

Early Prediction of Multiple Sclerosis Disability Progression via Multimodal Foundation Model Benchmarks

Maxime Usdin¹, Lito Kriara², Licinio Craveiro²

¹Genentech, Inc, South San Francisco, CA, USA

²F. Hoffmann-La Roche Ltd., Basel, Switzerland

usdin.maxime@gene.com, {lito.kriara, licinio.craveiro}@roche.com

Abstract

Early multiple sclerosis (MS) disability progression prediction is challenging due to disease heterogeneity. This work predicts 48- and 72-week disability using sparse baseline clinical data and 12 weeks of daily digital Floodlight data from the CONSONANCE clinical trial. We employed state-of-the-art tabular and time-series foundation models (FMs), a custom multimodal attention-based transformer, and machine learning methods. Despite the difficulty of early prediction (AUROC 0.63), integrating digital data via advanced models improved performance over clinical data alone. A transformer model using unimodal embeddings from the Moment FM yielded the best result, but our multimodal transformer consistently outperformed its unimodal counterpart, confirming the advantages of combining clinical with digital data. Our findings demonstrate the promise of FMs and multimodal approaches to extract predictive signals from complex and diverse clinical and digital life sciences data (e.g., imaging, omics), enabling more accurate prognostics for MS and potentially other complex diseases.

1 Introduction

Early prediction of multiple sclerosis (MS) disability progression is vital for personalized care, but is hindered by measurement limitations and disease heterogeneity [Oh *et al.*, 2024]. Numerous studies highlight this prognostic challenge, attempting to predict risk using diverse machine learning approaches on varied data (from baseline clinical, imaging and transcriptomic data [Eshaghi *et al.*, 2022; Yousef *et al.*, 2024; Gurevich *et al.*, 2025], to longitudinal clinical registries [De Brouwer *et al.*, 2024] and explorations of digital measure correlates [Graves *et al.*, 2023a]). Despite these varied efforts and the potential of advanced computational methods, achieving high predictive performance for early MS disability progression remains an unmet clinical and research need. These efforts, utilizing inputs ranging from purely baseline assessments to extensive clinical histories, underscore the difficulty of early prediction despite the potential of advanced computational methods.

Foundation models (FMs) show promise in single modalities [Bommasani *et al.*, 2022], and Large Language Models (LLMs) are being applied to diverse data types. However, few studies have systematically developed unified architectures for MS progression that combine sparse clinical data with dense, potentially noisy, digital biomarkers across different time scales [Dillenseger *et al.*, 2021; Simon *et al.*, 2024; Wenk *et al.*, 2024]. Our approach conceptually extends pre-trained multimodal large language models (LLMs) by treating sparse clinical assessments, dense digital biomarker streams, and other relevant structured data as varied tokens within a unified sequence, adaptable to sequence modeling.

This work addresses these gaps by comprehensively assessing early MS disability prediction using three months of observational data. We evaluated tabular and time-series FMs, and our novel multimodal transformer architecture (which fuses sparse clinical features with dense Floodlight time-series data inspired by the canonical transformer model [Vaswani *et al.*, 2017]), against state-of-the-art methods. Our findings reveal the potential of foundation models and multimodal data integration for early prediction, guiding the development of future clinical multimodal models.

2 Methods

2.1 Data

Data were sourced from 811 participants enrolled in the CONSONANCE Phase 3b clinical trial (NCT03523858) [Comi *et al.*, 2022]. This dataset uniquely combines infrequent clinical assessments with frequent, patient-generated digital biomarker data collected via the Floodlight smartphone technology [Montalban *et al.*, 2022]. Clinical data included baseline demographics, disease characteristics, and longitudinal assessments of neurological function and performance such as the Expanded Disability Status Scale (EDSS), Timed 25-Foot Walk (T25FWT), 9-Hole Peg Test (9HPT), and other standard clinical outcomes, typically collected every 24 weeks [Kurtzke, 1983; Cutter *et al.*, 1999]. Digital biomarker data consisted of CONSONANCE’s initial 12 weeks of daily time-series measurements of Floodlight tests designed to measure hand function and gait (i.e., Draw a Shape [Graves *et al.*, 2023b], Pinching Test [Graves *et al.*, 2023a], U-Turn Test [Cheng *et al.*, 2021], and the 2-Minute Walk Test [Montalban *et al.*, 2022]). The target

VARIABLE	WEEK 48	WEEK 72
PARTICIPANT NUMBER WITH CLINICAL DATA, N	406	394
AGE AT BASELINE (YEARS), MEAN (SD)	48.8 (9.3)	48.7 (9.4)
FEMALE, N (%)	219 (53.9)	210 (53.3)
EDSS, MEAN (SD)	4.8 (1.4)	4.9 (1.5)
T25FWT, MEAN (SD)	16.2 (26.3)	20.5 (35.1)
9HPT, MEAN (SD)	35.9 (29.1)	36.1 (27.3)
CCDP EVENTS, N (%)	108 (24)	155 (35)

Table 1: Demographic information of the CONSONANCE dataset.

