{-1} \tag{4}$$

where γ refers to the percentage of missing data, and M denotes the number of imputations. When setting $M = 15$, the efficiency was 99.9% by (4). As adding more imputations had not significantly improved prediction accuracy or efficiency and would incur unnecessary computing time, we chose $M = 15$ for the remaining of the paper.

III. Results

In this section, we present longitudinal MS prediction results using smartphone data. Throughout, to deal with uncertainty resulting from imputation, we generated 15 imputed values for each missing value. That is, for each experiment, 15 sets of complete data containing imputed values were used to train the model. Fifteen GEE models were subsequently fit for each of the imputed, and now full, training data sets. The resulting 15 sets of estimated parameters were pooled to form a unified, ensemble GEE predictive model. The final predictive model was then used to forecast disease outcomes in independent test PwMS. To ensure that the out-of-sample prediction was not affected by imputation mechanisms, we only considered imputation for the training data and evaluated the model performance on independent PwMS only on days where disease scores were available.

To demonstrate the efficacy of the proposed framework, we compared it with mixed-effect and GLM models in the same modelling and test strategy. For each model, we applied it to predict both the averaged outcomes and longitudinal outcomes, and for both EDSS and MSIS-29 scores. We recorded the accuracy statistics (correlation and MSE) from each model (see Table I). Our results showed that for both disease scores, GEE outperformed the mixed-effect model and was slightly better than GLM in mean assessment and longitudinal assessment. Since there were far fewer longitudinal measurements of EDSS score per individual compared to those of MSIS-29 score and the main theme herein is on longitudinal disease severity prediction, in the following we will focus on examining the performance of the proposed model with regards to longitudinal MSIS-29 score prediction.

Our results suggest features related to gait and upper extremity functions are useful to assess MS longitudinally (see Fig. 4). Specifically, the following DaS features [29] are consistently selected. "Figure-of-8 GOF" is the root-mean-squared error obtained from a linear regression between the reference trace of the figure-of-8 and the drawn trace (*i.e.*, the goodness of fit [GOF]). "Figure-of-8 AV" is the angular drawing velocity (AV). "Figure-of-8 trace celerity" is the ratio of trace accuracy over drawing time. "Figure-of-8 drawing error", "Circle drawing error", and "Square drawing error" quantify the error made when drawing a figure-of-8, circle, and square, respectively. "Figure-of-8 aDS" and "Line top to bottom aDS" quantify the absolute drawing speed (aDS) when drawing the figure-of-8 and line top to bottom, respectively. "Overall completion rate" describes the proportion of successfully drawn shapes. From the 2MWT [30], "Step power" is consistently selected. It quantifies the power, or energy, invested per step during a 2MWT.

Overall, the estimated longitudinal MSIS-29 scores in the new test sample were significantly correlated with their observed counterparts ($r = 0.81$, $p < 0.001$) (see left of Fig. 5). It, however, remained possible that the significant association was boosted by having repeated measurements in the test set. To check for this possibility, we took the means of the predicted and observed outcomes for each individual (thus there was only one estimated mean score and one predicted mean score for each participant) and calculated the association again; the result remained significant ($r = 0.88, p < 0.001$).

A. Bootstrap Experiments on Evaluating Predictions At the Group Level

Since training and test samples were obtained from a random split of the data, predictions made on some splits may perform better than those made on others. To evaluate the general performance of our model, we performed 1,000 bootstrap experiments. Specifically, each bootstrap began with a random draw of the data. The PwMS in the sampled data were subsequently split into 70% (training) and 30% (test), where repeated measurements of a subject were either in the training set or the test set. Next, for each bootstrap sample, multiple imputation ($M = 15$) was made on its training set, followed by model development using the GEE on each of the imputed full set. Subsequently, parameters from the M models were pooled to form an ensemble GEE model that was specific to each bootstrap training sample, with model performance evaluated on the test data. Each bootstrap experiment ended with one correlation value between the estimated and observed outcomes regarding the test set. Together, we obtained 1,000 correlation values from all bootstrap experiments. In parallel, we performed another 1,000 bootstrap experiments but with the personalized fine-tuning included as outlined in (3).

