Disability prediction in multiple sclerosis using performance outcome measures and demographic data

Subhrajit Roy[*] SUBHRAJITROY@GOOGLE.COM
Diana Mincu[*] DMINCU@GOOGLE.COM
Lev Proleev LEVP@GOOGLE.COM
Negar Rostamzadeh NROSTAMZADEH@GOOGLE.COM
Chintan Ghate CHINTANGHATE@GOOGLE.COM
Natalie Harris NATALIEHARRIS@GOOGLE.COM
Christina Chen CHRISTINIUM@GOOGLE.COM
Jessica Schrouff SCHROUFF@GOOGLE.COM
Google Research
Nenad Tomašev NENADT@DEEPMIND.COM
DeepMind
Fletcher Lee Hartsell FLETCHER.HARTSELL@DUKE.EDU
Duke University Health System
Katherine Heller KHELLER@GOOGLE.COM
Google Research
for MSOAC[†]

Abstract

Literature on machine learning for multiple sclerosis has primarily focused on the use of neuroimaging data such as magnetic resonance imaging and clinical laboratory tests for disease identification. However, studies have shown that these modalities are not consistent with disease activity such as symptoms or disease progression. Furthermore, the cost of collecting data from these modalities is high, leading to scarce evaluations. In this work, we used multi-dimensional, affordable, physical and smartphone-based performance outcome measures (POM) in conjunction with demographic data to predict multiple sclerosis disease progression. We performed a rigorous benchmarking exercise on two datasets and present results across 13 clinically actionable prediction endpoints and 6 machine learning models. To the best of our knowledge, our results are the first to show that it is possible to predict disease progression using POMs and demographic data in the context of both clinical trials and smartphone-based studies by using two datasets. Moreover, we investigate our models to understand the impact of different POMs and demographics on model performance through feature ablation studies. We also show that model performance is similar across different demographic subgroups (based on age and sex). To enable this work, we developed an end-to-end reusable pre-processing and machine learning framework which allows quicker experimentation over disparate MS datasets.

Data and Code Availability This paper uses two publicly available datasets: Multiple Sclerosis Outcome Assessments Consortium (MSOAC) (Rudick et al. (2014), https://c-path.org/programs/msoac/) and Floodlight (Baker et al., https://floodlightopen.com/en-US/for-scientists). While our code is not available at this time, we plan to open-source it in the future.

[*] These authors contributed equally

[†] Data used in the preparation of this article were obtained from the Multiple Sclerosis Outcome Assessments Consortium (MSOAC). As such, the investigators within MSOAC contributed to the design and implementation of the MSOAC Placebo database and/or provided placebo data, but did not participate in the analysis of the data or the writing of this report.

1. Introduction

Multiple sclerosis (MS) is a neurological disease that affects around 2.8 million people worldwide and is the leading cause of non-traumatic disability in young adults (The Multiple Sclerosis International Federation, 2020). The primary goal of a clinician treating MS is to manage disease activity and reduce the risk of disability. As such, the ability to accurately predict MS disease progression has the potential to guide therapy and may inform decisions about the most effective care. While machine learning (ML) models have been developed for predicting disease progression in MS (Pinto et al., 2020; Zhao et al., 2017; Rodriguez et al., 2012; Seccia et al., 2020; Tommasin et al., 2021), these approaches primarily rely on using clinically-acquired information such as magnetic resonance imaging (MRI) (Zhao et al., 2017), clinical laboratory tests or clinical history (Seccia et al., 2020). A lack of association between disease activity and these modalities has previously been identified (Whitaker et al., 1995), which in turn led to the development of multi-dimensional performance outcome measures (POMs) such as Multiple sclerosis functional composite (MSFC) scores to accurately track MS disease progression (Rudick et al., 2002). POMs are time-stamped responses collected from MS subjects either through assessment tests or questionnaires, which are used to track disease progression. These include tests to quantify walking ability, balance, cognition, and dexterity – physiological functions that are adversely affected by MS. The frequency of data collection may vary with intervals ranging from a day to multiple months. In addition, they also reduce costs related to personnel, equipment, space, and time requirements compared to neuroimaging or clinical laboratory tests. POMs have also been used alongside neuroimaging-derived data for predicting disability in MS (Law et al., 2019). Moreover, while POMs and demographic data have been used to diagnose MS (Schwab and Karlen, 2021), these have not been used for continually predicting MS disability progression.

In this work, we investigate the possibility of using POMs (physical or electronic) and demographic data for predicting disease progression (in particular disability scores), in MS subjects, in both a clinical and at-home setting. We proof-test this idea using two openly accessible MS datasets: MSOAC (LaRocca et al., 2018) and Floodlight (Baker et al.). Our contributions are as follows:

1. *Novel ML Health solution:* We present for the first time (to the best of our knowledge) that the conjunction of POMs and demographic data can be used to successfully and continually predict long- and short-term MS disease progression for clinical and smartphone-based datasets.
2. *Additional analysis:* We show that model performance is similar across different demographic subgroups (based on age and sex), and perform multiple feature ablations to understand the contributions of different POMS and demographics to the predictions.
3. *Reliability and scalability:* We present a reusable end-to-end pre-processing and machine learning modelling framework that enables benchmarking on different MS datasets. Our proposed framework focuses on reliable dataset ingestion through a common format, scalable label creation and metrics computation.

We envision that our work will not only serve as a first step towards development of machine learning models for monitoring MS, but also spur more ML research in this application area.

2. Methods

2.1. Data description

We looked at two datasets for benchmarking, one recorded in a clinical trial setting (MSOAC), and one from a mobile app in a clinically unsupervised manner (Floodlight). The MSOAC dataset records POMs from physical MSFC tests which were performed by the subjects in-clinic as a part of clinical trials, while the Floodlight dataset records outcome measures collected via an electronic equivalent of MSFC tests taken by the subjects on a smartphone. Both have been previously used for machine learning explorations (Schwab and Karlen, 2020; Walsh et al., 2020). A comparison of the two datasets can be seen in Table 1 and a set of statistics on the data can be found in Appendix E.

MSOAC Placebo Database: The Multiple Sclerosis Outcome Assessments Consortium (MSOAC, (Rudick et al., 2014)) was launched in 2012 to collect, standardize, and analyze data about MS. To that end, their Placebo Database collects data from the placebo arms of 9 different clinical trials (LaRocca et al., 2018) with 2,465 individual patient records.

It contains information on: demographics, medical history, POMs (e.g. timed walk test, dexterity tests,

Table 1: Comparison of the MSOAC and Floodlight datasets.

	MSOAC	**Floodlight**
Modality	Clinical visit	Smartphone
Cohort	MS subjects only	MS subjects + control
Frequency of assessment	3-monthly	Continuous
Test type	Physical Multiple Sclerosis Functional Composite (MSFC) scores	Smartphone-based
Clinician annotation	Expanded Disability Status Scale (EDSS)	None
Number of patients	2,465	2,339

auditory and visual acuity tests), patient reported outcome measures (e.g. health survey), relapse information and the MS sub-type, and clinically vetted measurements such as the Expanded Disability Status Scale (EDSS) (Kurtzke, 1983a).

Floodlight: Floodlight is a mobile app developed by Roche and Genentech (Baker et al.) designed to combat the infrequent measurements observed during clinical visits and allow healthcare professionals to have a greater understanding of the disease. It contains a set of active tests that measure brain function (daily mood question, symbol matching), hand function (draw a shape, on-screen pinching) and mobility (timed two minute walk, balance, u-turn). Unlike MSOAC, it does not contain any clinically or expert defined labels.

The app is still active at the time of writing and the number of enrolled patients is constantly growing. The data snapshot we use was taken on the 15th of June 2021 and has a total of 2,339 subjects, including both MS patients ($n = 1,236$) and control subjects ($n = 1,103$).

2.2. Data processing

Since the type and structure of the data contained in the two datasets are quite different, typically all data processing would be done individually for each dataset. This can lead to slight differences in the resulting model input, making direct comparison between datasets difficult. To avoid this pitfall, we have devised a general data processing, modelling and evaluation pipeline (shown in Figure 1) which enables us to reuse a number of downstream components and do a reliable cross-dataset comparison.

2.2.1. Pipeline overview

The raw data is taken through a set of processing steps into a common representation called Subject, inspired from (Tomašev et al., 2021). Once both the clinical dataset and the smartphone-based dataset are transformed into the Subject representation, a Label Creator runs on all processed data and enriches it with labels (see Sec. 2.3 for a description of the tasks), after which it gets transformed into model input. The labels can be dataset-specific, or common across multiple different datasets. The proposed pipeline is feature agnostic i.e. there is no specific pre-processing in one dataset or the other.

After training, a prediction format (Table A7, Appendix) is used to save all model output. This in turn gets fed into the metrics pipeline which can provide results at both population and subgroup levels.

2.2.2. Common format

A full description of each field present in this representation can be found in Appendix A. At a high level, each subject has some information that is constant across time (static) such as the medical history or the subject's sex [1], but also timestamp-based information (dynamic) encompassing all medical events - either outpatient or inpatient. Multiple such Events can be grouped into an Episode, but an individual Event can also form an Episode on its own.

The part that enables each dataset to be processed individually is the concept of Resources. These can be functional tests, questionnaires, medications or more, depending on the types of data available in each dataset. Each medical event has a set of resources associated with it, and since the types of resources depend on the dataset, this leaves a common overall structure while still allowing for variability.

To better illustrate this structure we can look at what this means for our two datasets. In the case of MSOAC, an Episode corresponds to a visit to the clinic, including all the tests performed. The data available at each visit consists of functional tests,

[1] In MSOAC this is clinician reported, while in Floodlight it is self reported.