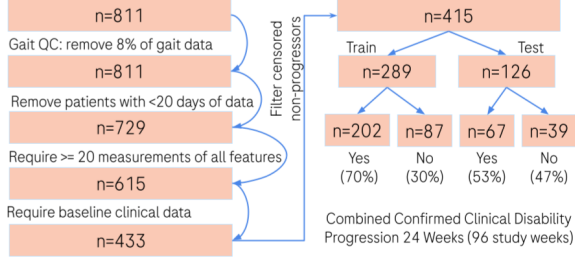


Figure 1: Study population selection. Gait features were filtered using pre-determined quality control flags in the Floodlight app.

variable for prediction was 24-Week composite Confirmed Disability Progression (24W cCDP), a standard composite endpoint in MS trials reflecting clinically meaningful worsening. The distribution of patient data used is summarized in Table 1. A complete list of the features used is shown in Table 2.

2.2 Preprocessing

Clinical data were one-hot encoding for categorical variables and standard scaling was applied to numerical features. Rigorous cohort filtering was performed to ensure baseline data availability and sufficient longitudinal density to model progression dynamics, resulting in the final analysis cohort of 415 out of originally 811 study participants (fig. 1). A fixed 80/20% train/test split was established based on a pre-defined data cut-off date to ensure temporal validity, preventing data leakage into the training set.

2.3 Feature Extraction and Gaussian Process Interpolation

To capture temporal dynamics from the Floodlight time series within the 12-week input window, feature engineering approaches were employed, including calculating windowed statistics (mean, variance, trend coefficients) and utilizing the *tsfresh* [Christ *et al.*, 2018] library for automated extraction of a comprehensive set of time-series characteristics. Feature selection methods, including ANOVA F-test and Recursive Feature Elimination (RFE), were used to reduce dimensionality and select informative features.

To address missing data within the dense but potentially incomplete daily Floodlight measurements, we employed patient-specific Gaussian Processes (GPs). GPs offer a robust probabilistic framework for interpolating and modeling temporal trajectories from such

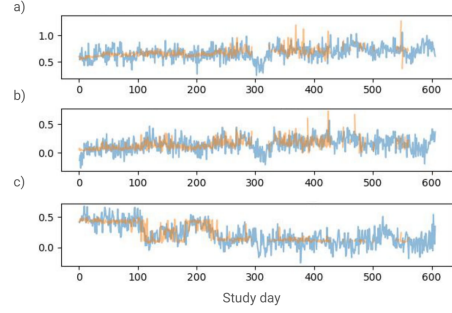


Figure 2: Illustrative examples of Floodlight digital biomarker time-series data from a single patient, augmented using patient-specific Gaussian Processes (GPs). The figure displays three trajectories over the study duration: (a) step duration (median), (b) step duration (interquartile range, IQR), and (c) step length (median). Each panel shows the original, more variable observed data points (orange) for the respective metric, alongside the smoother mean of the fitted GP model (blue).

sparse and noisy data. We fitted these patient-specific GPs to model the evolution of individual Floodlight features. The resulting models were then used to generate complete trajectories for downstream model training: observed Floodlight measurements were preserved, while any missing time-points in these original trajectories, along with entirely synthetic trajectories for data augmentation, were instantiated by sampling from the posterior predictive distribution of the fitted GPs. Illustrative examples of observed data points and their corresponding GP-imputed trajectories are provided in fig. 2.

To implement the GP methodology, we utilized a composite kernel combining a Radial Basis Function (RBF) for smoothness and a White Noise kernel to account for measurement noise. The covariance between two time points t_i and t_j is given by:

$$k(t_i, t_j) = \sigma_c^2 \exp\left(-\frac{(t_i - t_j)^2}{2l^2}\right) + \sigma_n^2 \delta_{ij}$$

where δ_{ij} is the Kronecker delta, which is 1 if $i = j$ and 0 otherwise. The hyperparameters $\theta = \{\sigma_c^2, l, \sigma_n^2\}$ represent output variance, characteristic length-scale, and noise level, respectively.