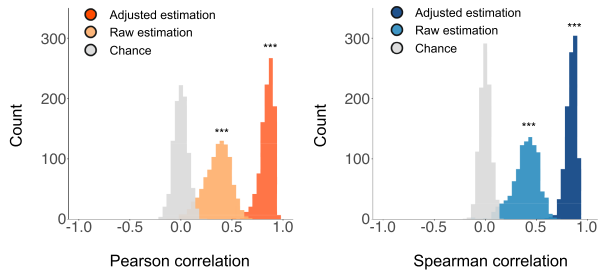
The bootstrap results showed that the proposed method was reproducible across all bootstrap samples with reasonable overall disease assessment results (with an average correlation of 0.34), and additionally, the adjusted personalized disease assessment further improved estimation accuracy (with an average correlation of $r = 0.81$) (see Fig. 6).

Yet, one may still question: how can one be certain that the above analyses justify the efficacy of the model for longitudinal prediction? It remained possible that, if the dataset contained many individuals without much longitudinal variability in their (within-subject) scores, then the above analyses only showed that the model was useful in estimating the average scores. To further examine the model performance in longitudinal disease prediction, we performed two additional analyses detailed in the next sub-section.

B. Evaluating the Longitudinal Model Performance Regarding Predicting the MSIS-29 Scores

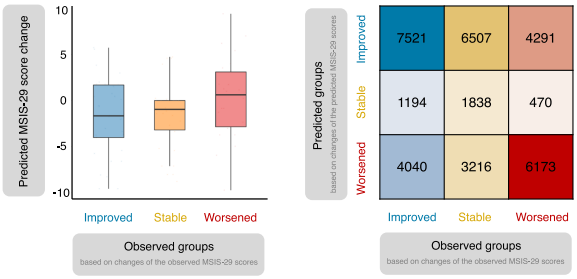
We considered an additional analysis using leave-one-subject-out cross-validation (LOOCV) to determine whether the predicted disease scores captured longitudinal trends. We begin by denoting τ ($\tau > 0$) as the grouping threshold. First, we group the PwMS into three categories: (1) Improved, (2) Stable, and (3) Worsened. The categories were determined by whether the change in MSIS-29 scores at the end of the study from baseline was smaller than $-\tau$ units (denoted as $\Delta < -\tau$, indicating improvement), within plus/minus τ units (denoted as $\Delta \in [-\tau, \tau]$, indicating stable), or greater than $+\tau$ units (denoted as $\Delta > +\tau$, indicating worsening). We set the threshold $\tau = 1/2$ for demonstration purposes. If the directions of the estimated score change agreed with the observed categories, it suggests that the model was able to pick up the longitudinal disease trend. Although the means of the three groups were not pairwise significant (this is, in part, due to the small sample size, and, in part, due to the small longitudinal changes during a period of six months), the LOOCV results suggested that the predicted score changes were generally in line with the observed categories (see Fig. 7). Note that a larger τ would yield more observations to be classified into the stable group; but across different sets of thresholds (*i.e.*, $\tau = 1$, 2, and 3), we observed similar results.

We then performed an additional 1,000 bootstrap experiments (with the same setting in Section III-A). First, individuals were classified into improved, stable, or worsened groups based on whether the disease scores at the end of the study were at least a half unit smaller than, similar to (within plus/minus a half unit), or at least a half unit larger than, the score at the baseline. The longitudinal prediction was carried out and the difference



During each bootstrap test, data from people with MS (PwMS) were randomly drawn from the sample. Each selected sample was then randomly split into a training set and a test set (containing 70% and 30% of the total samples, respectively). Next, multiple imputation was performed and an ensemble GEE was learned from the training data. The trained model was then evaluated on the test data. Subsequently, the (longitudinal) correlation between the estimated and observed disease scores was calculated. This was repeated $1,000$ times, yielding $1,000$ correlations. The light orange and light blue shaded histograms represent the distributions of Pearson and Spearman correlations from $1,000$ bootstrap tests; the dark orange and dark blue shaded histograms represent the distributions of the fine-tuned personalized prediction outcomes. Two null distributions (grey-shaded) were generated from a normal distribution with zero mean and the same standard deviations as the bootstrap Pearson and Spearman correlations.