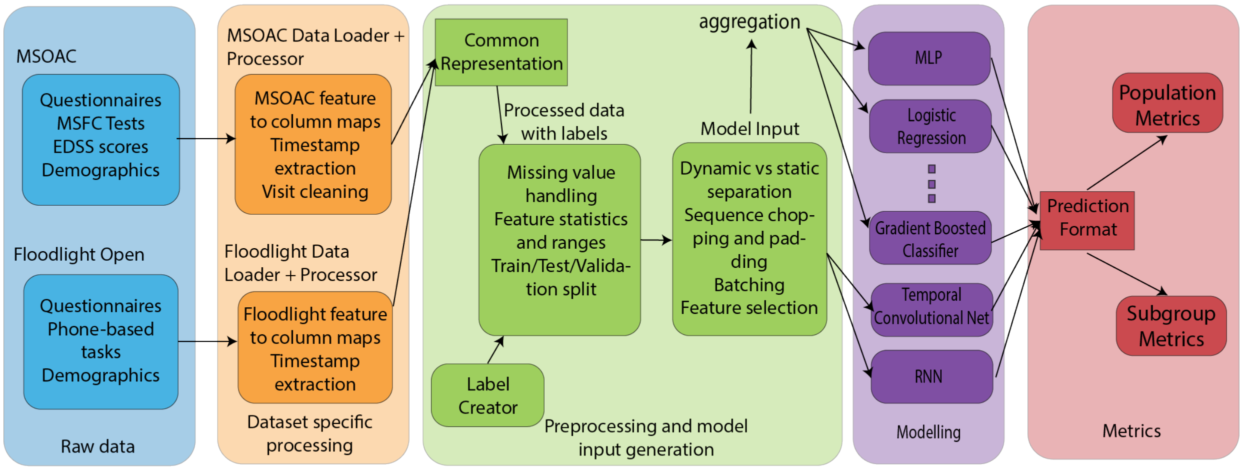
MSOAC
Questionnaires
MSFC Tests
EDSS scores
Demographics
Floodlight Open
Questionnaires
Phone-based tasks
Demographics
Raw data
MSOAC Data Loader + Processor
MSOAC feature to column maps
Timestamp extraction
Visit cleaning
Floodlight Data Loader + Processor
Floodlight feature to column maps
Timestamp extraction
Dataset specific processing
Common Representation
Processed data with labels
Missing value handling
Feature statistics and ranges
Train/Test/Validation split
Label Creator
aggregation
Model Input
Dynamic vs static separation
Sequence chopping and padding
Batching
Feature selection
Preprocessing and model input generation
MLP
Logistic Regression
Gradient Boosted Classifier
Temporal Convolutional Net
RNN
Modelling
Population Metrics
Prediction Format
Subgroup Metrics
Metrics

Figure 1: Dataset processing pipeline with common downstream modelling infrastructure

questionnaires and medications. Each of these is represented as a Resource. Given that MSOAC does not provide timestamps for these resources, they are considered as part of the same Event. For Floodlight, the dataset contains less but more frequent and timestamped measurements. In this case, we define an Episode as 24 hours of data, and each Event corresponds to a new POM. The two types of resources available are functional tests and questionnaires.

We emphasise that the datasets are not merged, but separately converted into the common Subject representation. Our goal is to demonstrate that ML modelling for disability prediction in MS is possible for two disparate datasets using the same modelling framework.

2.3. Prediction tasks

The disability prediction tasks are selected such that they are *clinically actionable* in the context of the particular dataset they are defined in.

2.3.1. MSOAC

2.3.1.1. EDSS: For MSOAC our primary prediction endpoints are derived from clinician-annotated EDSS scores which is a commonly used measure of long-term MS disability (Kurtzke, 1983b). The EDSS scale ranges from 0 to 10, in 0.5 unit increments. EDSS scores can be divided into distinct severity categories: 0-1 for no disability, 1.5-2.5 for mild disability, 3-4.5 for moderate disability, and 5-10 for severe disability. In this paper, we consider both the prediction of raw EDSS scores ($EDSS_{mean}$) and more clinically insightful tasks such as the severity of EDSS scores and whether it crosses a certain disability risk-threshold. Detecting a change in EDSS severity could signal to a clinician the need to change the medication a patient is on, or be used to check whether a treatment is effective.

All tasks are implemented as continuous predictions, triggered at every visit. Fixed prediction horizons are chosen for each task based on expert clinical input on the window of actionability (Table 2). These are 0 - 6 months, 6 - 12 months, 12 - 18 months and 18 - 24 months.

2.3.2. Floodlight

2.3.2.1. Disability scores: Floodlight does not contain any expert annotations, so we developed a score that closely mimics EDSS. EDSS is divided into multiple components measuring different kinds of disability: neural function, ambulatory, and walking. We categorize the assessment tests present in the Floodlight dataset into the above categories and perform a weighted combination of the tests in each category to develop three individual disability scores for separate functional systems. We also compute an overall disability score by taking the average of the individual functional scores. Given no literature exists on how to define disability scores from smartphone-based assessment tests, we rely on expert input and expect the score to be a close proxy of EDSS. These tasks are formulated as regression tasks since smartphone-based tests are relatively new and hence severity categories are not defined in literature. For the purpose of this work, we assume that continuous predictions of this derived disability score provides insights on the progression of the disease. While the EDSS-derived labels in Sec. 2.3.1.1 track long-term changes in disability, the higher frequency recordings of Floodlight enables the prediction of short-term changes.

Similar to the section above, all tasks are posed as continuous predictions, triggered after a new POM is available. Prediction horizons for these tasks are smaller than for MSOAC: 0 - 1 weeks, 1 - 2 weeks and 2 - 4 weeks (Table 3). This is because (a) frequent measurements are possible via a smartphone and hence short-term changes in disability trends can be predicted and (b) most Floodlight users stop using the app after a certain point, so long-term data is not available.

2.3.2.2. Smartphone-based diagnosis of MS: While predicting disability progression in MS is beneficial in both a clinical and at-home context, earlier diagnosis and treatment of MS is considered the best path to fighting it (Miller, 2004). Multiple studies have shown a delay between symptom development, to first medical visit and then finally to diagnosis, with an average of 1-2 years between symptom onset and diagnosis (Ghiasian et al., 2021; Fernández et al., 2010). We believe that the usage of smartphones can enable large-scale diagnosis of MS in a more timely manner.

To test this hypothesis, we use the Floodlight dataset to predict whether a subject has MS or not. This is not possible in MSOAC as we do not have control subjects. The problem statement is formulated as follows: we are given N POMs in total for each subject since the start of data collection, along with the self-reported ground truth on whether or not they have MS. The models have to predict whether the set of N tests belong to a subject with MS. We vary N as = 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100. Note that it is desirable to use less tests for making this prediction, not only to enable earlier access to clinical care, but also to tackle the adherence problem which plagues smartphone-based health studies (Patoz et al., 2021; Pathiravasan et al., 2021).

2.3.3. Cross-dataset

2.3.3.1. Disability progression: To investigate differences in signal between the two datasets, we also define a common label across both MSOAC and Floodlight (Figure A6, Appendix D). It is focused on forecasting disability by predicting substantial deviations of questionnaire and functional test values. This is because: (a) EDSS or other aggregated disability scores are combined across functional systems and (b) EDSS has been criticised to be focused more on mobility and less on cognitive abilities or dexterity (Meyer-Moock et al., 2014). Successfully predicting the deviation of individual functional tests can potentially be more informative for clinicians in understanding which functional systems of a subject are likely to contribute to a subject's future disability. We define the disability progression labels as a change (greater/lesser) of 20% (Goldman et al., 2019) from a baseline, where the baseline is updated as time progresses, for each subject and each feature. This led to a three-class classification problem where each timestamp was annotated with one of the three labels: disability unchanged, improved, or worsened.

2.4. Features and Missingness

We define a set of input features for each task to prevent any label leakage during the training and testing of the model. For MSOAC, we eliminate the EDSS score feature for EDSS-derived targets. For both datasets we only use the questionnaires, functional tests, and patient characteristics such as age, sex, weight, height (dataset dependent, where available). In total, there are 92 distinct features for MSOAC (POMs are multi-component) and 24 for Floodlight.

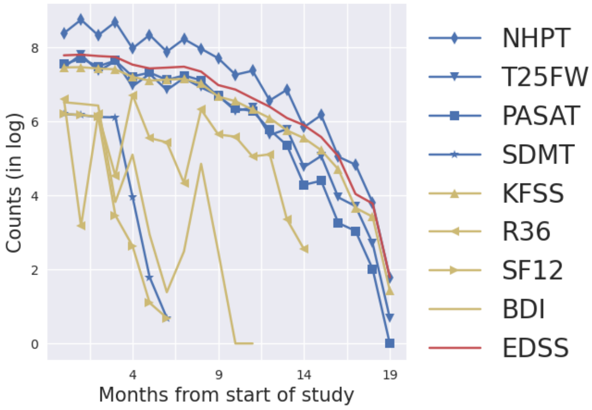

Figure 2: Feature sparsity in MSOAC

To evaluate feature sparsity in our datasets, we look at population level counts for each feature and a given time bucket. For MSOAC we use buckets of 1 month from the start of the study, while for Floodlight we use 1 week intervals from the time the subject joined the app. Based on figure 2 we see that in MSOAC functional tests (Nine Hole Peg Test (NHPT), Timed 25-Foot Walk (T25FW), Paced Auditory Serial Addition Test (PASAT), Symbol Digit Modalities Test (SDMT)) have a much higher count than questionnaires (Kurtzke Functional Systems Scores (KFSS), RAND-36 Item Health Survey (RAND-36), 12-Item Short Form Survey (SF-12), Beck Depression Inventory (BDI)) and are present until much later in the study as well. For Floodlight (Appendix B, Figure A5) we observe a similar downward trend as time progresses, but in this case the functional tests and questionnaires behave in a similar manner. This can be explained by the simplicity of the questionnaires in this dataset, since only one mood related question is available.

Missing values at various timestamps were replaced by 0 in the case of numerical features, or empty string for text-based features.

2.5. Models

For benchmarking purposes, we choose a few popular baseline models from Scikit-Learn (Logistic Regression, Linear Regression, Gradient Boosted Classifier, Gradient Boosted Regressor), a non-sequential (Multi-layer Perceptron), and a sequential deep neural network (Temporal Convolutional Neural Networks (TCN) (Bai et al., 2018).

For predictions, the models use information across a specified window before the prediction timepoint. While for Scikit-Learn models information is aggregated across the window by taking the mean, TCN processes them sequentially and hence retains the temporal information.

We report the area under the precision-recall curve (AU PRC) for the classification tasks since most prediction problems are imbalanced (see Appendix E, Table 9) and R-MSE for regression tasks. We perform 10-fold cross-validation for each model and report the mean and standard deviation of AU PRC / R-MSE. For TCN we use the Adam optimizer (Kingma and Ba, 2015). We performed a hyperparameter search for each model to find the optimal hyperparameters, and the search space is reported in Appendix C.

3. Experiments and results

3.1. Performance on the full feature set

3.1.1. MSOAC

Illustrative example: Figure 3 shows an example usage of our predictive models for the MSOAC dataset. The model is trained to predict whether the patient will transition to a state of moderate disability in the next 0-6 months and 6-12 months interval. Updated risk estimates of future disease worsening are made for every clinic visit throughout the course of the clinical study. Identifying an increased risk of decline sufficiently well in advance can enable early preventative action (Schlaeger et al., 2012). This is possible even when clinicians may not be monitoring a patient or actively intervening.

Performance on EDSS-derived labels: Table 2 summarizes model performance on the labels (see Sec. 2.3.1.1) derived from EDSS scores. We show the mean and standard deviation of the metrics across different folds. We observe that TCN consistently outperforms other ML algorithms by achieving superior performance in all regression and classification tasks. We also see that although the dataset is composed of data from 9 different clinical trials, the models still manage to learn meaningful representations, as demonstrated by the results. For MSOAC, we observe a general trend that, as the prediction interval slides into the future, the tasks progressively get more difficult, leading to a reduction in model performance. This phenomenon has also been shown in other sequential healthcare prediction problems (Tomašev et al., 2021; Nestor et al., 2019).

