# DeMMO: Cross-Disease Longitudinal Modelling of Digital Mobility Outcomes via Multi-Task Learning

**Anonymous authors**
Paper under double-blind review

## Abstract

Digital mobility outcomes (DMOs) derived from wearable sensors characterise mobility in daily life and offer a promising means of monitoring disease progression. However, existing DMO studies have typically focused on either a single disease or a single visit. To the best of our knowledge, this work is the first to define and study the practical problem of cross-disease longitudinal DMO modelling. We argue that this problem should satisfy at least two requirements. First, the temporal progression of DMOs should be modelled within each disease, as mobility-limiting diseases evolve over time. Second, multiple mobility-limiting diseases should be modelled jointly, as different diseases affect different aspects of human mobility. To address this problem, we propose DeMMO, an interpretable framework for longitudinal, multi-disease, and multi-outcome learning. Its central technical contribution is an interpretable cross-disease and cross-outcome relation-learning mechanism that infers signed relations directly from learned longitudinal DMO–outcome mappings, thereby enabling selective information sharing across cohorts without requiring paired participants. We evaluate DeMMO on the recently released, large-scale, multicentre Mobilise-D dataset, which presents a challenging modelling setting involving longitudinal observations, multiple clinical outcomes, and four participant-disjoint cohorts representing distinct mobility-limiting diseases. Compared with eight strong structural longitudinal and deep-regression baselines, DeMMO achieves the best overall predictive performance and outperforms the baselines for most individual outcomes. Stability selection further identifies reliable longitudinal DMO patterns that can inform subsequent clinical validation and disease monitoring.

## 1 Introduction

Disease monitoring typically relies on periodic clinical assessments that may require specialised resources, depend on evaluator judgement, or be affected by recall and individual perceptions (Delgado-Ortiz et al., 2023; Deane et al., 2014; Port et al., 2021; Prince et al., 2008). Assessments performed weeks or months apart provide only intermittent snapshots and may miss gradual, fluctuating, or transient changes (Babrak et al., 2019; Coran et al., 2020). These limitations are particularly consequential in mobility-limiting diseases such as Parkinson's disease (PD), multiple sclerosis (MS), chronic obstructive pulmonary disease (COPD), and proximal femoral fracture (PFF), which impose substantial personal and socioeconomic burdens (Yang et al., 2020; Bouleau et al., 2022; Boers et al., 2025; Williamson et al., 2017). Mobility reflects disease severity, progression, treatment response, and recovery, yet a brief supervised test measures what an individual is *capable of doing* under standardised conditions rather than what they *actually do* in daily life (Rochester et al., 2020). A six-minute walk test, for example, samples only approximately $0.06\%$ of a week and cannot capture the diversity of everyday mobility.

Wearable sensing makes it possible to study this otherwise unobserved free-living behaviour. Inertial sensors record movement over extended periods, and validated analytical algorithms transform acceleration and angular-velocity signals into digital mobility outcomes (DMOs) (Goldsack et al., 2020; Cerreta et al., 2020). These outcomes describe daily-life walking in terms of amount, pattern, pace, rhythm, and variability, capturing both *how much* and *how* a person walks (Rochester et al., 2020;



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Kluge et al., 2021). Unlike a controlled assessment, free-living monitoring samples walking bouts across days, environments, and behavioural contexts and can therefore provide a more representative view of everyday mobility (Alcock et al., 2025). This ecological richness also makes the modelling problem harder: free-living DMOs reflect not only clinical capacity but also context, behaviour, day-to-day variation, and incomplete follow-up. Clinically useful models must distinguish persistent disease-related patterns from this natural heterogeneity and determine whether the relevant DMO relationships remain stable or change over time.

Current evidence does not yet resolve this problem. Most clinical DMO studies evaluate individual measures at a single visit using correlations, linear models, or group comparisons (Polhemus et al., 2021; Mikolaizak et al., 2022; Megaritis et al., 2026; Eckert et al., 2026). The emerging longitudinal literature is still disease-specific and commonly based on small samples or a few prespecified measures: studies in PD, PFF, and MS have shown that selected free-living measures can track change, but do not model the evolving joint relationship between a comprehensive DMO profile and multiple clinical outcomes (Mirelman et al., 2024; Engdal et al., 2024; Poleur et al., 2026). Longitudinal DMO modelling therefore remains limited even within individual diseases (Rábano-Suárez et al., 2025; Kirk et al., 2025). To the best of our knowledge, cross-disease longitudinal DMO modelling has not previously been defined or studied, let alone in the more challenging free-living setting. This setting must accommodate diseases that affect different aspects of mobility, clinical outcomes measured on different scales, distinct participant cohorts, and the contextual heterogeneity of daily-life behaviour.

We therefore provide the first definition of cross-disease longitudinal DMO modelling, which should satisfy at least two requirements. First, as mobility-limiting diseases evolve over time, the temporal progression of DMOs should be modelled within each disease. Second, as different mobility-limiting diseases affect different aspects of human mobility, multiple diseases should be modelled jointly to provide a more robust and comprehensive understanding of mobility impairment.

A major barrier to studying this problem has been the lack of large, harmonised longitudinal datasets that combine repeated measurements of a common DMO set with multiple clinical outcomes across diseases (Rochester et al., 2020). Earlier resources were commonly cross-sectional, disease-specific, heterogeneous in sensing and aggregation, or limited to few assessment periods, preventing systematic analysis of stable, visit-specific, and cross-disease DMO patterns. The recently released Mobilise-D dataset was developed to address this gap (Mikolaizak et al., 2022; Rochester et al., 2020; Kirk et al., 2024). It is the principal data resource produced by a major European public–private programme involving 35 partners from 13 countries. Conducted from 2019 to 2024, the programme represented a total investment of approximately EUR 49.9 million. Mobilise-D first established the technical validity of its sensor and analytical pipeline through a separate multicentre study of 108 participants spanning healthy ageing and five mobility-limiting conditions (Mikolaizak et al., 2022). The subsequent CVS extended this foundation to clinical validation at an unprecedented large scale: 2,400 participants across PD, MS, COPD, and PFF underwent a comprehensive clinical assessment and seven days of real-world mobility monitoring on five occasions, separated by six-month intervals over two years (Kirk et al., 2024). The released resource combines harmonised walking-bout, daily, and weekly DMOs from a common lower-back sensor with clinical tests, clinician-reported assessments, and patient-reported outcomes. This combination of scale, longitudinal depth, cross-disease breadth, and standardised measurement makes Mobilise-D uniquely suited to studying how multivariate DMO–outcome relationships evolve during disease progression and functional recovery.

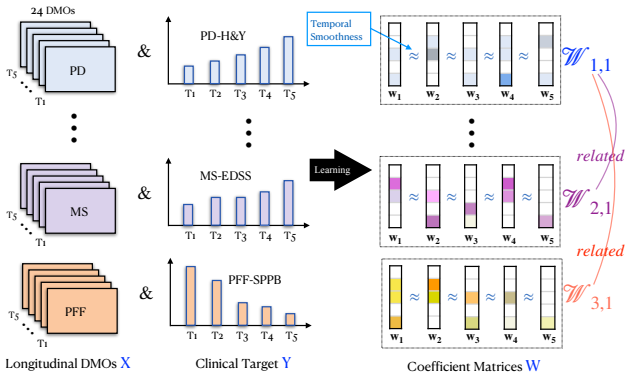


Figure 1: Overview of DeMMO. For each outcome, visit-specific coefficients map 24 longitudinal DMOs to clinical targets. DeMMO jointly learns temporally smooth mappings and selective within- and cross-disease relations.

We summarise the main contributions of this work as follows:

- **Conceptual contribution.** To the best of our knowledge, this work is the first to define cross-disease longitudinal DMO modelling. The problem establishes two requirements: modelling temporal DMO progression within each disease and jointly modelling multiple diseases to capture their complementary effects on human mobility. We further present the first systematic study of this problem in the more challenging free-living setting.
- **Technical contribution.** We expose a key limitation of existing temporal multi-task learning: current formulations do not jointly learn multiple outcomes across diseases with non-overlapping cohorts. To address this limitation, we introduce DeMMO and a novel cross-disease and cross-outcome relation-learning mechanism. As shown in Figure 1, DeMMO learns a signed relation graph directly from longitudinal DMO mappings, enabling selective knowledge transfer across unpaired cohorts while retaining outcome-specific temporal evolution and feature selection.
- **Practical contribution.** Using the recently released Mobilise-D dataset, we jointly study multiple clinical outcomes over five visits across several mobility-limiting diseases. Against eight strong baselines spanning structured longitudinal and recent deep-regression methods, DeMMO achieves the best overall predictive performance and outperforms the baselines for most individual outcomes. Longitudinal stability selection (Zhou et al., 2024; Fan et al., 2026) further identifies robust, outcome-specific DMO patterns as candidates for subsequent clinical validation and disease monitoring.

Taken together, these contributions establish cross-disease longitudinal DMO modelling as a concrete research problem and systematically investigate it in the challenging free-living setting. We emphasise that the simplicity, concreteness, and high interpretability of DeMMO make the framework readily understandable beyond the computer science community, thereby facilitating its broader acceptance and adoption by both computational researchers and non-computational domain experts, particularly clinicians. Further discussion of related work and DeMMO's potential use in interpretable mobility-monitoring software is provided in Appendices A and G, respectively.

## 2 METHOD

**Mobilise-D Longitudinal Dataset.** To ground the proposed methodology in a concrete setting, we first introduce the recently released Mobilise-D dataset, a multicentre longitudinal DMO study of people with mobility-limiting conditions (Mazzà et al., 2021; Mikolaizak et al., 2022). The dataset contains mutually exclusive PD, MS, COPD, and PFF cohorts. Participants were scheduled for five visits, $T_1, \ldots, T_5$, followed by real-world monitoring with a body-worn inertial sensor. We analyse all four cohorts at all five visits. We use filtered weekly aggregates containing at least three valid monitoring days. Each record contains 24 DMOs describing walking amount, bout pattern, pace, rhythm, and variability. The quality-control variable `n_days_w` is not used as a predictor. Definitions of all 24 DMOs are provided in Appendix B. We consider five regression objectives: H&Y stage (PD H&Y) (Hoehn & Yahr, 1967) and MDS–UPDRS Part III (PD MDS–UPDRS) (Goetz et al., 2008) for PD, EDSS for MS (MS EDSS) (Kurtzke, 1983), SPPB for PFF (PFF SPPB) (Bellettiere et al., 2020), and $\text{FEV}_1$ as a percentage of its predicted value for COPD (COPD $\text{FEV}_1$) (Agustí et al., 2023). Because higher SPPB indicates better function, we define SPPB impairment as $12 - \text{SPPB}$, while retaining PFF SPPB as the objective's short name. Invalid values, including MDS–UPDRS scores of $-1$, are treated as missing. Complete-case processing is applied at the participant–visit level. A record is retained only when its clinical outcome and all 24 DMOs are available. Table 1 summarises the resulting dataset. Sample sizes decrease across visits because of loss to follow-up, unavailable recordings, insufficient monitoring days, and incomplete clinical assessments.

Table 1: Mobilise-D analytical dataset after requiring a valid target and complete observations for all 24 weekly DMOs. $N$ denotes participants with at least one retained visit.

| Disease | Target | $N$ | $T_1$ | $T_2$ | $T_3$ | $T_4$ | $T_5$ | Total | Range |
|---|---|---|---|---|---|---|---|---|---|
| PD | H&Y stage | 574 | 528 | 439 | 400 | 363 | 346 | 2,076 | 0–4 |
| PD | MDS–UPDRS III | 574 | 528 | 439 | 400 | 363 | 346 | 2,076 | 1–84 |
| MS | EDSS | 578 | 535 | 448 | 402 | 346 | 352 | 2,083 | 0–7.5 |
| PFF | SPPB impairment | 469 | 418 | 358 | 282 | 146 | 108 | 1,312 | 0–11 |
| COPD | $\text{FEV}_1$ % predicted | 583 | 547 | 437 | 376 | 364 | 348 | 2,072 | 9.26–117.18 |

### 2.1 LONGITUDINAL CLINICAL-PROGRESSION FORMULATION

We first consider one clinical outcome from a single disease cohort. This simplified setting introduces the longitudinal modelling, temporal regularisation, and DMO-selection components before their extension to multiple diseases and outcomes in Section 2.2.