Prior to optimization, initial values and bounds for θ were estimated heuristically for each patient-feature pair based on data characteristics (overall feature variance for σ_c^2 , time intervals for l , and linear fit residuals for σ_n^2). Within these bounds, the hyperparameters were then optimized by maximizing the log marginal likelihood:

$$\log p(\mathbf{y}|\mathbf{t}, \theta) = -\frac{1}{2} \mathbf{y}^T K^{-1} \mathbf{y} - \frac{1}{2} \log |K| - \frac{n}{2} \log 2\pi$$

Here, \mathbf{y} is the vector of observed values, \mathbf{t} is the vector of corresponding time points, and K is the covariance matrix where $K_{ij} = k(t_i, t_j)$. The posterior predictive distribution at a new time point t_* given observed data (t, y) is:

$$p(f_*|t_*, \mathbf{t}, \mathbf{y}, \theta) = \mathcal{N}(\mathbf{k}_*^T K^{-1} \mathbf{y}, k(t_*, t_*) - \mathbf{k}_*^T K^{-1} \mathbf{k}_*)$$

where \mathbf{k}_* is the vector of covariances between the new time point t_* and the observed time points \mathbf{t} , i.e., $\mathbf{k}_* = [k(t_*, t_1), k(t_*, t_2), \dots, k(t_*, t_n)]^T$. This posterior distribution allows us to sample synthetic data f_* conditioned on our observed data and optimized hyperparameters θ .

2.4 Model Architectures

A range of unimodal and multimodal models was evaluated:

Machine Learning: AutoGluon [Erickson *et al.*, 2020], an automated machine learning framework, was used to establish robust baselines on tabular feature sets derived from clinical data and/or aggregated longitudinal Floodlight features. We also trained XGBoost, Logistic Regression, and Random Forest models as additional competitive baselines.

Time-Series Modeling: Clinical and Floodlight time-series features data were concatenated and fed into a comprehensive suite of state-of-the-art time-series classifiers from the *sktime* library [Löning *et al.*, 2019]. We utilized a diverse set of time-series classification models, encompassing kernel-based (RocketClassifier), interval-based (CanonicalIntervalForest, DrCIF), distance-based (KNeighborsTimeSeriesClassifier), hybrid (HIVECOTEV2), deep learning (MVTSTransformerClassifier, CNNClassifier, ResNetClassifier), dictionary-based (MUSE), and feature-based (FreshPRINCE) approaches, with model selection facilitated by grid search cross-validation on the training set.

Foundation Model Approaches (Unimodal):

- TabPFN [Hollmann *et al.*, 2025]: A transformer-based FM optimized for small tabular datasets, applied here to clinical data combined with engineered Floodlight features. Its inclusion aims to test the applicability of tabular FMs to this mixed-feature setting.
- Moment+Transformer: Precomputed time-series embeddings were generated using the publicly available MOMENT FM [Goswami *et al.*, 2024]. The embeddings, capturing rich temporal patterns from the Floodlight data, were fed as input sequences into a standard transformer encoder architecture for classification.
- GP-Transformer: A standard transformer encoder architecture applied directly to the GP-augmented Floodlight time-series data.

Custom Multimodal Transformer: An attention-based multimodal transformer model (fig. 3) was specifically engineered to seamlessly integrate heterogeneous inputs. It takes in three distinct data streams: Gaussian-Process augmented Floodlight data, projected static/clinical features, and a classification (CLS) token. These are concatenated along the sequence dimension, enhanced with positional embeddings, and then processed through standard transformer blocks. To ensure compatibility, the static features are projected to match the time-series embedding dimension using a feed-forward network. Finally, the CLS token’s representation, which aggregates information through self-attention across the entire fused

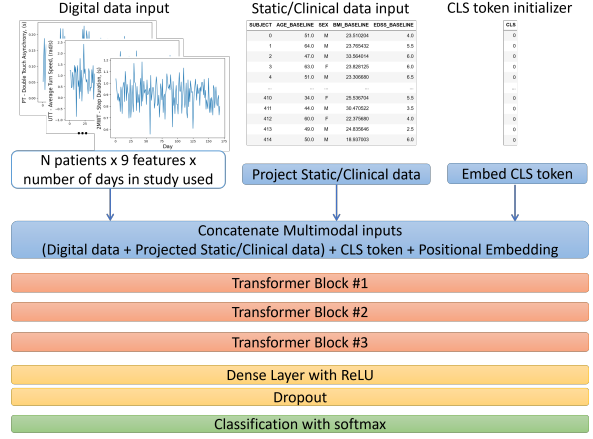


Figure 3: Schematic of the custom multimodal transformer, illustrating the integration of GP-processed time series, projected static/clinical features, and a CLS token prior to processing through transformer blocks and an MLP classification head.