3.1.2. Floodlight

Performance on disability scores: Table 3 reports model performance on disability scores defined on the Floodlight dataset (see Sec. 2.3.2.1). For Floodlight, we observe that Gradient Boosted Regressor outperforms the other models in all 12 tasks. We believe that this is due to the fact that Floodlight endpoints are zero-inflated (see Appendix E, Table A9) and we intend to explore zero-inflated versions of the models in future. The standard deviations show that the obtained results have tight bounds. Note that the inclusion of multiple families of models allows us to explore and find the best model per dataset or task. Performance on the Floodlight dataset remains relatively similar across different time hori-

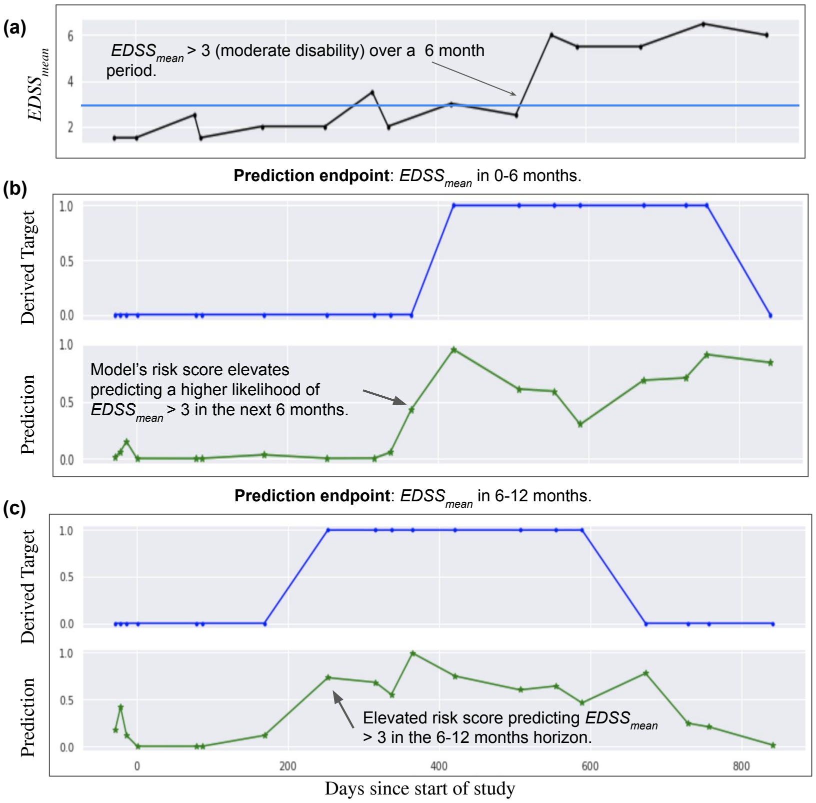

Figure 3: Illustrative example of the trained model successfully predicting that a female subject of age 29 is going to have moderate disability ($EDSS_{mean} > 3$) in the (b) 0-6 and (c) 6-12 months horizon.

zons, potentially since these are not as further out as MSOAC.

Performance on smartphone-based MS diagnosis: We summarize the results of diagnosing MS in this dataset (see Sec. 2.3.2.2) using smartphone-based tests in Figure 4. The results demonstrate that overall TCN achieves the best performance across various values of N (number of tests) and shows 71.67% accuracy at $N = 100$ (approximately 1.5 weeks of app usage). Moreover, while for all models the performance improves as N increases, the improvement is highest for TCN (8.89% increase from N=5 to N=100) and lowest for Logistic Regression (1.95% increase from N=5 to N=100).

3.1.3. Cross-dataset

Performance on disability progression labels: Next we look at the disability progression labels described in Sec. 2.3.3.1 for both the MSOAC and Floodlight dataset. For MSOAC we focus only on TCN, since TCN outperforms all other models for the previous endpoints (see Table 2). We observe that the disability progression tasks have a high class imbalance (92.86–98.27%), and hence we use both cross-entropy loss and focal loss (popular for imbalanced datasets) (Lin et al., 2018) during training. The latter provides a much better performance for this task, with an average AU PRC improvement of 10.27%. The mean and standard deviation of AU PRC are reported in Table 4. Note that we also explored focal loss for the classification labels listed in Table 2, however there was no observable improvement, potentially since these endpoints do not show high imbalance.

While results are promising on MSOAC, the models were not able to obtain significant results using Floodlight. This result might be due to the daily bucketing of measurements, which leads to a label prevalence lower than 0.01%. Weekly bucketing could be considered in the future.

Table 2: Performance (and standard deviation) obtained by machine learning models on a diverse set of prediction tasks for the MSOAC dataset. Best performance is reported in bold.

Dataset	**Prediction tasks**	**Prediction Window**	**Metric**	**Logistic Regression**	**Linear Regression**	**MLP**	**Gradient Boosted Classifier**	**Gradient Boosted Regressor**	**TCN**
MSOAC	$EDSS_{mean}$	0 - 6 mo	R-MSE	-	1.929 (0.098)	-	-	1.700 (0.045)	**1.264 (0.055)**
		6 - 12 mo		-	2.114 (0.111)	-	-	1.901 (0.057)	**1.650 (0.067)**
		12 - 18 mo		-	2.186 (0.115)	-	-	1.892 (0.065)	**1.725 (0.074)**
		18 - 24 mo		-	2.068 (0.133)	-	-	1.748 (0.062)	**1.666 (0.128)**
	$EDSS_{mean}$ > 3 (Moderate disability)	0 - 6 mo	AU PRC	0.803 (0.012)	-	0.826 (0.015)	0.843 (0.017)	-	**0.909 (0.014)**
		6 - 12 mo		0.707 (0.014)	-	0.731 (0.019)	0.756 (0.019)	-	**0.82 (0.027)**
		12 - 18 mo		0.605 (0.025)	-	0.664 (0.036)	0.706 (0.027)	-	**0.768 (0.031)**
		18 - 24 mo		0.502 (0.036)	-	0.594 (0.038)	0.641 (0.038)	-	**0.703 (0.038)**
	$EDSS_{mean}$ > 5 (Severe disability)	0 - 6 mo	AU PRC	0.695 (0.026)	-	0.727 (0.026)	0.785 (0.025)	-	**0.848 (0.035)**
		6 - 12 mo		0.576 (0.028)	-	0.597 (0.027)	0.676 (0.032)	-	**0.722 (0.039)**
		12 - 18 mo		0.457 (0.035)	-	0.504 (0.032)	0.594 (0.034)	-	**0.669 (0.037)**
		18 - 24 mo		0.362 (0.042)	-	0.421 (0.047)	0.536 (0.054)	-	**0.632 (0.037)**
	$EDSS_{mean}$ as severity category	0 - 6 mo	Avg. AU PRC	0.523 (0.015)	-	0.687 (0.01)	0.717 (0.015)	-	**0.782 (0.028)**
		6 - 12 mo		0.47 (0.015)	-	0.633 (0.018)	0.675 (0.016)	-	**0.709 (0.044)**
		12 - 18 mo		0.434 (0.017)	-	0.606 (0.012)	0.649 (0.019)	-	**0.674 (0.037)**
		18 - 24 mo		0.413 (0.017)	-	0.575 (0.02)	0.625 (0.02)	-	**0.632 (0.037)**

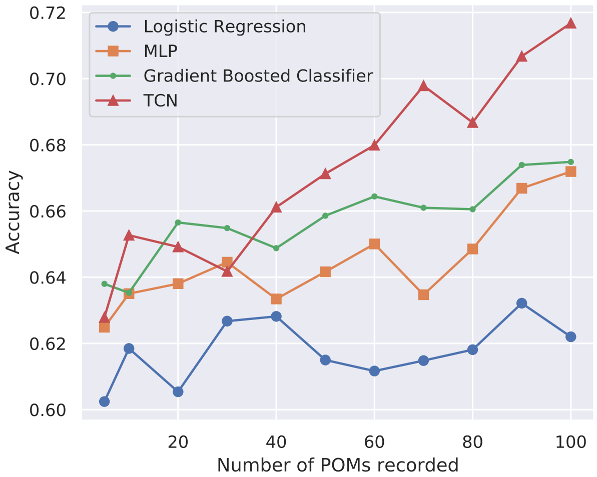

Figure 4: Comparison of model performance for diagnosing MS on Floodlight (smartphone-based dataset).

3.2. Feature ablation

To assess the signal related to disease progression in each feature group, we perform different feature ablation experiments. For both MSOAC and Floodlight we choose the following groups: demographics, questionnaires, functional tests, and all POMs (questionnaires + functional tests).

The results are presented in Table 5. Due to space constraints, we choose a single prediction horizon per dataset (6-12 months for MSOAC and 1-2 weeks for Floodlight), for all the tasks considered in Tables 2 and 3. While in Table 5 we show the results for only the top performing model per dataset (TCN for MSOAC and Gradient Booster Regressor for Floodlight), results for all models are reported in Table A10 (Appendix). The trends obtained for the top performing models are consistent across other models as well.

For MSOAC, the results depict that the feature groups are of following importance: demographics $<$ questionnaires $<$ functional tests $<$ all POMs $<$ full feature set. In line with expectations, the full feature set containing both POMs and demographic features produces the best performance. Demographics (static features) impact model performance the least, and are outperformed by functional tests and questionnaires (dynamic/temporal features). Between functional tests and questionnaires, we observe that the former outperforms the latter. We believe the reason is evident from Figure 2 which shows that the questionnaires are orders of magnitude sparser than functional tests thereby leading to less signal for the ML models.

Table 3: Performance (and standard deviation) obtained by machine learning models on a diverse set of prediction tasks for the Floodlight dataset. Best results are reported in bold.

Dataset	Prediction tasks	Prediction Window	Metric	Linear Regression	Gradient Boosted Regressor	TCN
Floodlight	Cognitive disability score	0 - 1 wk	R-MSE	0.275 (0.015)	**0.262 (0.016)**	0.313 (0.018)
		1 - 2 wks		0.285 (0.015)	**0.275 (0.014)**	0.337 (0.025)
		2 - 4 wks		0.286 (0.014)	**0.279 (0.013)**	0.353 (0.023)
	Dexterity disability score	0 - 1 wk	R-MSE	0.152 (0.011)	**0.146 (0.012)**	0.162 (0.018)
		1 - 2 wks		0.153 (0.012)	**0.148 (0.012)**	0.171 (0.022)
		2 - 4 wks		0.152 (0.011)	**0.149 (0.011)**	0.178 (0.019)
	Mobility disability score	0 - 1 wk	R-MSE	0.244 (0.017)	**0.226 (0.018)**	0.283 (0.021)
		1 - 2 wks		0.256 (0.018)	**0.244 (0.021)**	0.313 (0.026)
		2 - 4 wks		0.26 (0.017)	**0.249 (0.019)**	0.328 (0.027)
	Overall disability score	0 - 1 wk	R-MSE	0.192 (0.012)	**0.18 (0.013)**	0.225 (0.019)
		1 - 2 wks		0.206 (0.012)	**0.197 (0.013)**	0.246 (0.017)
		2 - 4 wks		0.209 (0.011)	**0.209 (0.011)**	0.265 (0.019)

Table 4: Performance (and standard deviation) of TCN on functional test specific disability progression labels defined on the MSOAC dataset.