At visit $t$, let $\mathbf{X}_t \in \mathbb{R}^{n_t \times p}$ denote the DMO feature matrix for $n_t$ retained records, where $p = 24$, and let $\mathbf{y}_t \in \mathbb{R}^{n_t}$ contain the corresponding clinical outcomes. Unlike conventional progression models that use baseline measurements to predict future outcomes, Mobilise-D provides both DMOs and clinical outcomes at each of $T = 5$ visits. We therefore treat clinical-status estimation at each visit as a separate but related regression task: $\mathbf{y}_t = \mathbf{X}_t\mathbf{w}_t + \boldsymbol{\epsilon}_t$, where $\mathbf{w}_t \in \mathbb{R}^p$ is the visit-specific coefficient vector and $\boldsymbol{\epsilon}_t$ denotes observational noise. The coefficients are collected as $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_T] \in \mathbb{R}^{p \times T}$. Each column represents the DMO–outcome relationship at one visit, while each row describes the longitudinal coefficient trajectory of one DMO. We use squared-error regression because the clinical outcomes are numerical scores. To prevent visits with larger samples from dominating estimation, the loss is normalised within each visit: $\mathcal{L}(\mathbf{W}) = \sum_{t=1}^{T} \frac{1}{2n_t} \|\mathbf{y}_t - \mathbf{X}_t\mathbf{w}_t\|_2^2$.

**Temporal Smoothness.** We model temporal smoothness using the fused Lasso, a well-established regulariser for longitudinal biomarker analysis (Zhou et al., 2012; 2023b). Let $\mathbf{R} \in \mathbb{R}^{(T-1)\times T}$ be the first-order difference matrix, with $R_{i,i} = -1$, $R_{i,i+1} = 1$, and all other entries zero. We use $\Omega_{\mathrm{FL}}(\mathbf{W}) = \|\mathbf{W}\mathbf{R}^{\mathsf{T}}\|_1$, which encourages adjacent visits to share coefficients while retaining temporally localised changes. This convex structure and its efficient optimisation have been established in previous work (Zhou et al., 2012; 2023b).

**Sparse Group Lasso.** Following Zhou et al. (2011), we use sparse group Lasso for interpretable DMO selection. Its element-wise term selects visit-specific effects, whereas its row-wise term selects DMOs across the longitudinal period: $\Omega_{\mathrm{SGL}}(\mathbf{W}) = \lambda_{\mathrm{L}}\|\mathbf{W}\|_1 + \lambda_{\mathrm{G}}\|\mathbf{W}\|_{2,1}$, where $\|\mathbf{W}\|_{2,1} = \sum_{j=1}^{p} \|\mathbf{w}_{j:}\|_2$. Thus, $\lambda_{\mathrm{L}}$ controls visit-specific sparsity and $\lambda_{\mathrm{G}}$ controls DMO-level sparsity. Sparse group Lasso is a mature convex feature-selection structure with efficient proximal algorithms (Tibshirani, 1996; Zhou et al., 2022a); selected DMOs are interpreted as candidates requiring subsequent clinical validation.

Combining the regression loss, fused temporal regularisation, and sparse group Lasso yields cFSGL, a classical temporal multi-task learning formulation (Zhou et al., 2012; 2013a):

$$\min_{\mathbf{W}} \quad \sum_{t=1}^{T} \frac{1}{2n_t}\|\mathbf{y}_t - \mathbf{X}_t\mathbf{w}_t\|_2^2 + \lambda_{\mathrm{FL}}\|\mathbf{W}\mathbf{R}^{\mathsf{T}}\|_1 + \lambda_{\mathrm{L}}\|\mathbf{W}\|_1 + \lambda_{\mathrm{G}}\|\mathbf{W}\|_{2,1}. \tag{1}$$

cFSGL has subsequently inspired numerous extensions of temporal multi-task learning (Cao et al., 2017; Zhou et al., 2022a; Zhou & Yang, 2023; Zhou et al., 2023b; Fan et al., 2026). However, this line of work models only a single clinical outcome within a single disease cohort. It cannot jointly model multiple outcomes across diseases, let alone accommodate disease cohorts with non-overlapping participants. We next extend this formulation to address both settings.

### 2.2 MULTI-DISEASE AND MULTI-OUTCOME EXTENSION

Our analysis includes multiple prediction objectives across four cohorts: PD contributes two clinical outcomes, whereas MS, PFF, and COPD each contribute one. The number of prediction objectives therefore differs from the number of diseases. Let $d \in \{1, \ldots, D\}$ index the disease cohorts, where $D = 4$, and let $Q_d$ denote the number of clinical outcomes for disease $d$. The total number of disease–outcome prediction objectives is $M = \sum_{d=1}^{D} Q_d = 5$.

For disease $d$, outcome $q$, and visit $t$, let $\mathbf{X}_{t,q}^{(d)} \in \mathbb{R}^{n_{dtq} \times p}$ contain the complete DMO observations and let $\mathbf{y}_{t,q}^{(d)} \in \mathbb{R}^{n_{dtq}}$ contain the corresponding clinical outcomes. Here, $n_{dtq}$ is the number of retained participant–visit records. Although all objectives share the same 24 DMO definitions, allowing $\mathbf{X}_{t,q}^{(d)}$ to depend on $q$ accommodates outcome-specific differences in data availability. The visit-specific regression model is

$$\mathbf{y}_{t,q}^{(d)} = \mathbf{X}_{t,q}^{(d)}\mathbf{w}_{t,q}^{(d)} + \boldsymbol{\epsilon}_{t,q}^{(d)}, \tag{2}$$

where $\mathbf{w}_{t,q}^{(d)} \in \mathbb{R}^p$. The longitudinal coefficient matrix for disease $d$ and outcome $q$ is $\mathbf{W}_{d,q} = [\mathbf{w}_{1,q}^{(d)}, \ldots, \mathbf{w}_{T,q}^{(d)}] \in \mathbb{R}^{p \times T}$. Thus, each $\mathbf{W}_{d,q}$ represents one longitudinal clinical prediction objective rather than an entire disease.

### 2.3 Automatic Relation Learning Across Prediction Objectives

PD, MS, PFF, and COPD have distinct pathological mechanisms and clinical outcomes, but their DMO–outcome mappings may share functional patterns. Similarly, the two PD outcomes may reflect related but non-identical aspects of disease severity and motor impairment. These relationships cannot be specified reliably in advance. Outcomes from different diseases are not jointly observed because the cohorts contain different participants, while correlations between outcomes within the same disease do not necessarily reflect similarities in their longitudinal DMO mappings. We therefore learn relationships among prediction objectives directly from their model parameters.

For disease $d$ and outcome $q$, let $\mathbf{u}_{d,q} = \mathrm{vec}(\mathbf{W}_{d,q}) \in \mathbb{R}^{pT}$. The representations of all $M$ prediction objectives are collected as $\mathbf{U} \in \mathbb{R}^{pT \times M}$, with each column corresponding to one disease–outcome pair. We assume that each longitudinal DMO mapping is related to the remaining mappings. After assigning each disease–outcome pair a unique objective index, this relation is expressed as

$$\mathbf{U} \approx \mathbf{U}\mathbf{A}, \qquad \mathbf{A} = \mathbf{A}^{\mathsf{T}}, \qquad \mathrm{diag}(\mathbf{A}) = \mathbf{0}, \tag{3}$$

where $\mathbf{A} \in \mathbb{R}^{M \times M}$ is the learned relation matrix. Symmetry produces an undirected relation graph, while the zero diagonal excludes trivial self-relations.

The matrix $\mathbf{A}$ can be partitioned into blocks $\mathbf{A}^{(d,k)} \in \mathbb{R}^{Q_d \times Q_k}$ according to disease membership. A diagonal block $\mathbf{A}^{(d,d)}$ captures relationships among outcomes within disease $d$, whereas an off-diagonal block $\mathbf{A}^{(d,k)}$ captures cross-disease relationships. Symmetry implies $\mathbf{A}^{(d,k)} = (\mathbf{A}^{(k,d)})^{\mathsf{T}}$. We estimate the relation structure using

$$\Omega_{\mathrm{R}}(\mathbf{U}, \mathbf{A}) = \frac{\lambda_{\mathrm{R}}}{2} \|\mathbf{U} - \mathbf{U}\mathbf{A}\|_F^2 + \frac{\lambda_{\mathrm{A}}}{2} \|\mathbf{A}\|_F^2, \tag{4}$$

where $\lambda_{\mathrm{R}} \geq 0$ controls cross-objective information sharing and $\lambda_{\mathrm{A}} > 0$ stabilises relation estimation. We use an $\ell_2$ penalty rather than sparsity because the model contains only five prediction objectives and weak relationships may still provide useful information.

### 2.4 Unified Multi-Disease and Multi-Outcome Objective

Combining longitudinal regression, fused temporal regularisation, sparse group DMO selection, and cross-disease and cross-outcome relation learning yields the proposed *DMO-enabled Multi-Disease and Multi-Outcome learning* framework (DeMMO):

$$\begin{aligned} \min_{\{\mathbf{W}_{d,q}\},\, \mathbf{A}} \quad & \sum_{d=1}^{D} \sum_{q=1}^{Q_d} \sum_{t=1}^{T} \frac{\|\mathbf{y}_{t,q}^{(d)} - \mathbf{X}_{t,q}^{(d)} \mathbf{w}_{t,q}^{(d)}\|_2^2}{2n_{dtq}} \\ & + \sum_{d=1}^{D} \sum_{q=1}^{Q_d} \big(\lambda_{\mathrm{FL}} \|\mathbf{W}_{d,q}\mathbf{R}^{\mathsf{T}}\|_1 + \lambda_{\mathrm{L}} \|\mathbf{W}_{d,q}\|_1 + \lambda_{\mathrm{G}} \|\mathbf{W}_{d,q}\|_{2,1}\big) \\ & + \frac{\lambda_{\mathrm{R}}}{2} \|\mathbf{U} - \mathbf{U}\mathbf{A}\|_F^2 + \frac{\lambda_{\mathrm{A}}}{2} \|\mathbf{A}\|_F^2 \\ \text{s.t.} \quad & \mathbf{A} = \mathbf{A}^{\mathsf{T}}, \qquad \mathrm{diag}(\mathbf{A}) = \mathbf{0}. \end{aligned} \tag{5}$$

It is worth noting that DeMMO does not require shared participants across disease cohorts, and each prediction objective retains its own longitudinal coefficient matrix. The objective is biconvex in $\{\mathbf{W}_{d,q}\}$ and $\mathbf{A}$. For fixed $\mathbf{A}$, the coefficient subproblem is convex but non-smooth because of the fused-Lasso, Lasso, and group-Lasso penalties. For fixed coefficient matrices, relation learning is a strongly convex quadratic problem over symmetric matrices with zero diagonal. We alternate between these two subproblems.

### 2.5 Data Standardisation

PD H&Y, PD MDS–UPDRS, MS EDSS, and COPD $\text{FEV}_1$ retain their original directions, while SPPB is transformed as $12 - \text{SPPB}$ to represent impairment. Consequently, higher values indicate greater severity or impairment for the first four objectives, whereas higher COPD $\text{FEV}_1$ indicates better pulmonary function. Each outcome is then standardised using the mean and standard deviation estimated from its training records pooled across visits. The 24 DMOs are similarly standardised using pooled training records from the corresponding disease cohort. Pooling across visits preserves longitudinal distributional changes that visit-specific standardisation could remove. All transformations and standardisation parameters are estimated exclusively from the training data and applied unchanged to the validation and test sets to prevent information leakage.

**Optimisation.** Note that optimising the DeMMO objective is challenging because it couples multiple parameter blocks under several non-smooth regularisers, precluding a straightforward closed-form solution. We therefore develop an efficient alternating optimisation algorithm that decomposes the original problem into tractable subproblems while preserving the intended structural constraints. Owing to space limitations, the complete algorithm, together with its convergence and computational complexity analyses, is provided in Appendix C. The resulting algorithm is computationally efficient: coefficient updates use established accelerated proximal-gradient operations, while relation learning adds little overhead because the five outcomes yield only ten free relation parameters.

**Additional results in the appendix.** It is well known that convex sparsity-inducing penalties can shrink nonzero coefficients and thereby introduce estimation bias (Fan & Li, 2001). To mitigate this bias, we develops two non-convex variants, DeMMO-var1 and DeMMO-var2. The technical details are in Appendix D. Their empirical results are reported in Appendix E. Despite their greater computational cost, both non-convex variants perform worse overall than the original DeMMO, suggesting that reducing sparsity-induced estimation bias is not beneficial in this setting. We conjecture that, because free-living DMOs contain substantial behavioural and measurement noise, the stronger shrinkage of the convex formulation provides useful regularisation, whereas the non-convex variants may retain more noise and generalise less effectively.

## 3 Experiments

We evaluate DeMMO on the Mobilise-D dataset described in Section 2. The analysis includes 24 weekly DMOs measured over five visits and five clinical prediction objectives: PD H&Y, PD MDS–UPDRS, MS EDSS, PFF SPPB, and COPD $\text{FEV}_1$. The resulting objective-specific sample sizes are reported in Table 1. For each experimental run, participants are divided into training, validation, and held-out test sets in a 70%/10%/20% ratio. All available visits from the same participant remain in one split to prevent information leakage. Data processing and standardisation follow Section 2.5, with all parameters estimated from the training data only. We repeat the complete procedure over five matched participant-level splits using random seeds 42–46. Following common practice (Zhou et al., 2024; Fan et al., 2026), we use normalised mean squared error (nMSE) and weighted correlation (wR) to assess overall performance, and root mean squared error (RMSE) for visit-specific performance. We also report $p$-values from two-sided paired $t$-tests comparing DeMMO with the strongest baseline.