Fig. 6. Bootstrap results of out-of-sample testing performance for MSIS-29 score prediction.



Left: We performed a leave-one-subject-out cross-validation (LOOCV) analysis and plotted the predicted disease score changes against patients' disease categories. Right: We performed a bootstrap analysis and presented the number of patients falling into each predicted *vs.* observed group (see text for details).

Fig. 7. Results of predicted score change against observed groups in out-of-sample people with MS (PwMS).

between the predicted scores at baseline and the end of the study for each individual was calculated. We then counted the number of individuals belonging to each predicted category against it falling in each observed category (see Fig. 7). Note that some patients showed improved scores due to therapeutic effects or recovery from relapse events.

We next sought to quantify the longitudinal prediction performance. To that end, we calculated the repeated measures correlation ($rmcorr$) [50] for the out-of-sample PwMS prediction to estimate the within-individual association between longitudinal observations and their estimates for multiple individuals. The $rmcorr$ was calculated using the R package $rmcorr$ [50]. Specifically, we considered the $rmcorr$ for estimated and observed MSIS-29 scores at baseline and their counterparts at the end of the study: a positive $rmcorr$ suggested there was evidence that the model was able to detect the longitudinal within-individual disease score change between the baseline and the end of the study. To account for variability due to random splits between training and test samples, we performed 1,000 bootstrap experiments and obtained 1,000 out-of-sample $rmcorr$ values (mean $rmcorr = 0.14$, $st.d. = 0.19$) and Pearson correlations (mean $r = 0.42$, $st.d. = 0.16$), respectively. We note that the variability in the $rmcorr$ estimates does not allow us to rule out the possibility of there being no positive longitudinal correlation; the estimate of $P(rmcorr < 0)$ from the bootstrap samples, however, was only 0.22. These explorations suggested that the model was overall promising to capture the longitudinal within-individual disease trend, with limitations discussed below.

IV. Discussion

The results from Sections III-A and III-B suggest that the utility of the proposed model is twofold. First, it balances group- and subject-level information. The parameters were learned from the training sample and contained information that may be manifested at the group level and can be extrapolated to previously unseen PwMS (see the light orange histogram in Fig. 6). The new PwMS, however, may contain information that may not be captured by the population parameters. The subject-specific fine-tuning using data from day one may narrow the gap (see the dark orange histogram in Fig. 6). Nevertheless, if the PwMS have scores that do not vary much over time, the model may underperform the one using the data from day one alone. Said differently, if the scores of PwMS do not vary much longitudinally, one may as well use a participant's score from day one, say, 50, to predict the future scores; that is, a string of constant scores. Yet, the analyses in Section III-B suggest that even when the individual scores had not progressed considerably, the model seemed to be able to pick up the longitudinal trend. Previous findings suggested that it is possible to obtain longitudinal trend approximate to EDSS scores using gait data (2MWT) collected by smartphones [35]. Our work confirms this, and further expands to the longitudinal proxies of both EDSS and MSIS-29 scores using gait (2MWT) and upper extremity function (DaS) features. Such converging evidence suggests an additional attractive (longitudinal) property of our model; it also suggests that when there are even minor disease dynamics, one should consider a longitudinal model and should refrain from making extrapolations that, since the disease (for an individual) is relatively stable, one can rely on the first disease score to infer (a string of constant) future scores. Finally, when dealing with longitudinal predictions with different levels of missing data, (4) provides a helpful reference to which one can choose the number of imputations based on the amount of missing data to achieve a desirable efficiency.