Functional test	Prediction horizon	Loss type	
		Cross-entropy loss	Focal loss
NHPT	0-6 mo	0.406 (0.030)	**0.583 (0.142)**
	6-12 mo	0.396 (0.021)	**0.534 (0.171)**
PASAT	0-6 mo	0.476 (0.016)	**0.533 (0.172)**
	6-12 mo	0.482 (0.012)	**0.567 (0.161)**
SDMT	0-6 mo	0.471 (0.058)	**0.522 (0.112)**
	6-12 mo	0.534 (0.084)	**0.535 (0.158)**
T25FW	0-6 mo	0.422 (0.017)	**0.558 (0.147)**
	6-12 mo	0.423 (0.015)	**0.600 (0.053)**

For Floodlight, the order of importance of feature groups is as follows: questionnaires < demographics < functional tests < all POMs < full feature set. Compared to MSOAC, the importance of questionnaires and demographics have flipped. This result is expected for Floodlight since it contains only one questionnaire feature (Mood Response) unlike MSOAC consisting of multiple questionnaire features.

3.3. Subgroup results

Apart from performance on the entire dataset, subgroup analysis enables researchers and clinicians to understand where models fall short.

For both datasets we look at the sex and age-bucketed subgroups. Given that MS is a disease that tends to affect more women than men, we assess potential discrepancies in model performance between males and females. Stratification on age is also a relevant evaluation, as MS is a long-term condition and younger patients typically have a less severe form of the disease. As the disease progresses, it can evolve from relapse-remitting to a progressive state (Meca-Lallana et al., 2021), which is often accompanied by an increase in symptom severity. Our age buckets were chosen based on expert input on what would be most clinically useful.

Table A11 (Appendix) contains the results on 3 predictions tasks for MSOAC, for the 6-12 month horizon on all models. We only report results where the subgroup was known, as this information is not present for all patients.

We see that for both males and females the models tend to have a similar, and sometimes identical performance (AU PRC), across all folds. Age on the

Table 5: Summary of feature ablation studies on MSOAC and Floodlight for 6-12 months and 1-2 weeks horizon respectively for the best performing models from Table 2 and 3 respectively. For MSOAC, will $EDSS_{mean}$ reports R-MSE (lower better), all other labels report AU PRC (higher better). For Floodlight, R-MSE is reported for all labels.

Dataset	**Prediction tasks**	**Feature Groups**				
		Demographics	*Functional Tests*	*Questionnaires*	*Performance Outcome Measures*	*Full feature set*
MSOAC	$EDSS_{mean}$	1.957 (0.051)	1.777 (0.091)	1.860 (0.049)	1.676 (0.077)	**1.650 (0.067)**
	$EDSS_{mean} > 3$ (Moderate disability)	0.678 (0.019)	0.766 (0.025)	0.789 (0.020)	0.816 (0.037)	**0.820 (0.027)**
	$EDSS_{mean} > 5$ (Severe disability)	0.456 (0.028)	0.665 (0.036)	0.608 (0.037)	0.691 (0.034)	**0.722 (0.039)**
	$EDSS_{mean}$ as severity category	0.520 (0.011)	0.672 (0.086)	0.659 (0.016)	0.686 (0.042)	**0.709 (0.044)**
Floodlight	Cognitive disability score	0.306 (0.017)	0.286 (0.012)	0.416 (0.037)	0.283 (0.014)	**0.275 (0.014)**
	Dexterity disability score	0.159 (0.014)	0.153 (0.012)	0.198 (0.021)	0.152 (0.012)	**0.148 (0.012)**
	Mobility disability score	0.278 (0.018)	0.249 (0.017)	0.381 (0.031)	0.248 (0.019)	**0.244 (0.021)**
	Overall disability score	0.220 (0.012)	0.206 (0.012)	0.308 (0.034)	0.205 (0.013)	**0.197 (0.013)**

other hand sees a discrepancy when it comes to the various subgroups, with people aged under 30 seeing the biggest decrease in AU PRC. For people aged 50-70 we see that performance is either on par (EDSS >3, EDSS >5) or slightly lower (EDSS as severity category) to that on the full dataset. This is surprising, as the overall composition of MSOAC is relapse-remitting patients and we would expect people in this age group to have a more stable form of the disease. We note that evaluation for patients aged 70 or older is not informative given the low number of cases (n=11) in this group. A distribution of label values based on age can be found in Appendix E.

4. Discussion and future work

In this paper, we show for the first time that it is possible to predict disease progression in MS using POMs, demographic information, and machine learning for both a clinical trial and smartphone-based dataset. Early prediction of disability in both settings has the potential to support MS subjects and healthcare professionals, since the best course of action is early diagnosis and symptom treatment. This in turn can lead to a slower disease progression and a better quality of life for a longer period of time (Cerqueira et al., 2018). Smartphone-based monitoring can also enable early diagnosis of MS.

Temporal patterns of POMs seem to play an important role in the prediction of longer term disability, as shown by the better performance of TCN in all tasks in the MSOAC dataset. Similarly, results on Floodlight display that short-term patterns (a couple of weeks) also carry predictive power. These results suggest that the continuous evaluation of POMs is a promising avenue for the monitoring and early detection of disease progression of MS patients. Both long- and short-term predictions are potentially clinically actionable since while the former may lead to individually-tailored disease-modifying therapies (Robertson and Moreo, 2016), the latter may prompt a clinician to prescribe medications to control symptoms (National MS Society, 2021).

While the full feature set leads to the highest model performance (according to AU PRC) when compared to feature ablations, we note that POMs without demographics perform on par. This result, although preliminary, questions the recording of demographic data for predictive purposes. Future work could further investigate whether demographic data is indeed beneficial in terms of model performance and patient outcome, considering the balance between data need and privacy. Moreover, the relatively higher sparsity of questionnaires compared to functional tests and its eventual impact in model performance points toward the need of collecting more user-friendly questionnaires, more reliably, and over a longer horizon.

The limitations of this study include (i) the disability scores defined for Floodlight, while inspired from EDSS, are experimental and would require further clinical validation to ensure soundness and clinical relevance, and (ii) the disability progression labels (Sec. 2.3.3.1) did not lead to a well-defined machine learning problem for the Floodlight dataset due to a high-frequency of short-term recordings.

The plans for future work for this study are multifold. First, we shall ingest more relevant features (e.g. medication, medical history), in addition to the current POMs and demographic data. Second, we intend to explore different time-bucketing techniques for Floodlight to tackle the imbalance of disability progression labels and more closely relate to clinical actionability. Third, we intend to handle the irregularity of the features by continuous time modelling instead of missing-value imputation (Kidger et al., 2020). Fourth, we plan to further evaluate the robustness of our models, especially across longer time horizons. Fifth, we intend to create a large-scale multi-site smartphone-based MS dataset for further evaluation and potential deployment of the developed models. As discussed in Sec. 3.3, we lack data in specific patient subgroups. Targeted data acquisition in e.g. patients over 70 could be considered for model evaluation, and potentially for model training.

Institutional Review Board (IRB)

Our research does not require IRB approval, since the datasets are publicly available.

Acknowledgments

We would like to thank the Multiple Sclerosis Outcome Assessments Consortium and Genentech for making the two multiple sclerosis datasets publicly available for research.

References

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, 2018.

Mike Baker, Johan van Beek, and Christian Gossens. Digital health: Smartphone-based monitoring of multiple sclerosis using Floodlight. *Nature Portfolio*. URL https://www.nature.com/articles/d42473-019-00412-0.

João J Cerqueira, D Alastair S Compston, Ruth Geraldes, Mario M Rosa, Klaus Schmierer, Alan Thompson, Michela Tinelli, and Jacqueline Palace. Time matters in multiple sclerosis: can early treatment and long-term follow-up ensure everyone benefits from the latest advances in multiple sclerosis? *Journal of Neurology, Neurosurgery & Psychiatry*, 89(8):844–850, 2018. ISSN 0022-3050. doi: 10.1136/jnnp-2017-317509. URL https://jnnp.bmj.com/content/89/8/844.

O Fernández, V Fernández, T Arbizu, G Izquierdo, I Bosca, R Arroyo, J A García Merino, E de Ramón, and The Novo Group. Characteristics of multiple sclerosis at onset and delay of diagnosis and treatment in Spain (The Novo Study). *Journal of Neurology*, 257(9):1500–1507, 2010. ISSN 1432-1459. doi: 10.1007/s00415-010-5560-1. URL https://doi.org/10.1007/s00415-010-5560-1.

Masoud Ghiasian, Mohammad Faryadras, Maryam Mansour, Elham Khanlarzadeh, and Shahir Mazaheri. Assessment of delayed diagnosis and treatment in multiple sclerosis patients during 1990–2016. *Acta Neurologica Belgica*, 121(1):199–204, 2021. ISSN 2240-2993. doi: 10.1007/s13760-020-01528-7. URL https://doi.org/10.1007/s13760-020-01528-7.

Myla D Goldman, Nicholas G LaRocca, Richard A Rudick, Lynn D Hudson, Peter S Chin, Gordon S Francis, Adam Jacobs, Raj Kapoor, Paul M Matthews, Ellen M Mowry, Laura J Balcer, Michael Panzara, Glenn Phillips, Bernard M J Uitdehaag, Jeffrey A Cohen, and on behalf of the Multiple Sclerosis Outcome Assessments Consortium. Evaluation of multiple sclerosis disability outcome measures using pooled clinical trial data. *Neurology*, 93(21):e1921—-e1931, 2019. ISSN 0028-3878. doi: 10.1212/WNL.0000000000008519. URL https://n.neurology.org/content/93/21/e1921.

Daniel Golovin, Benjamin Solnik, Subhodeep Moitra, Greg Kochanski, John Elliot Karro, and D. Sculley, editors. *Google Vizier: A Service for Black-Box Optimization*, 2017. URL http://www.kdd.org/kdd2017/papers/view/google-vizier-a-service-for-black-box-optimization

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series, 2020.

Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pages 1–15, 2015.

John F Kurtzke. Rating neurologic impairment in multiple sclerosis. *Neurology*, 33(11):1444, 1983a. ISSN 0028-3878. doi: 10.1212/WNL.33.11.1444. URL https://n.neurology.org/content/33/11/1444.

John F Kurtzke. Rating neurologic impairment in multiple sclerosis. *Neurology*, 33(11):1444, 1983b. ISSN 0028-3878. doi: 10.1212/WNL.33.11.1444. URL https://n.neurology.org/content/33/11/1444.