### 3.1 Compared Methods and Training Settings

We compare DeMMO with eight baselines spanning conventional interpretable longitudinal models, and recent deep regression methods. All methods follow the common data-splitting and preprocessing protocol described above, with hyperparameters selected on the validation set and the held-out test set reserved for final evaluation. Methods that do not explicitly model time are fitted independently to each of the 25 outcome–visit tasks.

**DeMMO.** DeMMO is trained using the alternating optimisation procedure described in Section C. Because a joint grid search over its five regularisation parameters is computationally expensive, we use two-stage selection. First, relation learning is disabled while $\lambda_{\text{FL}}$, $\lambda_{\text{L}}$, and $\lambda_{\text{G}}$ are selected from $\{10^{-3}, 10^{-2}, 10^{-1}, 1\}$. These parameters are then fixed, and $(\lambda_{\text{R}}, \lambda_{\text{A}})$ is selected from $(0, 0)$ and $\{10^{-3}, 10^{-2}, 10^{-1}\}^2$. The selected model is refitted on the combined training and validation sets.

Table 2: Summary test performance on Mobilise-D (mean ± standard deviation). The best mean in each row is shown in bold.

| Objective | Metric | cFSGL | FRoTS | MAGPP | MLP-MSE | MLP-L1 | Rank-N | RankSim | ACCon | DeMMO | p-value |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Overall | nMSE ↓ | 0.787 ± 0.020 | 0.795 ± 0.020 | 0.812 ± 0.021 | 0.811 ± 0.023 | 0.835 ± 0.015 | 0.836 ± 0.019 | 0.807 ± 0.017 | 0.828 ± 0.020 | **0.769 ± 0.021** | 0.0032 |
| | wR ↑ | 0.443 ± 0.019 | 0.448 ± 0.015 | 0.437 ± 0.016 | 0.417 ± 0.024 | 0.399 ± 0.014 | 0.372 ± 0.023 | 0.421 ± 0.019 | 0.405 ± 0.017 | **0.464 ± 0.023** | 0.0260 |
| PD H&Y | nMSE ↓ | 0.969 ± 0.021 | 0.970 ± 0.026 | 0.994 ± 0.029 | 0.990 ± 0.020 | 1.025 ± 0.013 | 1.035 ± 0.008 | 0.992 ± 0.021 | 1.008 ± 0.017 | **0.937 ± 0.020** | 0.0008 |
| | wR ↑ | 0.264 ± 0.025 | 0.275 ± 0.022 | 0.254 ± 0.023 | 0.214 ± 0.060 | 0.156 ± 0.021 | 0.095 ± 0.045 | 0.234 ± 0.042 | 0.234 ± 0.026 | **0.300 ± 0.033** | 0.0038 |
| PD MDS–UPDRS | nMSE ↓ | 0.899 ± 0.034 | 0.899 ± 0.035 | 0.917 ± 0.035 | 0.916 ± 0.030 | 0.956 ± 0.022 | 0.956 ± 0.029 | 0.928 ± 0.030 | 0.951 ± 0.036 | **0.885 ± 0.032** | 0.0025 |
| | wR ↑ | 0.347 ± 0.032 | 0.363 ± 0.027 | 0.351 ± 0.027 | 0.304 ± 0.034 | 0.285 ± 0.030 | 0.240 ± 0.052 | 0.298 ± 0.036 | 0.264 ± 0.032 | **0.369 ± 0.034** | 0.0496 |
| MS EDSS | nMSE ↓ | 0.497 ± 0.037 | 0.501 ± 0.037 | 0.514 ± 0.038 | 0.515 ± 0.034 | 0.527 ± 0.034 | 0.538 ± 0.036 | 0.508 ± 0.036 | 0.525 ± 0.036 | **0.482 ± 0.034** | 0.0053 |
| | wR ↑ | 0.715 ± 0.026 | 0.713 ± 0.025 | 0.706 ± 0.026 | 0.704 ± 0.023 | 0.703 ± 0.023 | 0.697 ± 0.027 | 0.708 ± 0.024 | 0.697 ± 0.023 | **0.731 ± 0.024** | 0.0459 |
| PFF SPPB | nMSE ↓ | 0.560 ± 0.039 | 0.571 ± 0.052 | 0.585 ± 0.055 | 0.585 ± 0.036 | 0.592 ± 0.036 | 0.593 ± 0.043 | 0.579 ± 0.039 | 0.594 ± 0.055 | **0.550 ± 0.040** | 0.0061 |
| | wR ↑ | 0.671 ± 0.027 | 0.668 ± 0.034 | 0.662 ± 0.035 | 0.661 ± 0.023 | 0.661 ± 0.019 | 0.655 ± 0.027 | 0.660 ± 0.026 | 0.650 ± 0.040 | **0.680 ± 0.029** | 0.069 |
| COPD $FEV_1$ | nMSE ↓ | 0.933 ± 0.021 | 0.955 ± 0.020 | 0.969 ± 0.023 | 0.972 ± 0.054 | 0.989 ± 0.033 | 0.977 ± 0.031 | 0.951 ± 0.036 | 0.982 ± 0.044 | **0.916 ± 0.017** | 0.0151 |
| | wR ↑ | 0.298 ± 0.027 | 0.296 ± 0.018 | 0.289 ± 0.019 | 0.285 ± 0.035 | 0.278 ± 0.029 | 0.267 ± 0.032 | 0.286 ± 0.037 | 0.265 ± 0.036 | **0.315 ± 0.027** | 0.0228 |

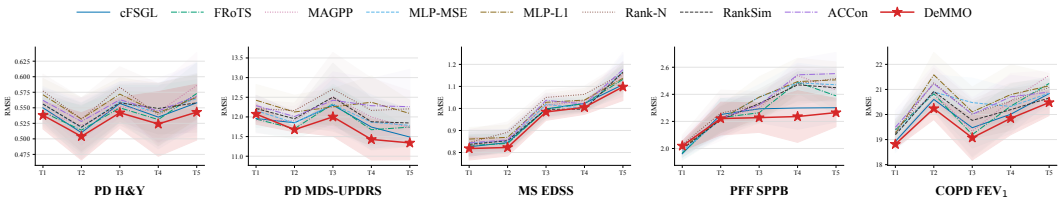


Figure 2: Visit-specific RMSE (mean ± standard deviation over five matched participant splits) for the five prediction objectives. Each panel reports T1–T5 performance for all baselines and DeMMO. Lower values indicate better predictive accuracy; no visit-level significance test is displayed.

Optimisation is limited to 50 outer and 500 inner iterations, with convergence tolerances of $10^{-6}$ and $10^{-7}$, respectively. A detailed hyperparameter sensitivity analysis of DeMMO is in Appendix F.1.

**Baselines.** For cFSGL (Zhou et al., 2012; 2013b), FRoTS (Zhou et al., 2023b), and MAGPP (Zhou et al., 2024), all penalty hyperparameters are selected from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ using validation nMSE. MLP-MSE uses a two-hidden-layer MLP with 32 ReLU units per layer, MSE loss, Adam optimisation, and validation-based early stopping. MLP-L1 uses the same setting but replaces MSE with MAE. Rank-N-Contrast (Rank-N) (Zha et al., 2024) pretrains a contrastive encoder using unit-width target bins and then fits a linear regression head. RankSim (Gong et al., 2022) combines MSE with a regulariser that aligns representation and continuous-label rankings. ACCon (Zhao et al., 2025) combines MSE with an angle-compensated contrastive regulariser, whose weight is selected from $\{10^{-2}, 10^{-1}, 1, 10, 100\}$ using validation MSE.

### 3.2 Prediction Performance and Learned Cross-Objective Relations

**Overall and outcome-specific performance.** DeMMO achieves the best mean nMSE and wR both overall and for each of the five prediction outcomes (Table 2). Relative to the strongest baseline for each metric, DeMMO reduces overall nMSE from 0.787 to 0.769 and increases overall wR from 0.448 to 0.464. Its advantage is consistent across outcomes: the nMSE improvements over the best corresponding baselines are 0.032, 0.014, 0.015, 0.010, and 0.017 for PD H&Y, PD MDS–UPDRS, MS EDSS, PFF SPPB, and COPD $FEV_1$, respectively. The corresponding wR improvements range from 0.006 to 0.025. The comparison with cFSGL most directly isolates the value of cross-objective relation learning. cFSGL uses the same temporal smoothing and stable and visit-specific feature-selection structure, but fits each clinical objective separately. By additionally learning and exploiting relations across objectives, DeMMO improves overall nMSE from 0.787 to 0.769 and wR from 0.443 to 0.464, while outperforming cFSGL on both metrics for all five objectives. This consistent gain indicates that DeMMO's advantage arises from selective cross-objective information sharing rather than temporal regularisation alone. More broadly, the structured longitudinal methods generally outperform the deep-regression baselines, suggesting that explicitly modelling the problem structure is more valuable here than increasing model capacity.

**Visit-specific performance.** Figure 2 shows that the aggregate improvement is not driven by a single visit or outcome. DeMMO obtains the lowest mean RMSE in 21 of the 25 outcome–visit

tasks: all five visits for PD H&Y and COPD $FEV_1$, T2–T5 for PD MDS–UPDRS, T1–T3 and T5 for MS EDSS, and T3–T5 for PFF SPPB. Other methods lead only at PD MDS–UPDRS T1, MS EDSS T4, and PFF SPPB T1–T2. The later-visit gains for PFF are particularly relevant because its follow-up sample size decreases substantially. More broadly, the heterogeneous error trajectories across outcomes show the need to preserve outcome-specific behaviour while sharing information selectively. Detailed values are provided in Appendix F, Tables 5–9.

**Learned cross-objective relations.** The optimiser learns the zero-diagonal cross-objective matrix $\mathbf{A}$. For interpretation, we report the complete matrix $\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$, whose unit diagonal represents each objective's relation with itself. Figure 3 shows its element-wise mean over the five matched splits. The strongest positive relation is between the two PD outcomes (0.37), consistent with their shared cohort and related measures of PD severity. Substantial positive relations also emerge between MS EDSS and PFF SPPB (0.34), PD H&Y and COPD $FEV_1$ (0.25), and PD H&Y and MS EDSS (0.22). These cross-disease relations demonstrate that DeMMO can identify transferable mobility structure even when cohorts share no participants, providing a mechanism for the gains over the independently fitted cFSGL models. The strongest negative relations are PD MDS–UPDRS–COPD $FEV_1$ ($-0.30$) and PD MDS–UPDRS–MS EDSS ($-0.19$). Note that relation signs describe the orientation of the learned longitudinal DMO effects; they do not represent associations between diseases or correlations between raw clinical scores. This distinction is especially important for COPD $FEV_1$, for which higher values indicate better pulmonary function, whereas higher values of the other outcomes indicate greater impairment. By learning both positive and negative relations, DeMMO can exploit aligned as well as complementary cross-objective structure rather than imposing uniformly positive information sharing.

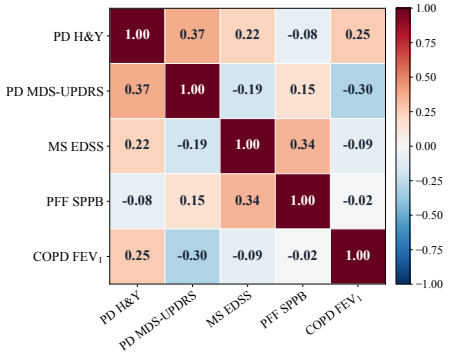


Figure 3: Mean complete relation matrix $\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ across the five matched splits.

### 3.3 Longitudinal Stability Selection

We perform longitudinal stability selection (Zhou et al., 2024; Fan et al., 2026; Zhou et al., 2023a) over ten randomly selected hyperparameter settings, each repeated on ten half-samples of the training participants. Figure 4 reports how often each DMO has a non-zero visit-specific coefficient in the resulting 100 fits. The profiles are broadly stable across visits but differ across objectives. The most stable DMOs are walking-bout-duration and cadence variability for PD H&Y; total walking-bout count for PD MDS–UPDRS; high-percentile walking speed and cadence for MS EDSS and PFF SPPB; and cadence P90, walking-bout-duration variability, and walking-bout count for COPD $FEV_1$. The strongest selection probabilities occur for MS EDSS, particularly for walking-speed and cadence P90 in bouts longer than 30 seconds (0.878 and 0.874, respectively). PFF SPPB shows a related emphasis on high-percentile pace and rhythm, whereas the two PD objectives favour partially overlapping but distinct measures. Consistently, variation across objectives is substantially greater than variation across visits (0.111 versus 0.019), and no DMO has a mean selection probability above 0.4 for all five objectives. This indicates that DMO relevance is longitudinally persistent but not universal across diseases or clinical measurements.