There are a few limitations of our study. First, the observed disease outcomes were sparse. Particularly limited were the sparsely observed EDSS scores, clinically more relevant than the MSIS-29 scores. Their scarcity obstructs model development (as there were no more than three outcomes per PwMS to train) and hinders model testing (as there were no more than three outcomes per PwMS to evaluate). Even though the proposed model trained on sparse EDSS scores hinted at longitudinal validity (see Table I), we expect the model performance would improve if more EDSS scores were available. Further studies should verify this expectation by training (and testing) the model using more outcomes (as the smartphone data were available at least weekly). By "more outcomes", we mean EDSS scores collected more frequently (*e.g.*, every fortnight), EDSS scores sparsely recorded but over a longer period, or EDSS scores frequently recorded over a long period. Once the model has been (better) trained, it can be applied to new subjects - either sparsely or frequently - to estimate their disease over time. Second, the data were acquired during a relatively short period (24 weeks) wherein the longitudinal scores might not substantially change. Future studies can verify the proposed method in data collected over a much longer period, during which there are likely more disease activity changes. Additionally, this study consists of subjects with relatively mild levels of MS-related functional impairment. Future studies may consider a broader population including healthy controls, patients with mild MS symptoms, and patients with severe, potentially fluctuating disease profiles. Beginnings are already

being made; see, for example, [51], [52]. Third, based on neurological insights and past findings, we were mainly interested in examining the efficacy of gait and upper extremity functions in predicting longitudinal MS progress. Although our findings have suggested the utility of these features in longitudinal MS prediction, we have naturally left out a large territory where other modalities, such as dexterity tests (e.g., pinching test), cognitive tests (e.g., the Symbol Digit Modalities Test (SDMT)), U-turn test, and passively collected digital data, may also be useful to assess MS over time. Further research may, on the one hand, explore the longitudinal prediction performance bestowed by each modality, and, on the other hand, investigate whether, and if so, to what extent, one may improve longitudinal MS prediction by integrating multivariate multi-modal features. Fourth, we used imputation to make use of all recorded features - some of which otherwise not utilized on days with missing values - during model development. During the test, we avoided imputation to prevent it from affecting the test results. This, however, made disease prediction on days with missing data impossible. Although one can, in practice, run a re-test by recording complete data on another day - as the disease is not likely to progress in a few days - future studies should explore advanced imputation methods that can yield complete test data not spuriously boosting predicting performance. Explorations can possibly be made by replacing missing values with data generated from a posterior learned from available test data (hence no leakage from the training data) and a flat prior (also no leakage from the training data) or a weak prior learned from the training data (the weak prior helps data generation, not prediction). Finally, although our model was able to significantly predict subject-specific mean scores and modestly identify the subject-specific longitudinal disease trend, the results of the latter were not statistically significant. This may be due, in part, to the short study period and small sample size and, in part, to the limitation of the model. Future studies should verify the proposed approach on larger datasets and explore additional techniques to improve our method.

V. Conclusion

Personalized longitudinal assessment of MS disease has the potential to inform clinical decisions, and thereby improve treatment outcomes. Smartphone sensor-based assessments offer a new cost-efficient approach to remotely and frequently assessing MS-related functional ability that can complement standard clinical assessments [22]. Smartphone devices are widely available and generate, through their embedded sensors, highly granular and meaningful data suitable for longitudinal modelling of MS. Before such sensor-based assessments can be routinely deployed in clinical practice, it is important to evaluate whether, and if so, to what extent, they can distinguish between-individual differences in disease profiles and uncover within-individual disease courses longitudinally.

In this study, we developed an automated personalized longitudinal framework to assess MS over time. The framework combines MI, GEE, ensemble learning, and subject-specific fine-tuning. MI was used to impute missing data entries; the ensemble GEE was employed for model development and longitudinal prediction of MS disease scores; fine-tuning was introduced to adjust for idiosyncratic disease trajectory.