Nicholas G LaRocca, Lynn D Hudson, Richard Rudick, Dagmar Amtmann, Laura Balcer, Ralph Benedict, Robert Bermel, Ih Chang, Nancy D Chiaravalloti, Peter Chin, Jeffrey A Cohen, Gary R Cutter, Mat D Davis, John DeLuca, Peter Feys, Gordon Francis, Myla D Goldman, Emily Hartley, Raj Kapoor, Fred Lublin, Gary Lundstrom, Paul M Matthews, Nancy Mayo, Richard Meibach, Deborah M Miller, Robert W Motl, Ellen M Mowry, Rob Naismith, Jon Neville, Jennifer Panagoulias, Michael Panzara, Glenn Phillips, Ann Robbins, Matthew F Sidovar, Kathryn E Smith, Bjorn Sperling, Bernard M J Uitdehaag, Jerry Weaver, and for the Multiple Sclerosis Outcome Assessments Consortium (MSOAC). The MSOAC approach to developing performance outcomes to measure and monitor multiple sclerosis disability. *Multiple Sclerosis Journal*, 24(11):1469–1484, 2018. doi: 10.1177/1352458517723718. URL https://doi.org/10.1177/1352458517723718.

Marco T K Law, Anthony L Traboulsee, David K B Li, Robert L Carruthers, Mark S Freedman, Shannon H Kolind, and Roger Tam. Machine learning in secondary progressive multiple sclerosis: an improved predictive model for short-term disability progression. *Multiple Sclerosis Journal - Experimental, Translational and Clinical*, 5(4):2055217319885983, 2019. doi: 10.1177/2055217319885983. URL https://doi.org/10.1177/2055217319885983.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. 2018.

Virginia Meca-Lallana, Leticia Berenguer-Ruiz, Joan Carreres-Polo, Sara Eichau-Madueño, Jaime Ferrer-Lozano, Lucía Forero, Yolanda Higueras, Nieves Téllez Lara, Angela Vidal-Jordana, and Francisco Carlos Pérez-Miralles. Deciphering multiple sclerosis progression. *Frontiers in Neurology*, 12, 2021. ISSN 1664-2295. doi: 10.3389/fneur.2021.608491. URL https://www.frontiersin.org/article/10.3389/fneur.2021.608491.

Sandra Meyer-Moock, You-Shan Feng, Mathias Maeurer, Franz-Werner Dippel, and Thomas Kohlmann. Systematic literature review and validity evaluation of the Expanded Disability Status Scale (EDSS) and the Multiple Sclerosis Functional Composite (MSFC) in patients with multiple sclerosis. *BMC Neurology*, 14(1):58, 2014. ISSN 1471-2377. doi: 10.1186/1471-2377-14-58. URL https://doi.org/10.1186/1471-2377-14-58.

J. R. Miller. The importance of early diagnosis of multiple sclerosis. *J Manag Care Pharm*, 10(3 Suppl B):4–11, Jun 2004.

National MS Society. Symptom management, 2021. URL https://www.nationalmssociety.org/For-Professionals/Clinical-Care/Managing-MS/Symptom-Management.

Bret Nestor, Matthew McDermott, Willie Boag, Gabriela Berner, Tristan Naumann, Michael Hughes, Anna Goldenberg, and Marzyeh Ghassemi. Feature robustness in non-stationary health records: Caveats to deployable model performance in common clinical machine learning tasks, 08 2019.

Chathurangi H Pathiravasan, Yuankai Zhang, Ludovic Trinquart, Emelia J Benjamin, Belinda Borrelli, David D McManus, Vik Kheterpal, Honghuang Lin, Mayank Sardana, Michael M Hammond, Nicole L Spartano, Amy L Dunn, Eric Schramm, Christopher Nowak, Emily S Manders, Hongshan Liu, Jelena Kornej, Chunyu Liu, and Joanne M Murabito. Adherence of mobile app-based surveys and comparison with traditional surveys: ecohort study. *J Med Internet Res*, 23(1): e24773, Jan 2021. ISSN 1438-8871. doi: 10.

2196/24773. URL http://www.jmir.org/2021/1/e24773/.

Marie-Camille Patoz, Diego Hidalgo-Mazzei, Bruno Pereira, Olivier Blanc, Ingrid de Chazeron, Andrea Murru, Norma Verdolini, Isabella Pacchiarotti, Eduard Vieta, Pierre-Michel Llorca, and Ludovic Samalin. Patients' adherence to smartphone apps in the management of bipolar disorder: a systematic review. *International Journal of Bipolar Disorders*, 9(1):19, 2021. ISSN 2194-7511. doi: 10.1186/s40345-021-00224-6. URL https://doi.org/10.1186/s40345-021-00224-6.

Mauro F. Pinto, Hugo Oliveira, Sónia Batista, Luís Cruz, Mafalda Pinto, Inês Correia, Pedro Martins, and César Teixeira. Prediction of disease progression and outcomes in multiple sclerosis with machine learning. *Scientific Reports*, 10 (1):1–13, 2020. ISSN 20452322. doi: 10.1038/s41598-020-78212-6. URL https://doi.org/10.1038/s41598-020-78212-6.

D. Robertson and N. Moreo. Disease-Modifying Therapies in Multiple Sclerosis: Overview and Treatment Considerations. *Fed Pract*, 33(6):28–34, Jun 2016.

Juan Diego Rodriguez, Aritz Perez, David Arteta, Diego Tejedor, and Jose A. Lozano. Using multidimensional bayesian network classifiers to assist the treatment of multiple sclerosis. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(6):1705–1715, 2012. doi: 10.1109/TSMCC.2012.2217326.

R A Rudick, G Cutter, and S Reingold. The Multiple Sclerosis Functional Composite: a new clinical outcome measure for multiple sclerosis trials. *Multiple Sclerosis Journal*, 8(5):359–365, 2002. doi: 10.1191/1352458502ms845oa. URL https://doi.org/10.1191/1352458502ms845oa.

Richard A Rudick, Nicholas LaRocca, Lynn D Hudson, and MSOAC. Multiple sclerosis outcome assessments consortium: Genesis and initial project plan. *Multiple Sclerosis Journal*, 20(1):12–17, 2014. doi: 10.1177/1352458513503392. URL https://doi.org/10.1177/1352458513503392. PMID: 24057430.

R Schlaeger, M D'Souza, C Schindler, L Grize, S Dellas, E W Radue, L Kappos, and P Fuhr. Prediction of long-term disability in multiple sclerosis. *Multiple Sclerosis Journal*, 18(1):31–38, 2012. doi: 10.1177/1352458511416836. URL https://doi.org/10.1177/1352458511416836.

Patrick Schwab and Walter Karlen. A deep learning approach to diagnosing multiple sclerosis from smartphone data. *CoRR*, abs/2001.09748, 2020. URL https://arxiv.org/abs/2001.09748.

Patrick Schwab and Walter Karlen. A deep learning approach to diagnosing multiple sclerosis from smartphone data. *IEEE Journal of Biomedical and Health Informatics*, 25(4):1284–1291, Apr 2021. ISSN 2168-2208. doi: 10.1109/jbhi.2020.3021143. URL http://dx.doi.org/10.1109/JBHI.2020.3021143.

Ruggiero Seccia, Daniele Gammelli, Fabio Dominici, Silvia Romano, Anna Chiara Landi, Marco Salvetti, Andrea Tacchella, Andrea Zaccaria, Andrea Crisanti, Francesca Grassi, and Laura Palagi. Considering patient clinical history impacts performance of machine learning models in predicting course of multiple sclerosis. *PLOS ONE*, 15(3): 1–18, 2020. doi: 10.1371/journal.pone.0230219. URL https://doi.org/10.1371/journal.pone.0230219.

The Multiple Sclerosis International Federation. Atlas of MS. 2020. URL https://www.msif.org/resource/atlas-of-ms-2020/. [Online; accessed 3-Sep-2020].

Nenad Tomašev, Natalie Harris, Sebastien Baur, Anne Mottram, Xavier Glorot, Jack W Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Valerio Magliulo, Clemens Meyer, Suman Ravuri, Ivan Protsyuk, Alistair Connell, Cían O Hughes, Alan Karthikesalingam, Julien Cornebise, Hugh Montgomery, Geraint Rees, Chris Laing, Clifton R Baker, Thomas F Osborne, Ruth Reeves, Demis Hassabis, Dominic King, Mustafa Suleyman, Trevor Back, Christopher Nielson, Martin G Seneviratne, Joseph R Ledsam, and Shakir Mohamed. Use of deep learning to develop continuous-risk models for adverse event prediction from electronic health records. *Nature Protocols*, 16(6): 2765–2787, 2021. ISSN 1750-2799. doi: 10.1038/s41596-021-00513-5. URL https://doi.org/10.1038/s41596-021-00513-5.

Silvia Tommasin, Sirio Cocozza, Alessandro Taloni, Costanza Giannì, Nikolaos Petsas, Giuseppe Pontillo, Maria Petracca, Serena Ruggieri, Laura De

Giglio, Carlo Pozzilli, Arturo Brunetti, and Patrizia Pantano. Machine learning classifier to identify clinical and radiological features relevant to disability progression in multiple sclerosis. *Journal of Neurology*, (0123456789), 2021. ISSN 14321459. doi: 10.1007/s00415-021-10605-7. URL https://doi.org/10.1007/s00415-021-10605-7.

Jonathan R. Walsh, Aaron M. Smith, Yannick Pouliot, David Li-Bland, Anton Loukianov, and Charles K. Fisher. Generating digital twins with multiple sclerosis using probabilistic neural networks, 2020.

JN Whitaker, HF McFarland, P Rudge, and SC Reingold. Outcomes assessment in multiple sclerosis clinical trials: a critical analysis. *Multiple Sclerosis Journal*, 1(1):37–47, 1995. doi: 10.1177/135245859500100107. URL https://doi.org/10.1177/135245859500100107. PMID: 9345468.

Yijun Zhao, Brian C. Healy, Dalia Rotstein, Charles R.G. Guttmann, Rohit Bakshi, Howard L. Weiner, Carla E. Brodley, and Tanuja Chitnis. Exploration of machine learning techniques in predicting multiple sclerosis disease course. *PLoS ONE*, 12(4):1–13, 2017. ISSN 19326203. doi: 10.1371/journal.pone.0174866.

Appendix A. Subject representation and MS dataset conversion

As described in Sec. 2.2, we convert the MS datasets into a common representation called Subject described in Table 6. The Subject representation allows us to map a diverse set of datasets into a common format consisting a pre-defined set of fields. This allows not only for an easier downstream processing of multiple datasets, but potentially joining multiple datasets into one.

The fields in Subject are chosen in a way that it stores all information relevant to MS datasets (both clinical and at-home), and to generalize to other healthcare datasets as well. The representation can also be expanded to include unique dataset-specific intricacies.

The Prediction format is described in Table 7, and while it's simplistic in its setup, it enables the downstream metrics pipeline development and cross-model comparisons. While in its current form it mainly focuses on time series (through the use of the timestamp field), we believe that it can be adapted to other types as well by simply ignoring the time value. We chose to store a series of label targets and predictions at once, to ease at-scale metrics computations. Thus, our pipeline computes a variety of metrics for all tasks at once.