These findings complement the learned relation matrix: stability selection identifies reliably selected DMOs, whereas the relation matrix captures the alignment of their longitudinal effects. Shared DMOs may therefore have either aligned or opposing effects across objectives. DeMMO thus identifies persistent, objective-specific mobility patterns while enabling selective information sharing. These patterns require further clinical validation.

A broader value of DeMMO is that its structured and interpretable longitudinal DMO mappings enable reliable feature discovery and can subsequently support concise explanations in monitoring software. As a proof of concept, Appendix G applies stability selection to the DeMMO mappings and uses the two most stable MS DMOs to construct interpretable binary and three-class EDSS rules. This

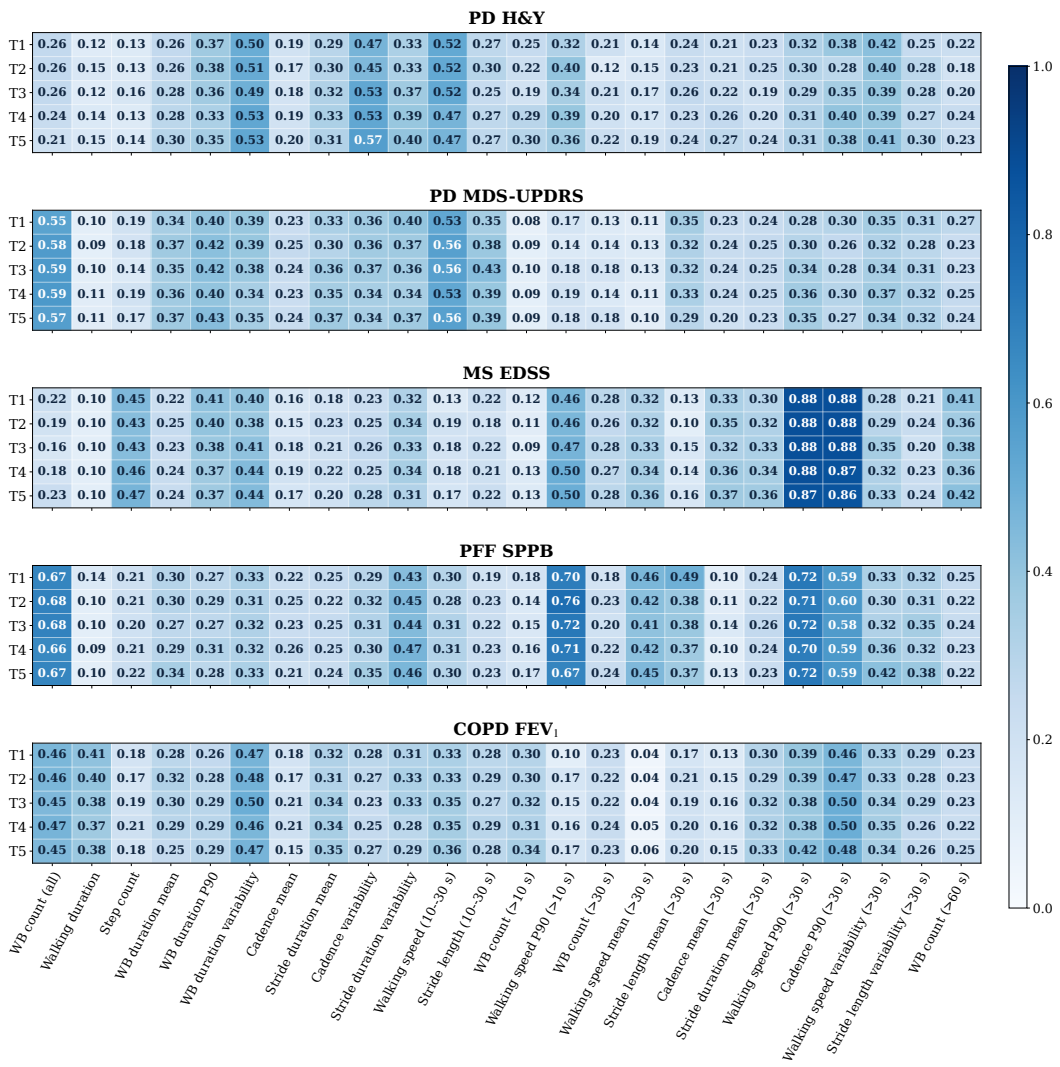


Figure 4: DMO selection probabilities across 100 half-sample fits (ten hyperparameter settings with ten repetitions) for each of the five prediction outcomes and visits T1–T5.

illustrates how the proposed framework can model all 24 DMOs while presenting users with a small and understandable set of indicators associated with a detected mobility change. The application remains exploratory and requires prospective validation.

## 4 Conclusion

We presented DeMMO, an interpretable framework for jointly modelling longitudinal DMOs across multiple diseases and clinical outcomes. By learning signed relations between longitudinal DMO mappings, DeMMO supports selective information sharing across participant-disjoint cohorts without requiring paired observations. On four Mobilise-D disease cohorts, five prediction objectives, and five visits, DeMMO achieved the best overall performance among the evaluated methods and led in most outcome- and visit-specific comparisons. The learned relations and stability analysis further show that informative DMO patterns are persistent across visits yet distinct across clinical objectives, supporting structured sharing rather than a universal mobility profile. Owing to its simple and interpretable formulation, DeMMO provides both accurate predictions and clinically inspectable longitudinal patterns. Future work will validate these patterns in independent cohorts and examine their generalisability to additional diseases, outcomes, and longer follow-up periods.

### AI USE STATEMENT

Generative AI tools were used to assist with language editing, LaTeX formatting, and code review. The authors reviewed all AI-assisted material, verified the reported claims and experimental results, and take responsibility for the final content of the paper and its associated artifacts.

### REPRODUCIBILITY STATEMENT

The paper specifies the model, optimisation procedure, data splitting, preprocessing, hyperparameter selection, and evaluation protocol. Source code is in the supplementary material. Additional analyses are reported in the appendix.

## REFERENCES

Alvar Agustí, Bartolome R Celli, Gerard J Criner, David Halpin, Antonio Anzueto, Peter Barnes, Jean Bourbeau, MeiLan K Han, Fernando J Martinez, Maria Montes de Oca, et al. Global initiative for chronic obstructive lung disease 2023 report: Gold executive summary. *European Respiratory Journal*, 61(4):2300239, 2023.

Lisa Alcock, Encarna Micó-Amigo, Tecla Bonci, and Silvia Del Din. Algorithms for gait: technical and clinical validity. In *Gait, Balance, and Mobility Analysis*, pp. 277–321. Elsevier, 2025.

Lmar M Babrak, Joseph Menetski, Michael Rebhan, Giovanni Nisato, Marc Zinggeler, Noé Brasier, Katja Baerenfaller, Thomas Brenzikofer, Laurenz Baltzer, Christian Vogler, et al. Traditional and digital biomarkers: two worlds apart? *Digital biomarkers*, 3(2):92–102, 2019.

Tian Bai, Shanshan Zhang, Brian L Egleston, and Slobodan Vucetic. Interpretable representation learning for healthcare via capturing disease progression through time. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 43–51, 2018.

John Bellettiere, Michael J Lamonte, Jonathan Unkart, Sandy Liles, Deepika Laddu-Patel, JoAnn E Manson, Hailey Banack, Rebecca Seguin-Fowler, Paul Chavez, Lesley F Tinker, et al. Short physical performance battery and incident cardiovascular events among older women. *Journal of the American Heart Association*, 9(14):e016845, 2020.

Elroy Boers, Angier Allen, Meredith Barrett, Adam V Benjafield, Mary B Rice, Jadwiga A Wedzicha, Leanne Kaye, Heather J Zar, Sanjeev Sinha, Obianuju Ozoh, et al. Forecasting the global economic and health burden of chronic obstructive pulmonary disease from 2025 through 2050. *Chest*, 2025.

A Bouleau, C Dulong, CA Schwerer, R Delgrange, K Bouaou, T Brochu, S Zinai, K Švecová, MJ Sá, A Petropoulos, et al. The socioeconomic impact of multiple sclerosis in france: Results from the petals study. *Multiple Sclerosis Journal–Experimental, Translational and Clinical*, 8(2): 20552173221093219, 2022.

Gavin Brittain, Michael Long, Ellen Buckley, Clemens Becker, Joren Buekers, Martí de Las Heras, Claudia Armengol, Sarah Koch, Julia Braun, Brian Caufield, et al. Testing the construct validity of real-world digital mobility outcomes in multiple sclerosis. *Multiple Sclerosis Journal*, pp. 13524585261464537, 2026.

Peng Cao, Xuanfeng Shan, Dazhe Zhao, Min Huang, and Osmar Zaiane. Sparse shared structure based multi-task learning for mri based cognitive performance prediction of alzheimer's disease. *Pattern Recognition*, 72:219–235, 2017.

Francesca Cerreta, Armin Ritzhaupt, Thomas Metcalfe, Scott Askin, João Duarte, Michael Berntgen, and Spiros Vamvakas. Digital technologies for medicines: shaping a framework for success. *Nature Reviews Drug Discovery*, 19(9):573–574, 2020.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pp. 301–318. PMLR, 2016.

Philip Coran, Jennifer C Goldsack, Cheryl A Grandinetti, Jessie P Bakker, Marisa Bolognese, E Dorsey, Kaveeta Vasisht, Adam Amdur, Christopher Dell, Jonathan Helfgott, et al. Advancing the use of mobile technologies in clinical trials: recommendations from the clinical trials transformation initiative. *Digital Biomarkers*, 3(3):145–154, 2020.

Marta Francisca Corrà, Arash Atrsaei, Ana Sardoreira, Clint Hansen, Kamiar Aminian, Manuel Correia, Nuno Vila-Chã, Walter Maetzler, and Luís Maia. Comparison of laboratory and daily-life gait speed assessment during on and off states in parkinson's disease. *Sensors*, 21(12):3974, 2021.

Katherine HO Deane, Helen Flaherty, David J Daley, Roland Pascoe, Bridget Penhale, Carl E Clarke, Catherine Sackley, and Stacey Storey. Priority setting partnership to identify the top 10 research priorities for the management of parkinson's disease. *BMJ open*, 4(12):e006434, 2014.

Laura Delgado-Ortiz, Ashley Polhemus, Alison Keogh, Norman Sutton, Werner Remmele, Clint Hansen, Felix Kluge, Basil Sharrack, Clemens Becker, Thierry Troosters, et al. Listening to the patients' voice: a conceptual framework of the walking experience. *Age and Ageing*, 52(1):afac233, 2023.

Tobias Eckert, Martin Aursand Berge, Michael Long, Martí de Las Heras, Paula Alvarez, Hubert Blain, Julia Braun, Joren Buekers, Brian Caulfield, Monika Engdal, et al. Construct validity of real-world digital mobility outcomes in patients after proximal femoral fracture: a cross-sectional observational study. *Scientific Reports*, 16(1):9535, 2026.

Monika Engdal, Kristin Taraldsen, Carl-Philipp Jansen, Raphael Simon Peter, Beatrix Vereijken, Clemens Becker, Jorunn Laegdheim Helbostad, and Jochen Klenk. Real-world mobility recovery after hip fracture: secondary analyses of digital mobility outcomes from four randomized controlled trials. *Age and ageing*, 53(10):afae234, 2024.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

Xuanhan Fan, Menghui Zhou, Yu Zhang, Jun Qi, Yun Yang, and Po Yang. Beyond single scores: A multi-cognitive objective learning for ad progression prediction. *Pattern Recognition*, pp. 114348, 2026.

Mostafa Mehdipour Ghazi, Mads Nielsen, Akshay Pai, M Jorge Cardoso, Marc Modat, Sébastien Ourselin, Lauge Sørensen, Alzheimer's Disease Neuroimaging Initiative, et al. Training recurrent neural networks robust to incomplete data: application to alzheimer's disease progression modeling. *Medical image analysis*, 53:39–46, 2019.

Christopher G Goetz, Barbara C Tilley, Stephanie R Shaftman, Glenn T Stebbins, Stanley Fahn, Pablo Martinez-Martin, Werner Poewe, Cristina Sampaio, Matthew B Stern, Richard Dodel, et al. Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): scale presentation and clinimetric testing results. *Movement disorders: official journal of the Movement Disorder Society*, 23(15):2129–2170, 2008.

Jennifer C Goldsack, Andrea Coravos, Jessie P Bakker, Brinnae Bent, Ariel V Dowling, Cheryl Fitzer-Attas, Alan Godfrey, Job G Godino, Ninad Gujar, Elena Izmailova, et al. Verification, analytical validation, and clinical validation (v3): the foundation of determining fit-for-purpose for biometric monitoring technologies (biomets). *npj digital Medicine*, 3(1):55, 2020.

Yu Gong, Greg Mori, and Frederick Tung. Ranksim: Ranking similarity regularization for deep imbalanced regression. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 7634–7649. PMLR, 2022.

Margaret M Hoehn and Melvin D Yahr. Parkinsonism: onset, progression, and mortality. *Neurology*, 17(5):427–427, 1967.