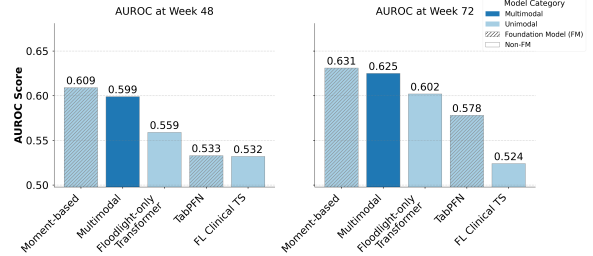


Figure 4: Model performance (AUROC) for predicting 48-week (left) and 72-week (right) cCDP using 12 weeks of input data with Moment-based, embeddings derived from the Moment time-series foundation model; Multimodal, our multimodal transformer; Floodlight-only transformer, our transformer model trained on Floodlight data only; TabPFN, a tabular foundation model trained on clinical and derived Floodlight features; and FL Clinical TS, which denotes the optimal machine learning model performance achieved on clinical and derived Floodlight features.

sequence, is passed through final MLP layers for classification.

2.5 Training and Evaluation

Models were trained on the 80% training set, with 5-fold CV for hyperparameter optimization for machine learning models. Final model performance was evaluated on the held-out 20% test set (fig. 1) using the AUROC as the primary metric, which is well-suited for imbalanced classification tasks such as early disability progression prediction.

3 Results

Our evaluation focused on the clinically relevant task of predicting 48-week (T48) and 72-week (T72) cCDP using only the initial 12 weeks of data. Overall performance was modest, reflecting the difficulty of long-term prediction from early data (see fig. 4 for T72 results; T48 trends were similar). The highest AUROC achieved with 12 weeks of input was 0.63 (for T72)

and 0.61 (for T48), obtained by a transformer using unimodal Moment embeddings, showcasing the potential of FM-derived representations in Floodlight data.

Despite the challenge of early prediction, we observed encouraging and consistent relative performance differences (fig. 4). A benefit from multimodality was seen in the transformer architecture: for T72, the multimodal transformer (AUROC 0.63) improved upon a comparable unimodal transformer using only GP-processed Floodlight data (AUROC 0.60), a relative gain suggesting value in combining data sources. A similar advantage was seen for T48 (Multimodal 0.60 vs Unimodal 0.56).

FMs demonstrated clear utility over standard machine and deep learning approaches in our disability progression prediction task. For T72, the transformer on Moment embeddings (Unimodal FM, AUROC 0.63) and TabPFN (Unimodal FM, AUROC 0.58) generally outperformed simpler non-FM baselines such as the best sktime ensemble (AUROC 0.52) and standard ML benchmarks (AUROC 0.55). Similar trends held for T48. The evaluated fusion strategies within the custom transformer, however, did not yield substantial gains over the best unimodal FM approach.

4 Discussion

This study demonstrates the potential of foundation models (FMs) and multimodal strategies for the challenging early prediction of MS disability progression. Our modest overall AUROC (0.61-0.63) using 12 weeks of input underscores the difficulty of long-term forecasting from limited early data, yet relative performance differences highlight the value of advanced modeling.

Our findings show FM-derived representations, particularly Moment embeddings with a transformer, effectively capture predictive signals from noisy digital biomarkers, outperforming other FMs on combined features and standard time-series methods. This underscores the importance of tailoring foundation models to the unique properties of different data modalities for effective signal extraction.

Integrating multimodal data also proved beneficial, though challenging. Our custom multimodal transformer (clinical + digital) outperformed its unimodal digital-only counterpart, demonstrating that clinical context enhances predictive signal even when early signals are weak. However, this multimodal approach did not surpass the best-performing unimodal FM (Moment + Transformer). This suggests that the primary difficulty lies not just in the fusion architecture but in the inherent signal-to-noise ratio and heterogeneity of early data, where sparse clinical signals may be underdeveloped and dense digital signals noisy.

From a clinical perspective, achieving an AUROC of 0.63 with only 12 weeks of data, is a promising step for earlier risk stratification. Although not yet enabling high-confidence individual predictions, this benchmark demonstrates the utility of FMs and multimodal learning to improve early risk identification compared to traditional approaches. Continued advancements in robust representation learning

from noisy digital data and self-supervised pre-training could further enhance clinical utility.