Appendix B. Feature statistics

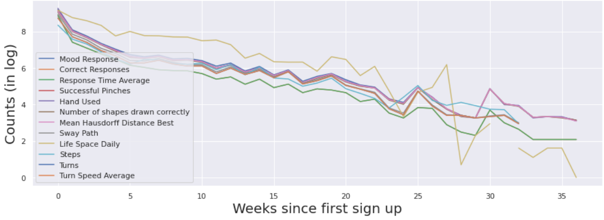

Figure 5: Feature sparsity in Floodlight

Figure 5 shows a summary of the Floodlight functional tests and questionnaires, and how often they appear at a population level, when each patient sequence is bucketed into 1-week intervals. The start of the sequence is when they joined the app.

We observe a similar trend to that on MSOAC, with time playing an important factor in feature sparsity. As mentioned in the main text, both functional tests and questionnaires follow a similar pattern due to the simplicity of questionnaires (only one mood related question). This is consistent with the patient drop-off that takes place in Floodlight, which in turn is consistent with other mobile-based studies.

We believe a focus on preventing attrition is needed for an increase in mobile dataset quality, as these types of datasets have the power to harness diverse information at-scale.

Appendix C. Hyperparameter search space

The hyperparameter search space for each model can be found in Table 8. Future work will look into using parameter auto-tuning tools such as Vizier (Golovin et al., 2017), in order to expand the search space and identify the optimal set-up for our suite of models.

Table 6: Description of the common Subject representation.

Field	Type	Format	Description
Subject: Defines all the data provided for a single subject.			
subject_id	string	Optional	Unique ID for each subject.
subject_characteristics	SubjectCharacteristics	Optional	Defines the subject's characteristics.
medical_history	MedicalHistory	Optional	Defines the subject's medical history.
episodes	Episode	Repeated	A sequence of encounters corresponding to a single subject.
SubjectCharacteristics: Defines all the data provided for a single subject.			
sex	Sex.Enum	Optional	The subject's sex.
MedicalHistory: Defines a sequence of outpatient events recorded before the current system started recording events. These typically are recorded with the same timestamp even though they took place over longer periods of time.			
clinical_events	ClinicalEvent	Repeated	A clinical event.
Episode: An abstraction of an event sequence.			
clinical_event	ClinicalEvent	Repeated	A sequence of clinical events.
changing_characteristics	ChangingSubjectCharacteristics	optional	Defines the subject's changing characteristics.
clinical_trial	ClinicalTrial	optional	Details about the clinical trial the subject was part of (if applicable).
ChangingSubjectCharacteristics: Defines the subject's changing characteristics.			
age	float	optional	The subject's age.
gender	Gender.Enum	optional	The subject's gender
race	string	optional	The subject's race.
weight	float	optional	The subject's weight.
height	float	optional	The subject's height.
country	string	optional	The subject's country.
Sex.Enum: Defines the subject's sex information.			
FEMALE	0	-	Female sex.
MALE	1	-	Male sex.
Gender.Enum: Defines the subject's gender information.			
UNKNOWN	0	-	Unknown gender.
FEMALE	1	-	Female gender.
MALE	2	-	Male gender.
OTHER	3	-	Other gender.
ClinicalEvent: Defines a single event or a set of coinciding individual events that happen at the same time.			
timestamp	int64	optional	Timestamp of event. While for MSOAC, this corresponds to the day of the clinic visit, for Floodlight, this is the timestamp recorded by the smartphone.
resources	Resource	repeated	A list of specific clinical entries recorded at this timestamp.
classification_labels	map<string, int64>	required	Classification labels for prediction.
regression_labels	map<string, float>	required	Regression labels for prediction.
Resource: Describes the various types of resources that can be contained within ClinicalEvent.			
functional_test	FunctionalTest	optional	Functional assessment test data.
questionnaire	Questionnaire	optional	Questionnaires filled by the subjects.
generic_resource	GenericResource	optional	Generic resource to store dataset-specific intricacies.
FunctionalTest: Functional assessment test data recorded from the subject.			
name	string	optional	Name of performance outcome measure.
category	string	optional	Category of performance outcome measure.
response	NumericalResponse	optional	Numerical response recorded from subject.
Questionnaire: Questionnaire data recorded from the subject.			
name	string	optional	Name of performance outcome measure.
category	string	optional	Category of performance outcome measure.
response	QuestionnaireResponse	optional	Questionnaire response recorded from subject.
NumericalResponse: Numeric response converted to standardized unit.			
numerical_response_std_unit	float	optional	Numeric response converted to standardized unit.
std_unit	string	optional	Standardized unit. This unit was used for homogenizing the data.
QuestionnaireResponse: Responses recorded from questionnaires.			
text_response_orig	string	optional	Response in original text.
text_response_std	string	optional	Response in standardized text.
numeric_response_std	float	optional	Standardized numeric response.
categorical_response_std	string	optional	Standardized categorical response.An example entry is EDSS.

Table 7: Description of the common Prediction representation.

Field	**Type**	**Description**
Prediction: Defines the model output information.		
subject_id	string	Unique ID for each subject.
timestamp	int64	Timestamp of event. While for MSOAC, this corresponds to the day of the clinic visit, for Floodlight, this is the . timestamp recorded by the smartphone.
label_targets	map<string, float>	A mapping from label name to the target value for the particular timestamp this is recorded for.
label_predictions	map<string, list<float>>	A mapping from label name to the predicted values for the particular timestamp this is recorded for. Multiple values are recorded to account for multi-class predictions.
subgroup_attributes	map<string, oneof<string, int, float>>	A mapping from subgroup name (such as Sex, or Age) to the exact value (e.g. Female, or 56).

Table 8: Hyperparameter search for models considered in this study.

Model	**Hyperparameter search space**
Logistic Regression	C = [1e-2, 1e-1, 1., 1e+1, 1e+2]
Linear Regression	-

MLP	network_size = [(16, 16), (16, 16, 16), (32, 32)]
	learning_rate = [0.001, 0.01]

Gradient Boosted Classifier	n_estimators = [100, 150]
	learning_rate = [0.001, 0.01]
	max_depth = [3, 5]

Gradient Boosted Regressor	n_estimators = [100, 150]
	learning_rate = [0.001, 0.01]
	max_depth = [3, 5]

TemporalConvNet	num_filters = [16, 32, 64]
	kernel_size = [3, 5]
	learning_rate = [0.001, 0.05, 0.01]
	dropout = [0.0, 0.5]

Appendix D. Disability progression labels

Figure 6 illustrates the computation of the baseline values and how they are used to create the final Worsening/Unchanged/Improved outcome, for each functional test and questionnaire. In the first c timesteps we compute a baseline value for each feature. For each following timestep we perform two actions:

- We compare the value at the new timestep with the baseline value we have for that feature. If the difference in value is greater than a threshold (in our case 20% increase or decrease), we set a label of Worsening/Improving. Otherwise the label value gets set to Unchanged.

- We update the baseline value for each feature based on this new information.

While this task could have been posed as "are any of the tests/questionnaires deviating from the baseline?", the per-test prediction was considered to be more clinically actionable as the actual test is more informative than "something is wrong".

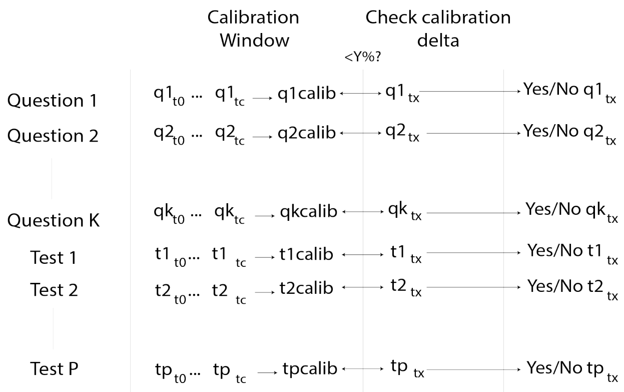
Calibration Window
Check calibration delta
<Y%?
Question 1
Question 2
Question K
Test 1
Test 2
Test P
$q1_{t0}$ ... $q1_{tc}$ → q1calib ↔ $q1_{tx}$ → Yes/No $q1_{tx}$
$q2_{t0}$ ... $q2_{tc}$ → q2calib ↔ $q2_{tx}$ → Yes/No $q2_{tx}$
qk_{t0} ... qk_{tc} → qkcalib ↔ qk_{tx} → Yes/No qk_{tx}
$t1_{t0}$ ... $t1_{tc}$ → t1calib ↔ $t1_{tx}$ → Yes/No $t1_{tx}$
$t2_{t0}$ ... $t2_{tc}$ → t2calib ↔ $t2_{tx}$ → Yes/No $t2_{tx}$
tp_{t0} ... tp_{tc} → tpcalib ↔ tp_{tx} → Yes/No tp_{tx}

Figure 6: Disability progression label definition

Appendix E. Statistics of prediction tasks

Table 9 shows the class distribution for both binary and multi-class classification tasks, while Figures 7 and 8 present a histogram of the label values for the regression tasks.

We note a higher class imbalance for $EDSS_{mean} > 5$, even for the shorter horizon of 0-6 months, while $EDSS_{mean} > 3$ starts off in a more balanced state for the same timeframe. $EDSS_{mean}$ as severity category shows a similar trend to $EDSS_{mean} > 3$, with the shorter time horizon being balanced and the longer time horizon seeing "No disability" as the most prevalent label. We believe this is due to the fact that we do not have information so far into the future, so the default values of "No disability" get used instead. Future work will look into handling this lack of future information. For Floodlight we note that the labels are zero-inflated.

For the classification tasks in MSOAC, we can see the class distribution for the age subgroups, by each cross-validation split, in Figures 9, 10 and 11. Note that splits tend to have very different distributions of age groups, which explains the higher standard deviation for the less prominent groups (age <30).

Appendix F. Feature ablation studies

Table 10 contains results for the feature ablation studies on both MSOAC and Floodlight. For MSOAC the label horizon chosen was 6-12 months, while for Floodlight it was 1-2 weeks.

Appendix G. Subgroup analysis

Table 11 presents subgroup results for the classification tasks performed on the MSOAC dataset, for the 6-12 month horizon.

Appendix H. Ethical considerations and broader impact

Employing easily accessible information in diagnosis and predicting the progression of MS, can have many advantages, including but not limited to better choices of treatment and interventions for each patient, and hopefully reducing the number of relapses in RRMS and hence, the disability of patients. However, these studies should be done with a great amount of care. First, multiple studies have shown great disparity of results between various demographics, typically stemming from representation issues in datasets. Besides potential disparity of results among demographics, we should pay great care about where and how this research is being used. This is of importance, especially where the resources are scarce. In addition, it is important to note, we intend the outcome of this study to be used for patients' access to better choices of treatment and not for this information to be used for unintended purposes such as insurance policies.