Alison Keogh, Lisa Alcock, Philip Brown, Ellen Buckley, Marina Brozgol, Eran Gazit, Clint Hansen, Kirsty Scott, Lars Schwickert, Clemens Becker, et al. Acceptability of wearable devices for measuring mobility remotely: observations from the mobilise-d technical validation study. *Digital Health*, 9:20552076221150745, 2023.

Cameron Kirk, Arne Küderle, M Encarna Micó-Amigo, Tecla Bonci, Anisoara Paraschiv-Ionescu, Martin Ullrich, Abolfazl Soltani, Eran Gazit, Francesca Salis, Lisa Alcock, et al. Mobilise-d insights to estimate real-world walking speed in multiple conditions with a wearable device. *Scientific reports*, 14(1):1754, 2024.

Cameron Kirk, Emma Packer, Ashley Polhemus, Mhairi K MacLean, Harry Bailey, Felix Kluge, Heiko Gaßner, Lynn Rochester, Silvia Del Din, and Alison J Yarnall. A systematic review of real-world gait-related digital mobility outcomes in parkinson's disease. *npj Digital Medicine*, 8 (1):585, 2025.

Felix Kluge, Silvia Del Din, Andrea Cereatti, Heiko Gaßner, Clint Hansen, Jorunn L Helbostad, Jochen Klucken, Arne Küderle, Arne Müller, Lynn Rochester, et al. Consensus based framework for digital mobility monitoring. *PloS one*, 16(8):e0256541, 2021.

John F Kurtzke. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (edss). *Neurology*, 33(11):1444–1452, 1983.

Hui Li, Yanlin Wang, Ziyu Lyu, and Jieming Shi. Multi-task learning for recommendation over heterogeneous information network. *IEEE Transactions on Knowledge and Data Engineering*, 34 (2):789–802, 2020a.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):7155, 2020b.

Tengfei Ma, Xuan Lin, Bosheng Song, Philip S Yu, and Xiangxiang Zeng. Kg-mtl: knowledge graph enhanced multi-task learning for molecular interaction. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):7068–7081, 2022.

Claudia Mazzà, Lisa Alcock, Kamiar Aminian, Clemens Becker, Stefano Bertuletti, Tecla Bonci, Philip Brown, Marina Brozgol, Ellen Buckley, Anne-Elie Carsin, et al. Technical validation of real-world monitoring of gait: a multicentric observational study. *BMJ open*, 11(12):e050785, 2021.

Dimitrios Megaritis, Michael Long, Martí de Las Heras, Victoria Alcaraz-Serrano, Paula Alvarez, Clemens Becker, Julia Braun, Joren Buekers, Sara Buttery, Brian Caulfield, et al. The construct validity of real-world digital mobility outcomes in people with copd. *ERJ Open Research*, 12(2): 00993–2025, 2026.

M Encarna Micó-Amigo, Tecla Bonci, Anisoara Paraschiv-Ionescu, Martin Ullrich, Cameron Kirk, Abolfazl Soltani, Arne Küderle, Eran Gazit, Francesca Salis, Lisa Alcock, et al. Assessing real-world gait with digital technology? validation, insights and recommendations from the mobilise-d consortium. *Journal of neuroengineering and rehabilitation*, 20(1):78, 2023.

A Stefanie Mikolaizak, Lynn Rochester, Walter Maetzler, Basil Sharrack, Heleen Demeyer, Claudia Mazzà, Brian Caulfield, Judith Garcia-Aymerich, Beatrix Vereijken, Valdo Arnera, et al. Connecting real-world digital mobility assessment to clinical outcomes for regulatory and clinical endorsement–the mobilise-d study protocol. *PloS one*, 17(10):e0269615, 2022.

Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6(1): 26094, 2016.

Anat Mirelman, Jana Volkov, Amit Salomon, Eran Gazit, Alice Nieuwboer, Lynn Rochester, Silvia Del Din, Laura Avanzino, Elisa Pelosin, Bastiaan R Bloem, et al. Digital mobility measures: a window into real-world severity and progression of parkinson's disease. *Movement Disorders*, 39 (2):328–338, 2024.

Y Nesterov. A method for solving a convex programming problem with convergence rate o (1/k2). In *Soviet Mathematics. Doklady*, volume 27, pp. 367–372, 1983.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

Minh Nguyen, Tong He, Lijun An, Daniel C Alexander, Jiashi Feng, BT Thomas Yeo, Alzheimer's Disease Neuroimaging Initiative, et al. Predicting alzheimer's disease progression using deep recurrent neural networks. *NeuroImage*, 222:117203, 2020.

Margaux Poleur, Barbara Willekens, Bertrand Degos, Damien Ricard, Vincent Van Pesch, Annick Mélin, Oihana Piquet, Alexis Tricot, Laurie Médard, Mona Michaud, et al. Stride-level measurement of gait as an early sensitive marker of disability progression in ambulatory patients with multiple sclerosis. *EClinicalMedicine*, 93, 2026.

Ashley Polhemus, Laura Delgado-Ortiz, Gavin Brittain, Nikolaos Chynkiamis, Francesca Salis, Heiko Gaßner, Michaela Gross, Cameron Kirk, Rachele Rossanigo, Kristin Taraldsen, et al. Walking on common ground: a cross-disciplinary scoping review on the clinical utility of digital mobility outcomes. *NPJ digital medicine*, 4(1):149, 2021.

Rebecca J Port, Martin Rumsby, Graham Brown, Ian F Harrison, Anneesa Amjad, and Claire J Bale. People with parkinson's disease: what symptoms do they most want to improve and how does this change with disease duration? *Journal of Parkinson's disease*, 11(2):715–724, 2021.

Stéphanie A Prince, Kristi B Adamo, Meghan E Hamel, Jill Hardt, Sarah Connor Gorber, and Mark Tremblay. A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. *International journal of behavioral nutrition and physical activity*, 5 (1):56, 2008.

Pablo Rábano-Suárez, Natalia Del Campo, Isabelle Benatru, Caroline Moreau, Clément Desjardins, Álvaro Sánchez-Ferro, and Margherita Fabbri. Digital outcomes as biomarkers of disease progression in early parkinson's disease: a systematic review. *Movement Disorders*, 40(2):184–203, 2025.

Lynn Rochester, Claudia Mazzà, Arne Mueller, Brian Caulfield, Marie McCarthy, Clemens Becker, Ram Miller, Paolo Piraino, Marco Viceconti, Wilhelmus P Dartee, et al. A roadmap to inform development, validation and approval of digital mobility outcomes: the mobilise-d approach. *Digital biomarkers*, 4(Suppl. 1):13–27, 2020.

Vrutangkumar V Shah, James McNames, Martina Mancini, Patricia Carlson-Kuhta, Rebecca I Spain, John G Nutt, Mahmoud El-Gohary, Carolin Curtze, and Fay B Horak. Laboratory versus daily life gait characteristics in patients with multiple sclerosis, parkinson's disease, and matched controls. *Journal of neuroengineering and rehabilitation*, 17(1):159, 2020.

Satya Narayan Shukla and Benjamin M Marlin. Multi-time attention networks for irregularly sampled time series. In ... *International Conference on Learning Representations*, volume 2021, pp. 14897, 2021.

Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of computational and graphical statistics*, 22(2):231–245, 2013.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

Sam Williamson, Filipa Landeiro, Thomas McConnell, Lucy Fulford-Smith, M Kassim Javaid, Andrew Judge, and José Leal. Costs of fragility hip fractures globally: a systematic review and meta-regression analysis. *Osteoporosis international*, 28(10):2791–2800, 2017.

Wenya Yang, Jamie L Hamilton, Catherine Kopil, James C Beck, Caroline M Tanner, Roger L Albin, E Ray Dorsey, Nabila Dahodwala, Inna Cintina, Paul Hogan, et al. Current and projected future economic burden of parkinson's disease in the us. *npj Parkinson's Disease*, 6(1):15, 2020.

Kaiwen Zha, Peng Cao, Jeany Son, Yuzhe Yang, and Dina Katabi. Rank-n-contrast: learning continuous representations for regression. *Advances in Neural Information Processing Systems*, 36, 2024.

Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

Botao Zhao, Xiaoyang Qu, Zuheng Kang, Junqing Peng, Jing Xiao, and Jianzong Wang. Accon: Angle-compensated contrastive regularizer for deep regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22750–22758, 2025.

Jiayu Zhou, Lei Yuan, Jun Liu, and Jieping Ye. A multi-task learning formulation for predicting disease progression. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 814–822, 2011.

Jiayu Zhou, Jun Liu, Vaibhav A Narayan, and Jieping Ye. Modeling disease progression via fused sparse group lasso. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1095–1103, 2012.

Jiayu Zhou, Jun Liu, Vaibhav A Narayan, Jieping Ye, Alzheimer's Disease Neuroimaging Initiative, et al. Modeling disease progression via multi-task learning. *NeuroImage*, 78:233–248, 2013a.

Menghui Zhou and Po Yang. Automatic temporal relation in multi-task learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3570–3580, 2023.

Menghui Zhou, Yu Zhang, Tong Liu, Yun Yang, and Po Yang. Multi-task learning with adaptive global temporal structure for predicting alzheimer's disease progression. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 2743–2752, 2022a.

Menghui Zhou, Yu Zhang, Tong Liu, Yun Yang, and Po Yang. Efficient multi-task learning with adaptive temporal structure for progression prediction. *Neural Computing and Applications*, pp. 3570–3580, 2023a.

Menghui Zhou, Yu Zhang, Yun Yang, Tong Liu, and Po Yang. Robust temporal smoothness in multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11426–11434, 2023b.

Menghui Zhou, Xulong Wang, Tong Liu, Yun Yang, and Po Yang. Integrating visualised automatic temporal relation graph into multi-task learning for alzheimer's disease progression prediction. *IEEE Transactions on Knowledge and Data Engineering*, 36(10):5206–5220, 2024.

Qiang Zhou, Gang Wang, Kui Jia, and Qi Zhao. Learning to share latent tasks for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2264–2271, 2013b.

Xiao Zhou, Yong Lin, Renjie Pi, Weizhong Zhang, Renzhe Xu, Peng Cui, and Tong Zhang. Model agnostic sample reweighting for out-of-distribution learning. In *International conference on machine learning*, pp. 27203–27221. PMLR, 2022b.

# Appendix

## A Related Work

### A.1 Real-World DMO Estimation and Clinical Validation

Research on DMOs has examined whether mobility can be estimated accurately from wearable-sensor signals and whether the resulting measures are clinically meaningful (Goldsack et al., 2020; Cerreta et al., 2020; Polhemus et al., 2021). Consensus definitions have improved comparability across sensors, algorithms, and walking-bout definitions (Kluge et al., 2021). The Mobilise-D Technical Validation Study further demonstrated the feasibility of deriving real-world DMOs from a single wearable device across several mobility-limiting conditions (Mazzà et al., 2021; Keogh et al., 2023). Detection and cadence estimation were generally robust, whereas stride length and walking speed were more affected by context and impairment severity (Micó-Amigo et al., 2023; Kirk et al., 2024). Clinical evidence remains dominated by known-groups comparisons and concurrent associations, with less support for predictive validity, responsiveness, and ecological validity (Polhemus et al., 2021). Differences between laboratory and daily-life gait further motivate real-world validation (Shah et al., 2020; Corrà et al., 2021). Recent Mobilise-D studies have related individual DMOs to clinical measures and severity or recovery groups (Megaritis et al., 2026; Eckert et al., 2026), but do not show how multiple DMOs should be combined or how their relationships with clinical status evolve over time.

This literature reflects the distinction between technical and clinical validation. Technical studies establish that a sensor and algorithm can measure a walking construct with acceptable accuracy, whereas clinical validation asks whether that construct is associated with an outcome of interest and is fit for a specified context of use (Goldsack et al., 2020; Rochester et al., 2020). Mobilise-D was designed around this staged process: its technical validation study evaluated the measurement pipeline across conditions, while the Clinical Validation Study linked harmonised real-world DMOs to disease-specific clinical outcomes (Mazzà et al., 2021; Mikolaizak et al., 2022). Construct-validity analyses in COPD, PFF, and MS provide important outcome-specific evidence (Megaritis et al., 2026; Eckert et al., 2026; Brittain et al., 2026), but are primarily cross-sectional. DeMMO complements these studies by analysing the joint longitudinal contribution of the complete DMO panel rather than testing each DMO separately.

### A.2 Longitudinal DMO Mining

Longitudinal DMO research remains limited. Reviews in PD found few studies with extended follow-up and a literature dominated by active or hospital-based tests, patient–control comparisons, and cross-sectional associations, leaving longitudinal change and predictive utility underexplored (Rábano-Suárez et al., 2025; Kirk et al., 2025). Emerging studies show that selected measures can capture change in PD, MS, and hip-fracture recovery (Mirelman et al., 2024; Poleur et al., 2026; Engdal et al., 2024). However, they are disease-specific and assess only a few prespecified outcomes; they do not jointly model how multivariate DMO–outcome relationships evolve across visits. To our knowledge, stable and visit-specific DMOs have not been systematically identified across multiple outcomes and mobility-limiting conditions. The Mobilise-D Clinical Validation Study provides the harmonised longitudinal data needed to address this gap (Mikolaizak et al., 2022).