Related work has shown similar performance for similar tasks, with Mostafa *et al.*, reporting an AUC of 0.63 to 0.67 for predicting only upper limb disability progression in MS [Mostafa *et al.*, 2021]. Other works that report higher performance in predicting future disability employ more detailed and harder to retrieve data like imaging (MRI and PET scans) [Pinto *et al.*, 2020], cerebrospinal fluid and blood biomarkers [Comi *et al.*, 2024], electronic health records in addition to clinical markers [De Brouwer *et al.*, 2024], or passive monitoring data (e.g., sleep quality, fatigue) [Chikersal *et al.*, 2022].

Limitations include reliance on the single CONSONANCE trial, a sample size (N=415) that may constrain highly complex models, and limited modalities (lacking imaging or genomics). Future research may prioritize validation on larger, multi-center datasets, incorporate broader biological modalities, and continue developing robust fusion techniques. Exploring self-supervised pre-training on larger mixed clinical/digital health data repositories before task-specific fine-tuning is a key direction for advancing multimodal FMs in life sciences. The benchmarking and findings presented have implications beyond MS. Integrating sparse clinical assessments with dense, noisy sensor data is a common challenge in managing chronic diseases via remote patient monitoring. Thus, our insights on FM and multimodal architecture utility could generalize, improving early risk detection and personalized interventions across various medical conditions.

5 Conclusion

In this study, we addressed the challenging task of predicting MS disability progression using only 12 weeks of multimodal data. We demonstrate that foundation model embeddings derived from digital biomarkers (Moment embeddings and a transformer model) yield the best predictive performance (AUROC 0.63) among the tested methods, outperforming standard ML and time-series approaches. Furthermore, integrating clinical context via a custom multimodal transformer architecture consistently improved upon its unimodal counterpart, confirming the value of multimodal data even in this difficult early prediction setting. While the inherent limitations of early, heterogeneous data capped overall performance and prevented simple fusion from surpassing the best unimodal FM, our findings clearly show that FMs and multimodal approaches, with potential future extensions towards LLM-based architectures (by treating disparate data types as tokens in a unified sequence, mirroring LLM tokenization strategies), can extract valuable predictive signals beyond traditional methods. These results underscore the potential of applying advanced models to complex clinical prediction tasks and provide a foundation for future research focused on enhancing signal extraction, representation learning, and fusion techniques for challenging real-world health data, ultimately moving towards more personalized and timely patient care.

Supplemental Material

Data and Feature Overview

Table 2: Data and Feature Overview

Feature	Category	Description
AGE	Demographic	Patient's age.
BBMI	Demographic	Baseline Body Mass Index (BMI).
BLSDMT	Clinical History	Baseline Symbol Digit Modalities Test (SDMT) score [Smith, 1973].
Average Time Taken for 25-Foot Walk Test	Standard Clinical Assessment	Average time taken to complete a 25-foot walk, assessing ambulation [Cutter <i>et al.</i> , 1999].
Average of 9 Hole Peg Tests	Standard Clinical Assessment	Average time to complete the 9-Hole Peg Test, assessing manual dexterity [Mathiowetz <i>et al.</i> , 1985].
Ambulation Score	Standard Clinical Assessment	Assesses patient's walking ability [Hauser <i>et al.</i> , 1983].
Bowel/Bladder Functional System Score	Standard Clinical Assessment	Assesses functional impairment in bowel and bladder systems [Kurtzke, 1983].
Brainstem Functional System Score	Standard Clinical Assessment	Assesses functional impairment in the brainstem [Kurtzke, 1983].
Cerebellar Functional System Score	Standard Clinical Assessment	Assesses functional impairment in the cerebellar system [Kurtzke, 1983].
Cerebral Functional System Score	Standard Clinical Assessment	Assesses functional impairment in the cerebral system [Kurtzke, 1983].
Expanded Disability Score (EDSS)	Standard Clinical Assessment	Standard measure of MS disability [Kurtzke, 1983].
NUMRLP	Clinical History	Number of documented relapses.
ONSETYRS	Clinical History	Years since the onset of MS symptoms.
Pyramidal Functional System Score	Standard Clinical Assessment	Assesses functional impairment in the pyramidal system [Kurtzke, 1983].
SEX	Demographic	Patient's sex.
Sensory Functional System Score	Standard Clinical Assessment	Assesses functional impairment in the sensory system [Kurtzke, 1983].
Visual Functional System Score	Standard Clinical Assessment	Assesses functional impairment in the visual system [Kurtzke, 1983].
Max 9HPT	Derived Clinical	Maximum score from the 9-Hole Peg Test.
Min 9HPT	Derived Clinical	Minimum score from the 9-Hole Peg Test.
Range 9HPT	Derived Clinical	Range of scores from the 9-Hole Peg Test.
Cerebellar White Matter Volume	Neuroimaging	Volume of white matter in the cerebellum from MRI.
Cerebral White Matter Volume	Neuroimaging	Volume of white matter in the cerebrum from MRI.
T2 Lesion Volume	Neuroimaging	Total volume of T2 hyperintense lesions from MRI.
Thalamic Volume	Neuroimaging	Volume of the thalamus from MRI.
Median step duration	Floodlight (Gait)	Median value of step duration from digital gait measurements [Rodrigues <i>et al.</i> , 2024].
Median step impulse	Floodlight (Gait)	Median value of step impulse/intensity from digital gait measurements [Rinderknecht <i>et al.</i> , 2023].
Median step length	Floodlight (Gait)	Median value of step length from digital gait measurements [Rodrigues <i>et al.</i> , 2024].
Sum of step length	Floodlight (Gait)	Sum of step lengths from digital gait measurements [Rodrigues <i>et al.</i> , 2024].
Median step velocity	Floodlight (Gait)	Median value of step velocity from digital gait measurements [Rodrigues <i>et al.</i> , 2024].
Median turn speed	Floodlight (Gait)	Median absolute turn speed from digital gait measurements [Cheng <i>et al.</i> , 2021].
Trace accuracy for DaS FIGURE 8 shape	Floodlight (Upper limb)	Aggregated tracing accuracy across attempts for the Figure-8 drawing task [Graves <i>et al.</i> , 2023b].
Double touching asynchrony in Pinching Test	Floodlight (Upper limb)	Mean time between the first and second finger touch during a pinching test [Graves <i>et al.</i> , 2023a].
Total number of pinches	Floodlight (Upper limb)	Total number of pinching attempts during a pinching test [Graves <i>et al.</i> , 2023a].