Using smartphone and clinical data, the framework showed promise to estimate individual longitudinal MS disease profiles in previously unseen PwMS. Particularly, the detected disease changes between baseline and the end of the study agreed in general with the observed changes in MSIS-29 scores.

Taken together, our analyses proved the concept of smartphone-based, personalized MS assessment and demonstrated the potential of the proposed model in longitudinal MS evaluation. Future research needs to test the model using independent datasets and verify if the framework can be extended to evaluate other MS-related clinical outcomes. Future studies may also examine the utility of the method in MS prognosis (namely, predicting the disease before its onset) and explore whether this approach is useful to investigate and forecast other neurodegenerative diseases longitudinally.

Acknowledgment

The authors thank Sven Holm and Guy Nagels for helpful comments.

References

[1] D. S. Reich, C. F. Lucchinetti, and P. A. Calabresi, "Multiple sclerosis," *New England J. Med.*, vol. 378, no. 2, pp. 169–180, 2018.
[2] R. Milo and E. Kahana, "Multiple sclerosis: Geoepidemiology, genetics and the environment," *Autoimmunity Rev.*, vol. 9, no. 5, pp. A387–A394, 2010.
[3] P. Browne et al., "Atlas of multiple sclerosis 2013: A growing global problem with widespread inequity," *Neurology*, vol. 83, no. 11, pp. 1022–1024, 2014.
[4] C. Walton et al., "Rising prevalence of multiple sclerosis worldwide: Insights from the Atlas of MS," *Mult. Scler. J.*, vol. 26, no. 14, pp. 1816–1821, 2020.
[5] I. Kister et al., "Natural history of multiple sclerosis symptoms," *Int. J. MS Care*, vol. 15, no. 3, pp. 146–156, 2013.
[6] N. Yozbatıran, F. Baskurt, Z. Baskurt, S. Ozakbas, and E. Idiman, "Motor assessment of upper extremity function and its relation with fatigue, cognitive function and quality of life in multiple sclerosis patients," *J. Neurological Sci.*, vol. 246, no. 1/2, pp. 117–122, 2006.
[7] J. L. Poole et al., "Dexterity, visual perception, and activities of daily living in persons with multiple sclerosis," *Occup. Ther. Health Care*, vol. 24, no. 2, pp. 159–170, 2010.
[8] K. Lam et al., "Real-world keystroke dynamics are a potentially valid biomarker for clinical disability in multiple sclerosis," *Mult. Scler. J.*, vol. 27, no. 9, pp. 1421–1431, 2021.
[9] A. Bisio, L. Pedullá, L. Bonzano, A. Tacchino, G. Brichetto, and M. Bove, "The kinematics of handwriting movements as expression of cognitive and sensorimotor impairments in people with multiple sclerosis," *Sci. Rep.*, vol. 7, no. 1, pp. 1–10, 2017.
[10] H. L. Zwibel, "Contribution of impaired mobility and general symptoms to the burden of multiple sclerosis," *Adv. Ther.*, vol. 26, no. 12, pp. 1043–1057, 2009.
[11] L. Hemmett, J. Holmes, M. Barnes, and N. Russell, "What drives quality of life in multiple sclerosis?," *Int. J. Med.*, vol. 97, no. 10, pp. 671–676, 2004.
[12] S. Johansson et al., "High concurrent presence of disability in multiple sclerosis," *J. Neurol.*, vol. 254, no. 6, pp. 767–773, 2007.
[13] R. Bertoni, I. Lamers, C. C. Chen, P. Feys, and D. Cattaneo, "Unilateral and bilateral upper limb dysfunction at body functions, activity and participation levels in people with multiple sclerosis," *Mult. Scler. J.*, vol. 21, no. 12, pp. 1566–1574, 2015.
[14] B. K. Tsang and R. Macdonell, "Multiple sclerosis: Diagnosis, management and prognosis," *Australian Fam. Physician*, vol. 40, no. 12, pp. 948–955, 2011.