Longitudinal evidence can address several distinct questions: whether a DMO changes over time, whether its change tracks clinical progression, whether it predicts a future endpoint, and whether its association with an endpoint is stable across visits. These questions are not interchangeable. A measure may show a population-level trend but contribute little to individual prediction, or remain predictive even when its marginal mean changes little. Existing studies have mainly examined trajectories of individual measures or their association with a single clinical endpoint (Mirelman et al., 2024; Engdal et al., 2024; Poleur et al., 2026). In contrast, our setting estimates a visit-specific multivariate mapping from 24 correlated DMOs to each clinical outcome. Stability selection then distinguishes effects that persist across visits and participant subsamples from those that are visit-specific, without interpreting statistical selection alone as clinical validation.

Attrition and incomplete follow-up add a further difficulty. Later visits often contain fewer observations, making independent visit-wise models unstable, whereas complete-case trajectory analysis would discard participants with partially observed follow-up. DeMMO retains all valid participant–visit records and borrows strength through coefficient-level temporal regularisation. It therefore targets the evolution of DMO–outcome associations rather than requiring every participant to have a complete five-visit trajectory.

### A.3 Multi-Task Learning for Longitudinal Modelling

Multi-task learning (MTL) (Zhang & Yang, 2021; Ma et al., 2022; Li et al., 2020a) has been applied to disease progression by treating predictions at different future time points as related tasks. Studies in Alzheimer's disease use temporal smoothness, structured sparsity, and adaptive or robust regularisers to identify shared and time-specific biomarkers (Zhou et al., 2011; 2012; 2022a; 2023b; 2024). These methods generally predict multiple future outcomes from baseline features (Zhou et al., 2011; 2012; 2022b; 2024; Fan et al., 2026); in Mobilise-D, both DMOs and outcomes are repeatedly observed, so the goal is instead to model how their relationships evolve across visits. Mobilise-D also spans multiple outcomes and diseases. Recent MTL studies often define outcome relationships through patient-level correlations within one disease (Fan et al., 2026); this is unsuitable for non-overlapping disease cohorts. Cross-disease relationships must instead be learned from shared DMO-based prediction structures while preserving disease-specific patterns. Fused Lasso and sparse group Lasso provide mechanisms for temporal smoothness and structured selection (Tibshirani et al., 2005; Zhou et al., 2012; Simon et al., 2013). Here, they identify stable and visit-specific DMOs among 24 measures for parsimony and clinical interpretability. Existing MTL methods do not jointly accommodate repeated DMO–outcome relationships, multiple outcomes, and non-overlapping disease cohorts, motivating a framework that learns temporal, outcome-specific, and cross-disease structures simultaneously.

The task definition is central to this distinction. In conventional longitudinal MTL, tasks commonly correspond to future time points, and a shared baseline representation is used to predict several follow-up outcomes (Zhou et al., 2011; 2012; 2022b). Robust and adaptive extensions allow task-specific deviations or learn relations among these time points (Zhou et al., 2023b; 2024). In DeMMO, each disease–outcome pair is instead one prediction objective, and its coefficient matrix contains the complete sequence of visit-specific DMO mappings. Temporal dependence is modelled within each objective, while relation learning operates between objectives. This separation prevents a strong temporal relation from being confused with evidence that two clinical outcomes share the same DMO structure.

Cross-outcome learning also differs from ordinary multi-output regression. The two PD outcomes are observed in the same cohort, but MS, PFF, and COPD outcomes come from different participants and have different clinical scales. Raw outcome covariance is therefore unavailable across diseases and would not, in any case, directly describe similarity between their DMO effects. DeMMO learns a symmetric signed relation graph from the longitudinal coefficient mappings. Positive and negative weights can represent aligned and oppositely oriented DMO patterns, while the zero diagonal excludes trivial self-relations. This parameter-level mechanism enables selective sharing across non-overlapping cohorts without pooling participant records or assuming identical effects.

### A.4 Deep Learning for Disease Progression Modelling

Deep learning has been widely applied to large-scale electronic health records, medical images, and physiological signals. Representative models learn patient representations or temporal patterns to predict diseases, diagnoses, medications, and future conditions (Miotto et al., 2016; Choi et al., 2016; Li et al., 2020b; Bai et al., 2018). Recurrent models have also jointly predicted longitudinal imaging, diagnoses, cognitive scores, and ventricular volume in Alzheimer's disease while accommodating missing observations (Ghazi et al., 2019; Nguyen et al., 2020). These methods can capture nonlinear temporal dependencies but generally benefit from high-dimensional inputs, dense observations, or large patient populations. Mobilise-D instead contains 24 structured DMOs, at most five visits per participant, and smaller disease- and outcome-specific samples. Its main challenge is to identify interpretable, time-varying DMO relationships across clinical outcomes and diseases. A structured

Table 3: Definitions of the 24 weekly DMO predictors. WB denotes walking bout.

| Variable | Meaning | Variable | Meaning |
|---|---|---|---|
| wb_all_sum_w | Number of WBs | wb_10_sum_w | Number of WBs longer than 10 s |
| walkdur_all_sum_w | Total walking duration | ws_10_p90_w | 90th percentile of walking speed in WBs longer than 10 s |
| wbsteps_all_sum_w | Total number of walking steps | wb_30_sum_w | Number of WBs longer than 30 s |
| wbdur_all_avg_w | Mean WB duration | ws_30_avg_w | Mean walking speed in WBs longer than 30 s |
| wbdur_all_p90_w | 90th percentile of WB duration | strlen_30_avg_w | Mean stride length in WBs longer than 30 s |
| wbdur_all_var_w | Variability of WB duration | cadence_30_avg_w | Mean cadence in WBs longer than 30 s |
| cadence_all_avg_w | Mean cadence across all WBs | strdur_30_avg_w | Mean stride duration in WBs longer than 30 s |
| strdur_all_avg_w | Mean stride duration across all WBs | ws_30_p90_w | 90th percentile of walking speed in WBs longer than 30 s |
| cadence_all_var_w | Variability of cadence across all WBs | cadence_30_p90_w | 90th percentile of cadence in WBs longer than 30 s |
| strdur_all_var_w | Variability of stride duration across all WBs | ws_30_var_w | Variability of walking speed in WBs longer than 30 s |
| ws_1030_avg_w | Mean walking speed in WBs lasting 10–30 s | strlen_30_var_w | Variability of stride length in WBs longer than 30 s |
| strlen_1030_avg_w | Mean stride length in WBs lasting 10–30 s | wb_60_sum_w | Number of WBs longer than 60 s |

multi-task model therefore provides more appropriate inductive biases and more interpretable effects than a high-capacity deep sequence model.

Deep models for irregular clinical time series can explicitly encode missing values, elapsed time, or attention across observations (Che et al., 2018; Shukla & Marlin, 2021). Such architectures are valuable when the principal signal lies in dense, heterogeneous event sequences. Here, however, the inputs are a small, clinically defined DMO panel observed on a fixed five-visit schedule, and the scientific objective includes identifying which DMO is relevant at which visit. Post-hoc feature attribution would not provide the same direct coefficient trajectories or structured sparsity.

We nevertheless include neural regression and continuous-label representation learning baselines in the empirical comparison. This tests whether nonlinear capacity or label-aware embeddings improve prediction in the available sample regime. The comparison is therefore not intended to claim that structured linear models dominate deep learning generally; rather, it evaluates whether the temporal, sparse, and cross-objective inductive biases of DeMMO are better matched to this particular longitudinal clinical problem.

## B Definitions of the 24 Digital Mobility Outcomes

Table 3 defines the 24 weekly aggregated DMOs used as predictors. A walking bout (WB) is a continuous period of detected walking. In the variable names, `all`, `10`, `30`, and `60` denote all WBs or WBs longer than 10, 30, and 60 seconds, respectively; `1030` denotes WBs lasting 10–30 seconds. The suffixes `avg`, `p90`, `var`, and `sum` denote the mean, 90th percentile, variability, and amount/count aggregation, while `w` indicates aggregation over the monitoring week.

Together, these variables characterise five complementary aspects of real-world mobility: walking amount, bout pattern, pace, rhythm, and variability. Bout-duration thresholds distinguish short habitual walking from longer, sustained walking, while mean, upper-percentile, and variability summaries capture typical, high-performance, and within-person heterogeneous mobility behaviour, respectively.

## C Optimization

The unified objective in Equation 5 is not jointly convex in the longitudinal coefficient matrices and relation matrix because of the multiplicative term $\mathbf{UA}$. However, it is convex in either block when the other is fixed. We therefore use alternating optimisation to update the longitudinal coefficient matrices $\{\mathbf{W}_{d,q}\}$ with $\mathbf{A}$ fixed, followed by the symmetric relation matrix $\mathbf{A}$ with $\{\mathbf{W}_{d,q}\}$ fixed.

For simplicity, we assign each disease–outcome pair a unique index $m \in \{1, \ldots, M\}$, where $M = 5$, and denote its coefficient matrix by $\mathbf{W}_m = [\mathbf{w}_{m1}, \ldots, \mathbf{w}_{mT}] \in \mathbb{R}^{p \times T}$. The corresponding data at visit $t$ are denoted by $\mathbf{X}_{mt}$, $\mathbf{y}_{mt}$, and $n_{mt}$.

### C.1 Updating the Longitudinal Coefficient Matrices

For fixed $\mathbf{A}$, the coefficient subproblem is decomposed into the smooth component

$$f(\{\mathbf{W}_m\}; \mathbf{A}) = \sum_{m=1}^{M} \sum_{t=1}^{T} \frac{1}{2n_{mt}} \|\mathbf{y}_{mt} - \mathbf{X}_{mt}\mathbf{w}_{mt}\|_2^2 + \frac{\lambda_{\mathrm{R}}}{2} \|\mathbf{U} - \mathbf{U}\mathbf{A}\|_F^2, \tag{6}$$

and the non-smooth component

$$g(\{\mathbf{W}_m\}) = \sum_{m=1}^{M} \left[\lambda_{\mathrm{FL}} \|\mathbf{W}_m \mathbf{R}^{\mathsf{T}}\|_1 + \lambda_{\mathrm{L}} \|\mathbf{W}_m\|_1 + \lambda_{\mathrm{G}} \|\mathbf{W}_m\|_{2,1}\right]. \tag{7}$$

The relation penalty $\lambda_{\mathrm{A}} \|\mathbf{A}\|_F^2/2$ is constant in this subproblem and is therefore omitted. We minimise $f + g$ using accelerated proximal gradient with backtracking line search.

The regression gradient for objective $m$ at visit $t$ is

$$\nabla_{\mathbf{w}_{mt}} f_{\mathrm{reg}} = \frac{1}{n_{mt}} \mathbf{X}_{mt}^{\mathsf{T}} \left(\mathbf{X}_{mt}\mathbf{w}_{mt} - \mathbf{y}_{mt}\right). \tag{8}$$

For the relation term, define $\mathbf{B} = \mathbf{I} - \mathbf{A}$. Then $f_{\mathrm{rel}} = \lambda_{\mathrm{R}} \|\mathbf{U}\mathbf{B}\|_F^2/2$, with gradient

$$\nabla_{\mathbf{U}} f_{\mathrm{rel}} = \lambda_{\mathrm{R}} \mathbf{U}\mathbf{B}\mathbf{B}^{\mathsf{T}}. \tag{9}$$

Because column $m$ of $\mathbf{U}$ is the vectorised form of $\mathbf{W}_m$, its relation gradient is reshaped into a $p \times T$ matrix. The complete gradient is therefore

$$\nabla_{\mathbf{W}_m} f = \nabla_{\mathbf{W}_m} f_{\mathrm{reg}} + \mathrm{unvec}_{p \times T} \left([\nabla_{\mathbf{U}} f_{\mathrm{rel}}]_{:m}\right). \tag{10}$$

Let $\mathbf{Z}_m^{(k)}$ be the extrapolated coefficient matrix at inner iteration $k$, and let $L_k$ be the current estimate of the Lipschitz constant (Nesterov, 1983; 2003). The proximal-gradient step is

$$\mathbf{V}_m^{(k)} = \mathbf{Z}_m^{(k)} - \frac{1}{L_k} \nabla_{\mathbf{W}_m} f(\{\mathbf{Z}_m^{(k)}\}; \mathbf{A}), \tag{11}$$

followed by

$$\mathbf{W}_m^{(k+1)} = \mathrm{prox}_{g_m/L_k} \left(\mathbf{V}_m^{(k)}\right). \tag{12}$$

**Composite proximal operator.** The coefficient update involves a composite penalty consisting of fused Lasso, element-wise Lasso, and group Lasso terms. Although proximal operators cannot generally be composed in an arbitrary order, this fused sparse group-Lasso penalty admits an exact and efficient decomposition (Zhou et al., 2023b; 2012).