References

- [Bommasani *et al.*, 2022] Rishi Bommasani, Drew A. Hudson, and et al. On the Opportunities and Risks of Foundation Models, 2022.
- [Cheng *et al.*, 2021] Wei-Yi Cheng, Alan Bourke, Florian Lipsmeier, Corrado Bernasconi, Solomon Belachew, Caroline Gossens, Jennifer S Graves, Xavier Montalban, and Michael Lindemann. U-turn speed is a valid and reliable smartphone-based measure of multiple sclerosis-related gait and balance impairment. *Gait & Posture*, 84:120–127, 2021.
- [Chikersal *et al.*, 2022] Prerna Chikersal, Shruthi Venkatesh, Karman Masown, Elizabeth Walker, Danyal Quraishi, Anind Dey, Mayank Goel, and Zongqi Xia. Predicting multiple sclerosis outcomes during the covid-19 stay-at-home period: Observational study using passively sensed behaviors and digital phenotyping. *JMIR Ment Health*, 9(8), Aug 2022.
- [Christ *et al.*, 2018] Maximilian Christ, Nils Braun, Julius Neuffer, and Andreas W. Kempa-Liehr. Time series feature extraction on basis of scalable hypothesis tests (tsfresh – a python package). *Neurocomputing*, 307:72–77, 2018.
- [Comi *et al.*, 2022] Giancarlo Comi, Robert A Bermel, Amit Bar-Or, Marisa P McGinley, Douglas L Arnold, Roland G Henry, Ralph HB Benedict, Pavan Bhargava, Helmut Butzkueven, Declan T Chard, Guy Gherardi, Owain W Howell, Christine Lebrun-Frenay, Letizia Leocani, Agne Kazlauskaitė, Thomas Künzel, Xavier Montalban, Maria Pia Sormani, and Finn Sellebjerg. A multicentre, open label, single-arm, phase 3b study (consonance) to assess the effectiveness and safety of ocrelizumab in patients with primary and secondary progressive multiple sclerosis: year-1 interim analysis (p1-1.virtual). *Neurology*, 98, 2022.
- [Comi *et al.*, 2024] Giancarlo Comi, Gloria Dalla Costa, Bruno Stankoff, Hans-Peter Hartung, Per Soelberg Sørensen, Patrick Vermersch, and Letizia Leocani. Assessing disease progression and treatment response in progressive multiple sclerosis. *Nature Reviews Neurology*, 20:573–586, 2024.
- [Cutter *et al.*, 1999] Gary R. Cutter, Monika L. Baier, Richard A. Rudick, Diane L. Cookfair, Jill S. Fischer, John Petkau, Karl Syndulko, Brian G. Weinshenker, Jack P. Antel, Christian Confavreux, George W. Ellison, Fred Lublin, Aaron E. Miller, Stephen M. Rao, Stephen Reingold, Alan Thompson, and Ernest Willoughby. Development of a multiple sclerosis functional composite as a clinical trial outcome measure. *Brain*, 122(5):871–882, May 1999.