[15] J. F. Kurtzke, "Rating neurologic impairment in multiple sclerosis: An expanded disability status scale (EDSS)," *Neurology*, vol. 33, no. 11, pp. 1444–1452, 1983.

[16] A. Riazi, J. Hobart, D. Lamping, R. Fitzpatrick, and A. Thompson, "Multiple sclerosis impact scale (MSIS-29): Reliability and validity in hospital based samples," *J. Neurol., Neurosurgery Psychiatry*, vol. 73, no. 6, pp. 701–704, 2002.

[17] J. Hobart, D. Lamping, R. Fitzpatrick, A. Riazi, and A. Thompson, "The multiple sclerosis impact scale (MSIS-29): A new patient-based outcome measure," *Brain*, vol. 124, no. 5, pp. 962–973, 2001.

[18] C. McGuigan and M. Hutchinson, "The multiple sclerosis impact scale (MSIS-29) is a reliable and sensitive measure," *J. Neurol., Neurosurgery Psychiatry*, vol. 75, no. 2, pp. 266–269, 2004.

[19] A. Rae-Grant, A. Bennett, A. E. Sanders, M. Phipps, E. Cheng, and C. Bever, "Quality improvement in neurology: Multiple sclerosis quality measures: Executive summary," *Neurology*, vol. 85, no. 21, pp. 1904–1908, 2015.

[20] A. J. Steelman, "Infection as an environmental trigger of multiple sclerosis disease exacerbation," *Front. Immunol.*, vol. 6, 2015, Art. no. 520.

[21] E. A. Mills, A. Mirza, and Y. Mao-Draayer, "Emerging approaches for validating and managing multiple sclerosis relapse," *Front. Neurol.*, vol. 8, 2017, Art. no. 116.

[22] X. Montalban et al., "A smartphone sensor-based digital outcome assessment of multiple sclerosis," *Mult. Scler. J.*, vol. 28, no. 4, pp. 654–664, 2021.

[23] O. Y. Chén and B. Roberts, "Personalized health care and public health in the digital age," *Front. Digit. Health*, vol. 3, 2021, Art. no. 595704.

[24] L. Midaglia et al., "Adherence and satisfaction of smartphone-and smartwatch-based remote active testing and passive monitoring in people with multiple sclerosis: Nonrandomized interventional feasibility study," *J. Med. Internet Res.*, vol. 21, no. 8, 2019, Art. no. e14863.

[25] J. Poushter, "Smartphone ownership and internet usage continues to climb in emerging economies," 2016. Accessed: Sep. 15, 2022. [Online]. Available: https://policycommons.net/artifacts/618628/smartphone-ownership-and-internet-usage-continues-to-climb-in-emerging-economies/1599614/

[26] S. P. Bhavnani, J. Narula, and P. P. Sengupta, "Mobile technology and the digitization of healthcare," *Eur. Heart J.*, vol. 37, no. 18, pp. 1428–1438, 2016.

[27] B. C. Kieseier and C. Pozzilli, "Assessing walking disability in multiple sclerosis," *Mult. Scler. J.*, vol. 18, no. 7, pp. 914–924, 2012.

[28] M. Kierkegaard, U. Einarsson, K. Gottberg, L. von Koch, and L. W. Holmqvist, "The relationship between walking, manual dexterity, cognition and activity/participation in persons with multiple sclerosis," *Mult. Scler. J.*, vol. 18, no. 5, pp. 639–646, 2012.

[29] A. Creagh et al., "Smartphone-based remote assessment of upper extremity function for multiple sclerosis using the draw a shape test," *Physiol. Meas.*, vol. 41, no. 5, 2020, Art. no. 054002.

[30] A. P. Creagh et al., "Smartphone-and smartwatch-based remote characterisation of ambulation in multiple sclerosis during the two-minute walk test," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 3, pp. 838–849, 2021.