Specifically, the proximal problem is separable across prediction objectives and DMO rows. For objective $m$ and DMO $j$, let $\mathbf{v}_{mj} \in \mathbb{R}^T$ denote the $j$-th row of $\mathbf{V}_m^{(k)}$. The corresponding update is

$$\mathbf{w}_{mj}^{(k+1)} = \arg\min_{\mathbf{w} \in \mathbb{R}^T} \frac{1}{2} \|\mathbf{w} - \mathbf{v}_{mj}\|_2^2 + \frac{\lambda_{\mathrm{L}}}{L_k} \|\mathbf{w}\|_1 + \frac{\lambda_{\mathrm{FL}}}{L_k} \|\mathbf{R}\mathbf{w}\|_1 + \frac{\lambda_{\mathrm{G}}}{L_k} \|\mathbf{w}\|_2. \tag{13}$$

This problem is solved exactly in two stages. First, the fused and element-wise sparse solution is obtained from

$$\widetilde{\mathbf{w}}_{mj} = \arg\min_{\mathbf{w} \in \mathbb{R}^T} \frac{1}{2} \|\mathbf{w} - \mathbf{v}_{mj}\|_2^2 + \frac{\lambda_{\mathrm{L}}}{L_k} \|\mathbf{w}\|_1 + \frac{\lambda_{\mathrm{FL}}}{L_k} \|\mathbf{R}\mathbf{w}\|_1. \tag{14}$$

Second, row-wise group shrinkage is applied:

$$\mathbf{w}_{mj}^{(k+1)} = \left(1 - \frac{\lambda_{\mathrm{G}}}{L_k \|\widetilde{\mathbf{w}}_{mj}\|_2}\right)_+ \widetilde{\mathbf{w}}_{mj}, \tag{15}$$

where $(a)_+ = \max(a, 0)$. The first stage identifies temporally fused and visit-specific coefficients, while the second removes the complete DMO trajectory when its magnitude is insufficient. This decomposition exactly evaluates the composite proximal operator and can be implemented using established fast fused-Lasso algorithms.

Backtracking line search is used to select $L_k$, and Nesterov acceleration is applied after each accepted update. The inner iterations terminate when the relative change in the coefficient-block objective falls below a prescribed tolerance.

### C.2 Updating the Symmetric Relation Matrix

For fixed coefficient matrices, $\mathbf{U}$ is constant and the relation subproblem becomes

$$\begin{aligned} \min_{\mathbf{A}} \quad & \frac{\lambda_{\mathrm{R}}}{2} \|\mathbf{U} - \mathbf{U}\mathbf{A}\|_F^2 + \frac{\lambda_{\mathrm{A}}}{2} \|\mathbf{A}\|_F^2 \\ \text{s.t.} \quad & \mathbf{A} = \mathbf{A}^\mathsf{T}, \qquad \mathrm{diag}(\mathbf{A}) = \mathbf{0}. \end{aligned} \tag{16}$$

This is a strongly convex quadratic problem because $\lambda_{\mathrm{A}} > 0$. To satisfy symmetry and the zero-diagonal constraint exactly, we parameterise only the $K = M(M-1)/2$ upper-triangular entries.

For each pair $i < j$, define $\mathbf{E}_{ij} = \mathbf{e}_i \mathbf{e}_j^\mathsf{T} + \mathbf{e}_j \mathbf{e}_i^\mathsf{T}$, where $\mathbf{e}_i$ is the $i$-th standard basis vector in $\mathbb{R}^M$. Any feasible relation matrix can then be written as $\mathbf{A}(\mathbf{a}) = \sum_{i<j} a_{ij} \mathbf{E}_{ij}$, where $\mathbf{a} \in \mathbb{R}^K$. Let $\mathbf{b} = \mathrm{vec}(\mathbf{U})$ and define

$$\mathbf{C} = \left[\mathrm{vec}(\mathbf{U}\mathbf{E}_{12}), \mathrm{vec}(\mathbf{U}\mathbf{E}_{13}), \ldots, \mathrm{vec}(\mathbf{U}\mathbf{E}_{M-1,M})\right]. \tag{17}$$

Since $\mathrm{vec}(\mathbf{U}\mathbf{A}(\mathbf{a})) = \mathbf{C}\mathbf{a}$ and $\|\mathbf{A}(\mathbf{a})\|_F^2 = 2\|\mathbf{a}\|_2^2$, the constrained matrix problem reduces to

$$\min_{\mathbf{a} \in \mathbb{R}^K} \quad \frac{\lambda_{\mathrm{R}}}{2} \|\mathbf{b} - \mathbf{C}\mathbf{a}\|_2^2 + \lambda_{\mathrm{A}} \|\mathbf{a}\|_2^2. \tag{18}$$

Its unique minimiser is obtained from

$$\left(\lambda_{\mathrm{R}} \mathbf{C}^\mathsf{T}\mathbf{C} + 2\lambda_{\mathrm{A}} \mathbf{I}\right) \mathbf{a} = \lambda_{\mathrm{R}} \mathbf{C}^\mathsf{T} \mathbf{b}. \tag{19}$$

The relation matrix is then formed as $\mathbf{A} = \sum_{i<j} a_{ij} \mathbf{E}_{ij}$. This parameterisation satisfies both constraints by construction. In this study, $M = 5$ and hence $K = 10$, so the update can be computed efficiently using a standard linear-system solver.

### C.3 Alternating Optimisation Algorithm

The complete optimisation procedure is summarised in Algorithm 1. We initialise the relation matrix as $\mathbf{A}^{(0)} = \mathbf{0}$. The coefficient matrices may be initialised either at zero or using independently fitted visit-specific regression models. During subsequent outer iterations, each coefficient subproblem is warm-started from the solution obtained in the preceding iteration.

The outer iterations terminate when the relative change in the complete objective satisfies

$$\frac{\left|\mathcal{J}^{(r+1)} - \mathcal{J}^{(r)}\right|}{\max\left(1, \left|\mathcal{J}^{(r)}\right|\right)} \leq \varepsilon, \tag{20}$$

where $\mathcal{J}^{(r)}$ denotes the unified objective value at outer iteration $r$. A maximum number of outer iterations is imposed as an additional safeguard.

**Algorithm 1** Alternating Optimisation of the Longitudinal Coefficient Matrices and Cross-Objective Relation Matrix.

**Require:** Training data $\{\mathbf{X}_{mt}, \mathbf{y}_{mt}\}_{m=1,t=1}^{M,T}$; regularisation parameters $\lambda_{\mathrm{FL}}, \lambda_{\mathrm{L}}, \lambda_{\mathrm{G}}, \lambda_{\mathrm{R}}, \lambda_{\mathrm{A}}$; maximum number of outer iterations $R_{\max}$

Initialise $\{\mathbf{W}_m^{(0)}\}_{m=1}^{M}$ and $\mathbf{A}^{(0)} = \mathbf{0}$
**for** $r = 0, \ldots, R_{\max} - 1$ **do**
  Fix $\mathbf{A}^{(r)}$ and update $\{\mathbf{W}_m^{(r+1)}\}_{m=1}^{M}$ using accelerated proximal-gradient iterations with backtracking line search
  Evaluate the composite proximal operator row-wise using fused-Lasso signal approximation followed by group shrinkage
  Construct $\mathbf{U}^{(r+1)}$ by vectorising $\{\mathbf{W}_m^{(r+1)}\}_{m=1}^{M}$
  Construct $\mathbf{C}^{(r+1)}$ and $\mathbf{b}^{(r+1)}$ from $\mathbf{U}^{(r+1)}$
  Solve

$$\left(\lambda_{\mathrm{R}}\mathbf{C}^{\mathsf{T}}\mathbf{C} + 2\lambda_{\mathrm{A}}\mathbf{I}\right)\mathbf{a} = \lambda_{\mathrm{R}}\mathbf{C}^{\mathsf{T}}\mathbf{b}$$

  Form the symmetric, zero-diagonal relation matrix $\mathbf{A}^{(r+1)}$ from $\mathbf{a}$
  **if** the outer convergence criterion is satisfied **then**
    **break**
  **end if**
**end for**

**Ensure:** $\{\mathbf{W}_m\}_{m=1}^{M}$ and $\mathbf{A}$

### C.3.1 Convergence and Computational Complexity

For fixed $\mathbf{A}$, the coefficient subproblem is convex but non-smooth. Because its composite proximal operator can be computed exactly, the accelerated proximal-gradient procedure with backtracking converges to the global minimiser of this subproblem. For fixed coefficient matrices, the relation-matrix subproblem is strongly convex when $\lambda_{\mathrm{A}} > 0$ and therefore has a unique minimiser.

If both block subproblems are solved exactly, or to sufficient numerical accuracy, each outer iteration does not increase the unified objective. Because the objective is bounded below, the sequence of objective values converges. Under standard regularity conditions, every accumulation point of the alternating sequence is a block-coordinate stationary point. However, the complete objective is biconvex rather than jointly convex because of the interaction between $\mathbf{U}$ and $\mathbf{A}$. Convergence to a global minimiser is therefore not guaranteed and may depend on the initialisation.

For the coefficient update, evaluating the regression gradients across all objectives and visits requires

$$\mathcal{O}\left(p\sum_{m=1}^{M}\sum_{t=1}^{T} n_{mt}\right)$$

operations. The relation-gradient term involves multiplication between the $pT \times M$ coefficient representation and the $M \times M$ relation matrices, requiring $\mathcal{O}\left(pTM^2\right)$ operations.

The composite proximal operator is separable across the $Mp$ longitudinal coefficient trajectories. When a linear-time one-dimensional fused-Lasso solver is used, its overall cost is

$$\mathcal{O}\left(MpT\right).$$

Thus, if $I_{\mathrm{in}}$ accelerated proximal-gradient iterations are required, the coefficient-update cost per outer iteration is approximately

$$\mathcal{O}\left[I_{\mathrm{in}}\left(p\sum_{m=1}^{M}\sum_{t=1}^{T} n_{mt} + pTM^2 + MpT\right)\right]. \tag{21}$$

The relation update contains $K = M(M-1)/2$ free edge variables. Constructing its quadratic system depends on the $pT$-dimensional representations of the $M$ objectives, while solving the resulting $K \times K$ linear system by a direct method requires $\mathcal{O}(K^3)$ operations. Because $K$ depends only on the number of prediction objectives, this update remains inexpensive when $M$ is small.

In the present study, $p = 24$, $T = 5$, $M = 5$, and $K = 10$. Consequently, the composite proximal operations and relation-matrix update are small, and the overall computational cost is dominated by repeated regression-gradient evaluations across participants, visits, and prediction objectives.

# D Non-Convex Extensions of DeMMO

The fused-Lasso, Lasso, and group-Lasso regularisers in DeMMO provide a convex and computationally tractable formulation for temporal smoothing and DMO selection. However, convex sparsity penalties may over-shrink the coefficients of informative DMOs, particularly when their effects are weak, correlated, or confined to specific visits, thereby introducing estimation bias (Fan & Li, 2001). Such shrinkage may reduce both predictive performance and the ability to recover clinically meaningful longitudinal patterns.

Motivated by non-convex extensions of the fused sparse-group Lasso (Zhou et al., 2012), we introduce two variants, DeMMO-var1 and DeMMO-var2. Both use a concave square-root penalty to reduce shrinkage of sufficiently strong DMO effects while retaining sparsity. They differ in how DMO selection and temporal smoothness are organised. DeMMO-var1 treats them as separate components, allowing their strengths to be controlled independently. DeMMO-var2 couples them within a single DMO-level penalty, such that retention of a DMO depends jointly on its coefficient magnitude and temporal variation. Comparing these variants allows us to examine whether longitudinal DMO modelling benefits more from independent or coupled regularisation.

For concise notation, let $\mathbf{w}_{d,q,j:} \in \mathbb{R}^T$ denote the coefficients of DMO $j$ across all visits for disease $d$ and outcome $q$. Both variants retain the longitudinal regression loss $\mathcal{L}(\mathcal{W})$ and the relation-learning regulariser

$$\mathcal{R}_{\text{rel}}(\mathbf{U}, \mathbf{A}) = \frac{\lambda_{\text{R}}}{2}\|\mathbf{U} - \mathbf{U}\mathbf{A}\|_F^2 + \frac{\lambda_{\text{A}}}{2}\|\mathbf{A}\|_F^2. \tag{22}$$

## D.1 DeMMO-var1: Separate Selection and Temporal Smoothing

DeMMO-var1 replaces the Lasso and group-Lasso penalties with a composite $\ell_{(0.5,1)}$ penalty while retaining fused temporal regularisation as a separate component:

$$\begin{aligned} \min_{\mathcal{W},\mathbf{A}} \quad & \mathcal{L}(\mathcal{W}) + \lambda_{\text{NC}} \sum_{d,q,j} \sqrt{\|\mathbf{w}_{d,q,j:}\|_1} + \lambda_{\text{FL}} \sum_{d,q} \|\mathbf{W}_{d,q}\mathbf{R}^\top\|_1 + \mathcal{R}_{\text{rel}}(\mathbf{U}, \mathbf{A}) \\ \text{s.t.} \quad & \mathbf{A} = \mathbf{A}^\top, \qquad \text{diag}(\mathbf{A}) = \mathbf{0}. \end{aligned} \tag{23}$$

The outer square root promotes DMO-level sparsity, whereas the inner $\ell_1$ norm permits visit-specific coefficients to be removed. Temporal smoothness is controlled independently by $\lambda_{\text{FL}}$.