Table 9: Label distributions for classification tasks

Dataset	Prediction tasks	Prediction Window	Class	Percentage
MSOAC	$EDSS_{mean}$ > 3 (Moderate disability)	0 - 6 mo	False	53.49
			True	46.51
		6 - 12 mo	False	60.02
			True	39.98
		12 - 18 mo	False	69.47
			True	30.53
		18 - 24 mo	False	77.28
			True	22.72
	$EDSS_{mean}$ > 5 (Severe disability)	0 - 6 mo	False	77.3
			True	22.7
		6 - 12 mo	False	79.8
			True	20.2
		12 - 18 mo	False	83.83
			True	16.17
		18 - 24 mo	False	87.73
			True	12.27
	$EDSS_{mean}$ as severity category	0 - 6 mo	No disability	24.52
			Mild	24.63
			Moderate	26.17
			Severe	24.66
		6 - 12 mo	No disability	36.72
			Mild	20.05
			Moderate	21.57
			Severe	21.65
		12 - 18 mo	No disability	53.04
			Mild	14.12
			Moderate	15.47
			Severe	17.35
		18 - 24 mo	No disability	64.90
			Mild	10.52
			Moderate	11.42
			Severe	13.14

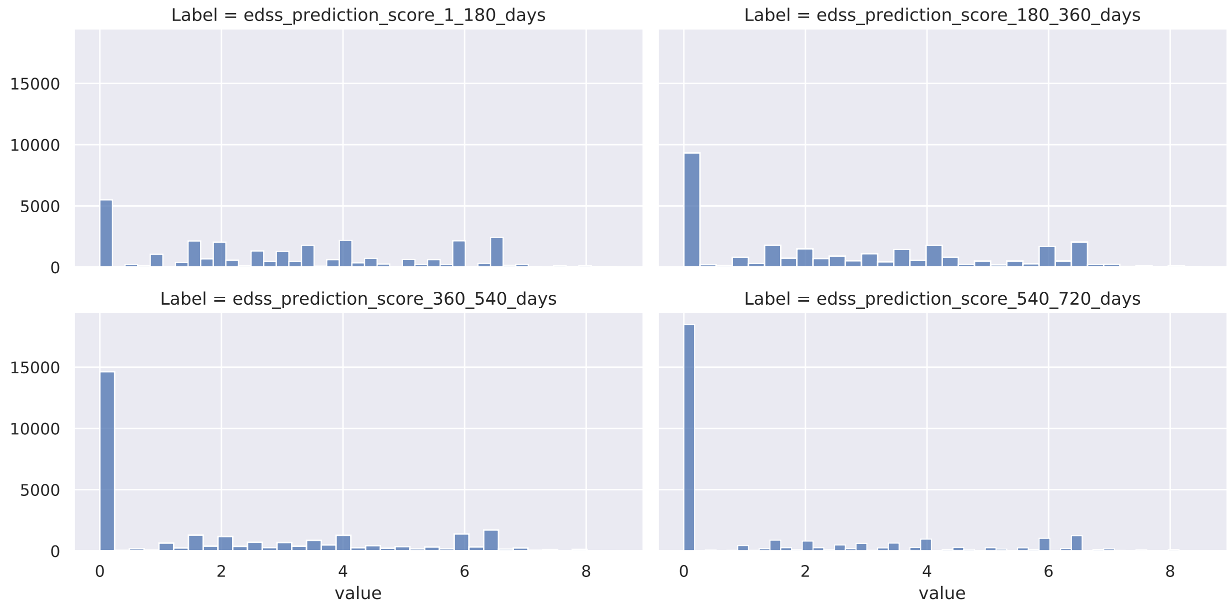

Figure 7: Histograms of regression labels derived from EDSS scores recorded in the MSOAC dataset.

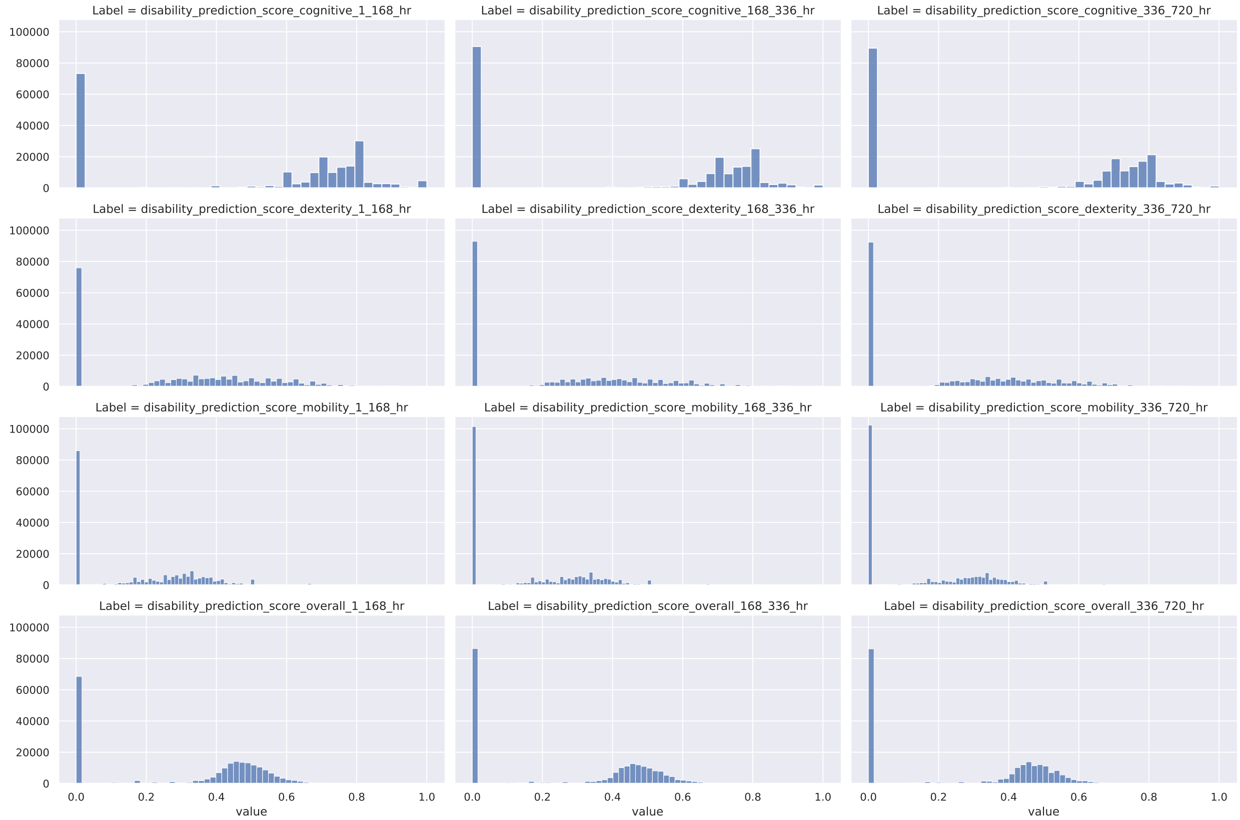

Figure 8: Histograms of regression labels derived from the Floodlight dataset.

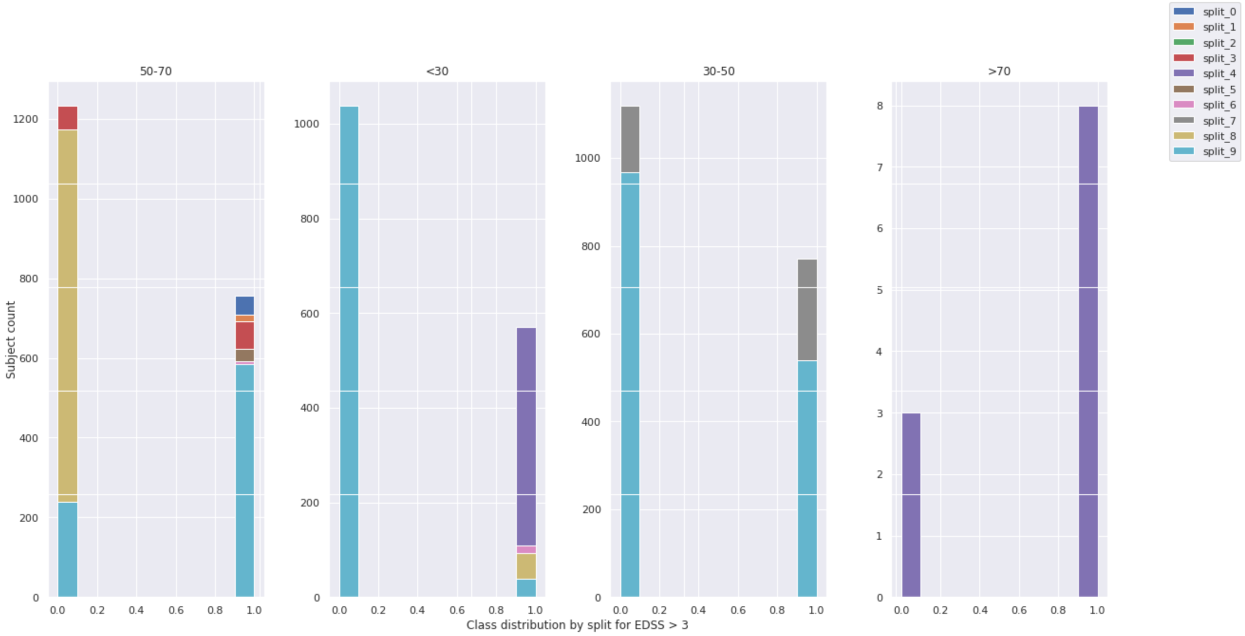

Figure 9: Class distribution by split for EDSS >3.