[31] A. K. Bourke, A. Scotland, F. Lipsmeier, C. Gossens, and M. Lindemann, "Gait characteristics harvested during a smartphone-based self-administered 2-minute walk test in people with multiple sclerosis: Test-retest reliability and minimum detectable change," *Sensors*, vol. 20, no. 20, 2020, Art. no. 5906.

[32] A. Scalfari et al., "The natural history of multiple sclerosis, a geographically based study 10: Relapses and long-term disability," *Brain*, vol. 133, no. 7, pp. 1914–1929, 2010.

[33] P. Feys et al., "The Nine-Hole Peg Test as a manual dexterity performance measure for multiple sclerosis," *Mult. Scler. J.*, vol. 23, no. 5, pp. 711–720, 2017.

[34] J. Prince, F. Andreotti, and M. D. Vos, "Multi-source ensemble learning for the remote prediction of Parkinson's disease in the presence of source-wise missing data," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 5, pp. 1402–1411, 2019.

[35] A. P. Creagh, F. Dondelinger, F. Lipsmeier, M. Lindemann, and M. D. Vos, "Longitudinal trend monitoring of multiple sclerosis ambulation using smartphones," *IEEE Open J. Eng. Med. Biol.*, vol. 3, pp. 202–210, 2022, doi: 10.1109/OJEMB.2022.3221306.

[36] J. W. Graham, "Missing data analysis: Making it work in the real world," *Annu. Rev. Psychol.*, vol. 60, no. 1, pp. 549–576, 2009.

[37] J. A. Sterne et al., "Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls," *BMJ*, vol. 338, 2009, Art. no. b2393.

[38] P. T. von Hippel, "New confidence intervals and bias comparisons show that maximum likelihood can beat multiple imputation in small samples," *Struct. Equation Model.: A Multidisciplinary J.*, vol. 23, no. 3, pp. 422–437, 2016.

[39] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*. Hoboken, NJ, USA: Wiley, 2004.

[40] D. B. Rubin, "Multiple imputation after 18 years," *J. Amer. Stat. Assoc.*, vol. 91, no. 434, pp. 473–489, 1996.

[41] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[42] P. D. Allison, *Missing Data*. Thousand Oaks, CA, USA: Sage Publications, 2001.

[43] C. K. Enders, *Applied Missing Data Analysis*. New York, NY, USA: Guilford Press, 2010.

[44] R. J. Little and D. B. Rubin, *Statistical Analysis with Missing Data*. Hoboken, NJ, USA: Wiley, 2019.

[45] J. C. Jakobsen, C. Gluud, J. Wetterslev, and P. Winkel, "When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowcharts," *BMC Med. Res. Methodol.*, vol. 17, no. 1, pp. 1–10, 2017.

[46] R. R. Andridge and R. J. Little, "A review of hot deck imputation for survey non-response," *Int. Stat. Rev.*, vol. 78, no. 1, pp. 40–64, 2010.

[47] S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in R," *J. Stat. Softw.*, vol. 45, pp. 1–67, 2011.

[48] K.-Y. Liang and S. L. Zeger, "Longitudinal data analysis using generalized linear models," *Biometrika*, vol. 73, no. 1, pp. 13–22, 1986.

[49] O. Y. Chén et al., "Building a machine-learning framework to remotely assess Parkinson's disease using smartphones," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 12, pp. 3491–3500, 2020.

[50] J. Z. Bakdash and L. R. Marusich, "Repeated measures correlation," *Front. Psychol.*, vol. 8, 2017, Art. no. 456.

[51] J. van Beek et al., "Floodlight open-a global, prospective, open-access study to better understand multiple sclerosis using smartphone technology," in *Proc. Annu. Meeting Consortium Mult. Scler. Centers*, 2019, Poster QOL10.

[52] S. Roy et al., "Disability prediction in multiple sclerosis using performance outcome measures and demographic data," in *Proc. Conf. Health, Inference, Learn.*, 2022, pp. 375–396.