- [De Brouwer *et al.*, 2024] Edward De Brouwer, Thijs Becker, Lorin Werthen-Brabants, Pieter Dewulf, Dimitrios Iliadis, Cathérine Dekeyser, Guy Laureys, Bart Van Wijmeersch, Veronica Popescu, Tom Dhaene, Dirk Deschrijver, Willem Waegeman, Bernard De Baets, Michiel Stock, Dana Horakova, Francesco Patti, Guillermo Izquierdo, Sara Eichau, Marc Girard, Alexandre Prat, Alessandra Lugaresi, Pierre Grammond, Tomas Kalincik, Raed Alroughani, Francois Grand’Maison, Olga Skibina, Murat Terzi, Jeannette Lechner-Scott, Oliver Gerlach, Samia J. Khoury, Elisabetta Cartechini, Vincent Van Pesch, Maria José Sà, Bianca Weinstock-Guttman, Yolanda Blanco, Radek Ampapa, Daniele Spitaleri, Claudio Solaro, Davide Maimone, Aysun Soysal, Gerardo Iuliano, Riadh Gouider, Tamara Castillo-Triviño, José Luis Sánchez-Menoyo, Guy Laureys, Anneke van der Walt, Jiwon Oh, Eduardo Aguera-Morales, Ayse Altintas, Abdullah Al-Asmi, Koen de Gans, Yara Fragoso, Tunde Csepány, Suzanne Hodgkinson, Norma Deri, Talal Al-Harbi, Bruce Taylor, Orla Gray, Patrice Lalive, Csilla Rozsa, Chris McGuigan, Allan Kermode, Angel Pérez Sempere, Simu Mihaela, Magdolna Simo, Todd Hardy, Danny Decoo, Stella Hughes, Nikolaos Grigoriadis, Attila Sas, Norbert Vella, Yves Moreau, and Liesbet Peeters. Machine-learning-based prediction of disability progression in multiple sclerosis: An observational, international, multi-center study. *PLOS Digital Health*, 3(7):1–25, July 2024.
- [Dillenseger *et al.*, 2021] Anja Dillenseger, Marie Luise Weidemann, Katrin Trentzsch, Hernan Inojosa, Rocco Haase, Dirk Schriefer, Isabel Voigt, Maria Scholz, Katja Akgün, and Tjalf Ziemssen. Digital Biomarkers in Multiple Sclerosis. *Brain Sciences*, 11(11), November 2021.
- [Erickson *et al.*, 2020] Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data, 2020.
- [Eshaghi *et al.*, 2022] Arman Eshaghi, Peter A Wijeratne, Neil P Oxtoby, Douglas L Arnold, Louis Collins, Sridar Narayanan, Charles R. G. Guttman, Alan J Thompson, Daniel C Alexander, Frederik Barkhof, Declan Chard, and Olga Ciccarelli. Predicting personalised risk of disability worsening in multiple sclerosis with machine learning. *medRxiv*, January 2022.
- [Goswami *et al.*, 2024] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.
- [Graves *et al.*, 2023a] Jennifer S. Graves, Marcín Elantkowski, Yan-Ping Zhang, Frank Dondelinger, Florian Lipsmeier, Corrado Bernasconi, Xavier Montalban, Luciana Midaglia, and Michael Lindemann. Assessment of Upper Extremity Function in Multiple Sclerosis: Feasibility of a Digital Pinching Test. *JMIR formative research*, 7, October 2023.
- [Graves *et al.*, 2023b] Jennifer S. Graves, Marco Ganzetti, Frank Dondelinger, Florian Lipsmeier, Shibeshih Belachew, Corrado Bernasconi, Xavier Montalban, Johan van Beek, Michael Baker, Christian Gossens, and Michael Lindemann. Preliminary