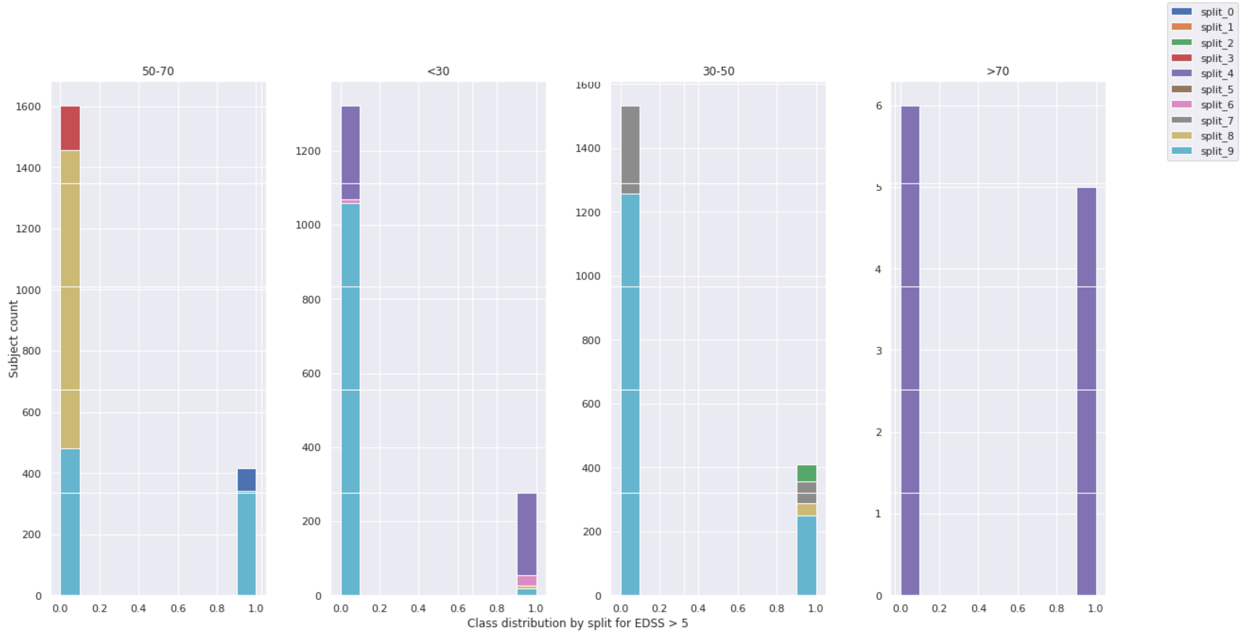

Figure 10: Class distribution by split for EDSS >5.

Table 10: Summary of feature ablation studies on MSOAC and Floodlight for 6-12 months and 1-2 weeks horizon respectively.

Dataset	Prediction tasks	Models	Feature Groups				
			Demographics	*Functional Tests*	Questionnaires	*Performance Outcome Measures*	*Full feature set*
MSOAC	$EDSS_{mean}$	Linear Regression	2.497 (0.091)	2.138 (0.050)	2.271 (0.052)	2.131 (0.083)	2.114 (0.111)
		Gradient Boosted Regressor	2.342 (0.062)	2.045 (0.058)	2.32 (0.063)	1.951 (0.088)	1.901(0.057)
		TCN	1.957 (0.051)	1.777 (0.091)	1.860 (0.049)	1.676 (0.077)	1.650 (0.067)
	$EDSS_{mean}$ >3 (Moderate disability)	Logistic Regression	0.529 (0.022)	0.597 (0.025)	0.631 (0.014)	0.693 (0.011)	0.707 (0.014)
		MLP	0.558 (0.020)	0.662 (0.025)	0.630 (0.027)	0.716 (0.017)	0.731 (0.019)
		Gradient Boosted Classifier	0.573 (0.020)	0.67 (0.018)	0.713 (0.015)	0.75 (0.016)	0.756 (0.019)
		TCN	0.678 (0.019)	0.766 (0.025)	0.789 (0.020)	0.816 (0.037)	0.82 (0.027)
	$EDSS_{mean}$ >5 4*(Severe disability)	Logistic Regression	0.344 (0.036)	0.467 (0.032)	0.491 (0.032)	0.549 (0.031)	0.576 (0.028)
		MLP	0.368 (0.038)	0.557 (0.023)	0.440 (0.038)	0.556 (0.04)	0.597 (0.027)
		Gradient Boosted Classifier	0.425 (0.04)	0.570 (0.098)	0.553 (0.012)	0.641 (0.032)	0.676 (0.032)
		TCN	0.456 (0.028)	0.665 (0.036)	0.608 (0.037)	0.691 (0.034)	0.722 (0.039)
	$EDSS_{mean}$ as severity category	Logistic Regression	0.302 (0.009)	0.384 (0.017)	0.423 (0.014)	0.457 (0.014)	0.470 (0.015)
		MLP	0.397 (0.013)	0.521 (0.016)	0.475 (0.019)	0.582 (0.014)	0.633 (0.018)
		Gradient Boosted Classifier	0.400 (0.010)	0.515 (0.015)	0.561 (0.031)	0.630 (0.013)	0.675 (0.016)
		TCN	0.520 (0.011)	0.572 (0.086)	0.659 (0.016)	0.686 (0.042)	0.709 (0.044)
Floodlight	Cognitive disability score	Linear Regression	0.308 (0.018)	0.305 (0.008)	0.322 (0.021)	0.303 (0.016)	0.285 (0.015)
		Gradient Boosted Regressor	0.306 (0.017)	0.286 (0.012)	0.312 (0.019)	0.283 (0.014)	0.275 (0.014)
		TCN	0.306 (0.017)	0.286 (0.012)	0.416 (0.037)	0.283 (0.014)	0.275 (0.014)
	Dexterity disability score	Linear Regression	0.161 (0.015)	0.163 (0.011)	0.165 (0.012)	0.161 (0.012)	0.153 (0.012)
		Gradient Boosted Regressor	0.159 (0.014)	0.153 (0.012)	0.163 (0.013)	0.152 (0.012)	0.148 (0.012)
		TCN	0.159 (0.014)	0.153 (0.012)	0.198 (0.021)	0.152 (0.012)	0.148 (0.012)
	Mobility disability score	Linear Regression	0.283 (0.016)	0.269 (0.017)	0.298 (0.019)	0.268 (0.021)	0.256 (0.018)
		Gradient Boosted Regressor	0.278 (0.018)	0.249 (0.017)	0.292 (0.018)	0.248 (0.019)	0.244 (0.021)
		TCN	0.278 (0.018)	0.249 (0.017)	0.381 (0.031)	0.248 (0.019)	0.244 (0.021)
	Overall disability score	Linear Regression	0.224 (0.014)	0.222 (0.008)	0.237 (0.017)	0.220 (0.014)	0.206 (0.012)
		Gradient Boosted Regressor	0.220 (0.012)	0.206 (0.012)	0.230 (0.016)	0.205 (0.013)	0.197 (0.013)
		TCN	0.220 (0.018)	0.206 (0.012)	0.308 (0.034)	0.205 (0.013)	0.197 (0.013)

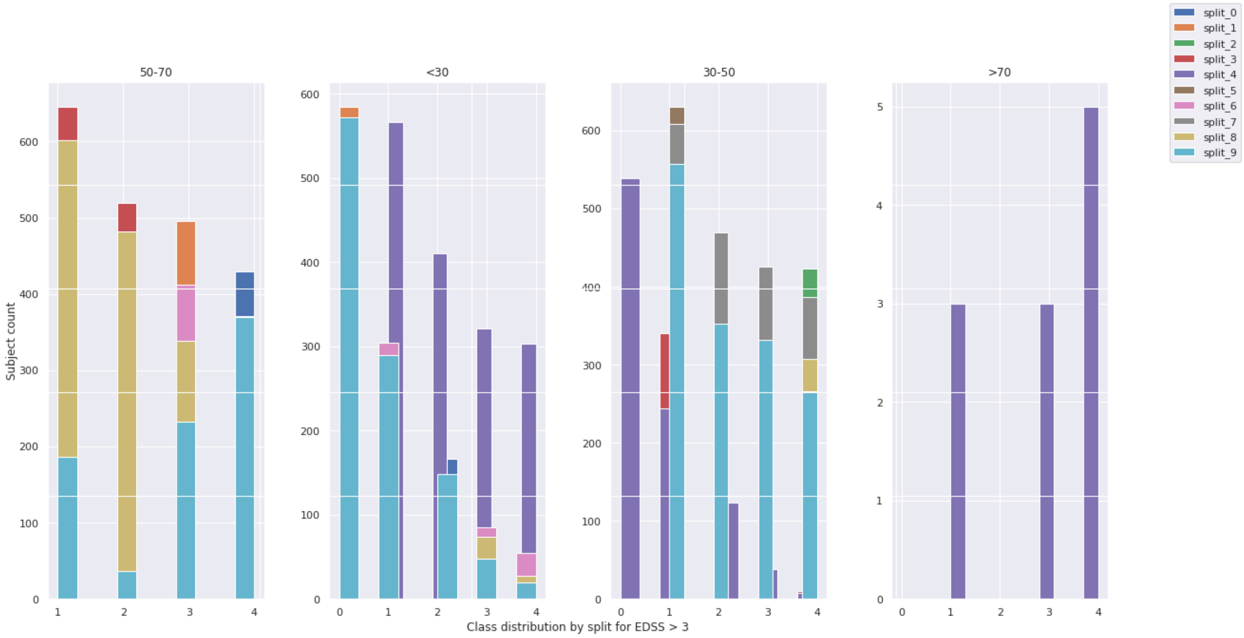

Figure 11: Class distribution by split for EDSS as severity category.

Table 11: Subgroup results for prediction tasks in MSOAC on the 6-12 months horizon.

Tasks	**Models**	**Female**	**Male**	**Age <30**	**Age 30-50**	**Age 50-70**	**Age >70**
Instance count	-	23028	11604	3643	17151	7489	11
$EDSS_{mean} > 3$	Logistic Regression	0.71 (0.027)	0.72 (0.016)	0.63 (0.055)	0.65 (0.017)	0.77 (0.022)	1.0 (0.0)
	MLP	0.74 (0.024)	0.72 (0.027)	0.70 (0.090)	0.67 (0.029)	0.71 (0.032)	0.90 (0.0)
	Gradient Boosted Classifier	0.76 (0.029)	0.75 (0.025)	0.76 (0.068)	0.71 (0.023)	0.78 (0.031)	1.0 (0.0)
	TCN	0.81 (0.032)	0.84 (0.034)	0.61 (0.171)	0.78 (0.048)	0.88 (0.033)	0.81 (0.0)
$EDSS_{mean} > 5$	Logistic Regression	0.56 (0.035)	0.60 (0.024)	0.41 (0.095)	0.52 (0.043)	0.63 (0.045)	NaN (NaN)
	MLP	0.60 (0.053)	0.58 (0.040)	0.57 (0.0125)	0.54 (0.028)	0.55 (0.049)	NaN (NaN)
	Gradient Boosted Classifier	0.68 (0.043)	0.67 (0.058)	0.66 (0.116)	0.64 (0.042)	0.71 (0.045)	NaN (NaN)
	TCN	0.72 (0.061)	0.72 (0.041)	0.43 (0.324)	0.68 (0.057)	0.78 (0.051)	0.58 (0.0)
$EDSS_{mean}$ as severity category	Logistic Regression	0.47 (0.018)	0.48 (0.023)	0.48 (0.033)	0.48 (0.027)	0.48 (0.028)	0.70 (0.0)
	MLP	0.63 (0.019)	0.63 (0.028)	0.61 (0.045)	0.60 (0.019)	0.58 (0.028)	0.67 (0.0)
	Gradient Boosted Classifier	0 (0.0)	0 (0.0)	0.69 (0.039)	0.60 (0.026)	0.62 (0.031)	0.62 (0.0)
	TCN	0.71 (0.050)	0.71 (0.041)	0.57 (0.085)	0.62 (0.056)	0.60 (0.065)	0.52 (0.0)