- validity of the Draw a Shape Test for upper extremity assessment in multiple sclerosis. *Annals of Clinical and Translational Neurology*, 10(2):166–180, February 2023.
- [Gurevich *et al.*, 2025] Michael Gurevich, Rina Zilkha-Falb, Jia Sherman, Maxime Usdin, Catarina Raposo, Licinio Craveiro, Polina Sonis, David Magalashvili, Shay Menascu, Mark Dolev, and Anat Achiron. Machine learning-based prediction of disease progression in primary progressive multiple sclerosis. *Brain communications*, 7(1), 2025.
- [Hauser *et al.*, 1983] S L Hauser, D M Dawson, J R Leich, M F Beal, S Schuffler, W J Koroshetz, and H L Weiner. The ambulation index in multiple sclerosis. a new measure of disability. *Neurology*, 33(11):1504–1507, 1983.
- [Hollmann *et al.*, 2025] Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeyer, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.
- [Kurtzke, 1983] John F Kurtzke. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (edss). *Neurology*, 33(11):1444–1452, 1983.
- [Löning *et al.*, 2019] Markus Löning, Anthony Bagnall, Sajaysurya Ganesh, Viktor Kazakov, Jason Lines, and Franz J. Király. sktime: A Unified Interface for Machine Learning with Time Series. *arXiv*, 2019.
- [Mathiowetz *et al.*, 1985] Virgil Mathiowetz, Karen Weber, Nancy Kashman, and Gloria Volland. Adult norms for the nine hole peg test of finger dexterity. *Occupational Therapy Journal of Research*, 5(1):24–38, 1985.
- [Montalban *et al.*, 2022] Xavier Montalban, Jennifer Graves, Luciana Midaglia, Patricia Mulero, Laura Julian, Michael Baker, Jan Schadrack, Christian Gossens, Marco Ganzetti, Alf Scotland, Florian Lipsmeier, Johan van Beek, Corrado Bernasconi, Shibeshih Belachew, Michael Lindemann, and Stephen L. Hauser. A smartphone sensor-based digital outcome assessment of multiple sclerosis. *Multiple Sclerosis Journal*, 28(4):654–664, April 2022.
- [Mostafa *et al.*, 2021] Sally Mostafa, Isabel Hyo Jung Song, Ahmed A. Metwally, Nicolas Strauli, Nehemiah Sewde, Michel Friesenhahn, Maxime Usdin, and Xiaoming Jia. Predicting upper limb disability progression in primary progressive multiple sclerosis using machine learning and statistical methods. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2021.
- [Oh *et al.*, 2024] Jiwon Oh, Luca Capezzuto, Lito Kriara, Jens Schjodt-Eriksen, Johan van Beek, Corrado Bernasconi, Xavier Montalban, Helmut Butzkueven, Ludwig Kappos, Gavin Giovannoni, Riley Bove, Laura Julian, Mike Baker, Christian Gossens, and Michael Lindemann. Use of smartphone-based remote assessments of multiple sclerosis in Floodlight Open, a global, prospective, open-access study. *Scientific Reports*, 14(1), January 2024.
- [Pinto *et al.*, 2020] 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):21038, December 2020.
- [Rinderknecht *et al.*, 2023] Mike D Rinderknecht, Mattia Zanon, Tjitske A Boonstra, Lorenza Angelini, Dimitar Stanev, Gabriela González Chan, Lisa Bunn, Frank Dondelinger, Richard Hosking, Jenny Freeman, Jeremy Hobart, Jonathan Marsden, and Licinio Craveiro. An observational study to assess validity and reliability of smartphone sensor-based gait and balance assessments in multiple sclerosis: Floodlight gaitlab protocol. *Digital Health*, 9, 2023.
- [Rodrigues *et al.*, 2024] J Rodrigues, G Bogaarts, A Festanti, C. Simillion, S Hubeaux, VP Illiano, L Leocani, M McGinley, MP Sormani, G Comi, A Kazlauskaitė, L Craveiro, MD Rinderknecht, M Zanon, and H Butzkueven. Smartphone-based assessment of gait with floodlight in multiple sclerosis shows sensitivity to change over a follow-up period of 3 years. *Multiple Sclerosis Journal (ECTRIMS 2024 – ePoster)*, 30(3-suppl), 2024.
- [Simon *et al.*, 2024] Benjamin D. Simon, Kutsev Bengisu Ozyoruk, David G. Gelikman, Stephanie A. Harmon, and Barış Türkbey. The future of multimodal artificial intelligence models for integrating imaging and clinical metadata: a narrative review. *Diagnostic and Interventional Radiology*, October 2024.
- [Smith, 1973] Aaron Smith. *Symbol Digit Modalities Test*. Western Psychological Services, Los Angeles, CA, 1973.
- [Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.
- [Wenk *et al.*, 2024] Judith Wenk, Isabel Voigt, Hernan Inojosa, Hannes Schlieter, and Tjalf Ziemssen. Building digital patient pathways for the management and treatment of multiple sclerosis. *Frontiers in Immunology*, 15, 2024.
- [Yousef *et al.*, 2024] Hibba Yousef, Brigitta Malagurski Tortei, and Filippo Castiglione. Predicting multiple sclerosis disease progression and outcomes with machine learning and MRI-based biomarkers: a review. *Journal of neurology*, 271(10), October 2024.