## D.2 DeMMO-var2: Coupled Selection and Temporal Smoothing

DeMMO-var2 instead incorporates coefficient sparsity and temporal variation within the same DMO-level penalty. Define

$$\phi_{d,q,j}(\mathcal{W}) = \|\mathbf{w}_{d,q,j:}\mathbf{R}^\top\|_1 + \beta\|\mathbf{w}_{d,q,j:}\|_1, \tag{24}$$

where $\beta$ controls their relative contributions. The objective becomes

$$\begin{aligned} \min_{\mathcal{W},\mathbf{A}} \quad & \mathcal{L}(\mathcal{W}) + \lambda_{\text{NC}} \sum_{d,q,j} \sqrt{\phi_{d,q,j}(\mathcal{W})} + \mathcal{R}_{\text{rel}}(\mathbf{U}, \mathbf{A}) \\ \text{s.t.} \quad & \mathbf{A} = \mathbf{A}^\top, \qquad \text{diag}(\mathbf{A}) = \mathbf{0}. \end{aligned} \tag{25}$$

Unlike DeMMO-var1, DeMMO-var2 determines whether to retain a DMO according to both the magnitude and temporal variation of its longitudinal coefficients.

## D.3 Optimisation

Unlike convex DeMMO, the two variants require an additional difference-of- convex outer loop. We use the iterative reweighting procedure established for non-convex fused sparse-group Lasso models

**Algorithm 2** Non-Convex DeMMO Optimisation.

**Require:** Longitudinal data, variant parameters, $\epsilon$, and tolerances
1: Initialise $(\mathcal{W}^{(0)}, \mathbf{A}^{(0)})$ from convex DeMMO
2: **for** $k = 0, 1, \ldots, K_{\mathrm{DC}} - 1$ **do**
3: Compute $a_{d,q,j}^{(k)}$ for var1 or $b_{d,q,j}^{(k)}$ for var2 using Eq. (27)
4: **repeat**
5: Update $\mathcal{W}$ by accelerated proximal gradient on Eq. (28) or Eq. (29)
6: Update the symmetric zero-diagonal $\mathbf{A}$ by solving the relation-learning linear system
7: **until** the convex majoriser satisfies the block-convergence criterion
8: Evaluate the original non-convex objective
9: **if** its relative change is below $\varepsilon_{\mathrm{DC}}$ **then**
10: **break**
11: **end if**
12: **end for**
**Ensure:** $\mathcal{W}$ and $\mathbf{A}$

(Zhou et al., 2012; 2023b). For a non-negative quantity $s$, the concavity of the square root gives the majoriser

$$\sqrt{s+\epsilon} \le \sqrt{s^{(k)}+\epsilon} + \frac{s - s^{(k)}}{2\sqrt{s^{(k)}+\epsilon}}, \tag{26}$$

where equality holds at $s = s^{(k)}$. Thus, at reweighting iteration $k$, the non-convex penalties are replaced by weighted convex $\ell_1$ and fused-Lasso penalties. Define

$$a_{d,q,j}^{(k)} = \frac{\lambda_{\mathrm{NC}}}{2\sqrt{\|\mathbf{w}_{d,q,j:}^{(k)}\|_1 + \epsilon}}, \qquad b_{d,q,j}^{(k)} = \frac{\lambda_{\mathrm{NC}}}{2\sqrt{\phi_{d,q,j}(\mathcal{W}^{(k)}) + \epsilon}}, \tag{27}$$

where $\epsilon > 0$ prevents unbounded weights. Ignoring terms independent of $\mathcal{W}$, the coefficient majorisers for DeMMO-var1 and DeMMO-var2 are, respectively,

$$\mathcal{Q}_{\mathrm{var1}}^{(k)} = \mathcal{L}(\mathcal{W}) + \sum_{d,q,j} a_{d,q,j}^{(k)} \|\mathbf{w}_{d,q,j:}\|_1 + \lambda_{\mathrm{FL}} \sum_{d,q,j} \|\mathbf{w}_{d,q,j:}\mathbf{R}^{\mathsf{T}}\|_1 + \mathcal{R}_{\mathrm{rel}}, \tag{28}$$

$$\mathcal{Q}_{\mathrm{var2}}^{(k)} = \mathcal{L}(\mathcal{W}) + \sum_{d,q,j} b_{d,q,j}^{(k)} \left(\|\mathbf{w}_{d,q,j:}\mathbf{R}^{\mathsf{T}}\|_1 + \beta\|\mathbf{w}_{d,q,j:}\|_1\right) + \mathcal{R}_{\mathrm{rel}}. \tag{29}$$

Both are weighted convex fused-Lasso problems. For fixed $\mathbf{A}$, we solve the coefficient block using accelerated proximal-gradient iterations with backtracking. The proximal step is separable across DMO trajectories and uses the same one-dimensional fused-Lasso solver as convex DeMMO, but with the row-specific weights in Eq. (27). For fixed $\mathcal{W}$, the symmetric relation matrix is updated using the same closed-form quadratic subproblem described in Section C.2.

Algorithm 2 summarises the complete procedure. We initialise both variants from the fitted convex DeMMO solution. This warm start provides a stable coefficient pattern and avoids starting the reweighting scheme near the singular point of the square-root penalty.

The majorisation step is tight and each inner block update decreases its convex surrogate. Consequently, the original objective is non-increasing and the procedure converges to a stationary point under the usual boundedness and accurate-subproblem assumptions. Global optimality is not guaranteed because the objectives remain non-convex.

**Computational cost.** Let $K_{\mathrm{DC}}$ be the number of reweighting iterations, $K_{\mathrm{B}}$ the number of alternating block updates per majoriser, and $K_{\mathrm{PG}}$ the number of proximal-gradient iterations. One gradient/proximal update has the same leading cost as convex DeMMO,

$$\mathcal{O}\left(p \sum_{m=1}^{M} \sum_{t=1}^{T} n_{mt} + pTM^2 + MpT\right). \tag{30}$$

The complete coefficient cost is therefore

$$\mathcal{O}\left[K_{\mathrm{DC}} K_{\mathrm{B}} K_{\mathrm{PG}} \left(p \sum_{m,t} n_{mt} + pTM^2 + MpT\right)\right], \tag{31}$$

Table 4: Comparison of the two non-convex DeMMO variants over five matched participant splits. Values are mean ± standard deviation. T1–T5 report visit-specific RMSE. The best mean between the two variants is shown in bold.

| Objective | Method | nMSE ↓ | wR ↑ | T1 | T2 | T3 | T4 | T5 |
|---|---|---|---|---|---|---|---|---|
| Overall | DeMMO-var1 | 0.773 ± 0.035 | 0.459 ± 0.040 | – | – | – | – | – |
| | DeMMO-var2 | **0.772 ± 0.033** | **0.459 ± 0.038** | – | – | – | – | – |
| PD H&Y | DeMMO-var1 | 0.946 ± 0.039 | 0.291 ± 0.058 | 0.536 ± 0.036 | **0.504 ± 0.058** | 0.549 ± 0.043 | **0.522 ± 0.078** | 0.555 ± 0.072 |
| | DeMMO-var2 | **0.944 ± 0.036** | **0.291 ± 0.056** | **0.536 ± 0.035** | 0.506 ± 0.058 | **0.546 ± 0.039** | 0.522 ± 0.077 | **0.552 ± 0.069** |
| PD MDS–UPDRS | DeMMO-var1 | **0.889 ± 0.051** | **0.364 ± 0.059** | **11.930 ± 0.475** | 11.689 ± 0.106 | 12.107 ± 0.697 | 11.492 ± 0.787 | 11.519 ± 0.769 |
| | DeMMO-var2 | 0.889 ± 0.048 | 0.358 ± 0.053 | 11.991 ± 0.533 | **11.689 ± 0.159** | **12.064 ± 0.639** | **11.480 ± 0.764** | **11.497 ± 0.737** |
| MS EDSS | DeMMO-var1 | **0.485 ± 0.054** | **0.727 ± 0.039** | **0.820 ± 0.082** | 0.830 ± 0.058 | **0.991 ± 0.061** | 1.013 ± 0.033 | **1.093 ± 0.086** |
| | DeMMO-var2 | 0.487 ± 0.049 | 0.727 ± 0.035 | 0.822 ± 0.078 | **0.828 ± 0.058** | 0.997 ± 0.063 | **1.010 ± 0.046** | 1.098 ± 0.090 |
| PFF SPPB | DeMMO-var1 | **0.547 ± 0.060** | **0.681 ± 0.044** | **1.997 ± 0.058** | 2.213 ± 0.189 | 2.262 ± 0.221 | **2.213 ± 0.327** | **2.221 ± 0.165** |
| | DeMMO-var2 | 0.550 ± 0.065 | 0.679 ± 0.047 | 2.016 ± 0.053 | **2.205 ± 0.187** | **2.249 ± 0.221** | 2.229 ± 0.330 | 2.256 ± 0.133 |
| COPD $FEV_1$ | DeMMO-var1 | 0.920 ± 0.031 | 0.310 ± 0.042 | 18.870 ± 0.320 | 20.264 ± 0.641 | 19.179 ± 1.203 | 19.778 ± 1.036 | **20.579 ± 0.608** |
| | DeMMO-var2 | **0.914 ± 0.031** | **0.317 ± 0.046** | **18.787 ± 0.331** | **20.158 ± 0.686** | **19.030 ± 1.307** | **19.758 ± 1.025** | 20.643 ± 0.554 |

in addition to the small relation-matrix solves. Convex DeMMO requires only one alternating optimisation sequence, whereas each non-convex variant solves a sequence of weighted convex DeMMO-like problems. In our implementation, $K_{\mathrm{DC}}$, $K_{\mathrm{B}}$, and $K_{\mathrm{PG}}$ are capped at 15, 12, and 300, respectively, with early stopping at every level. The convex warm start substantially reduces the realised number of iterations, but the two variants remain more computationally expensive. Their additional predictive flexibility therefore comes at the cost of longer model selection and training, which is the principal trade-off examined in our experiments.

# E Experimental Evaluation of the Non-Convex Variants

Table 4 reports the results of DeMMO-var1 and DeMMO-var2. The corresponding results for the original DeMMO are reported in Table 2 and Tables 5–9.

DeMMO performs best overall. It achieves an nMSE of $0.769 \pm 0.021$, compared with $0.773 \pm 0.035$ for DeMMO-var1 and $0.772 \pm 0.033$ for DeMMO-var2, while its wR of $0.464 \pm 0.023$ exceeds $0.459 \pm 0.040$ and $0.459 \pm 0.038$, respectively. DeMMO also obtains the best nMSE and wR for both PD outcomes and MS EDSS. DeMMO-var1 is marginally better for PFF SPPB, whereas DeMMO-var2 is marginally better for COPD $FEV_1$. At the visit level, DeMMO gives the lowest RMSE in 12 of the 25 tasks, compared with seven for DeMMO-var1 and six for DeMMO-var2.

The two variants retain DeMMO's outcome-specific sparsity, temporal continuity, and cross-objective relation learning, but use adaptive non-convex reweighting to reduce the estimation bias induced by the convex Lasso, group-Lasso, and fused-Lasso penalties. Their lack of a consistent aggregate improvement suggests that shrinkage-induced bias is not a major limitation in this setting. We conjecture that this result reflects the substantial behavioural and measurement noise in free-living DMOs. By weakening shrinkage, the non-convex penalties may retain more noise together with the DMO signal, leading to poorer generalisation; the stronger shrinkage of convex DeMMO may instead provide beneficial regularisation. Convex DeMMO also shows lower overall variability and requires substantially less computation because it avoids the additional difference-of-convex iterations. These results favour the original DeMMO formulation for this noisy longitudinal DMO setting.

# F Additional Experimental Results

## F.1 Hyperparameter Sensitivity

We vary each parameter over $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ while fixing the others at $10^{-3}$, and report mean validation nMSE across the 25 outcome–visit tasks for seed 42 (Fig. 5); the test set is not used. All five penalties affect validation performance. The sparsity penalties are most influential and become excessive at 1, $\lambda_{\mathrm{FL}}$ favours moderate temporal smoothing, and the U-shaped $\lambda_{\mathrm{R}}$ curve supports selective rather than unrestricted cross-objective sharing. Performance is comparatively insensitive to $\lambda_{\mathrm{A}}$ over the tested range, likely because standardising the predictors and outcomes places the coefficient matrices across disease–outcome objectives on comparable scales.

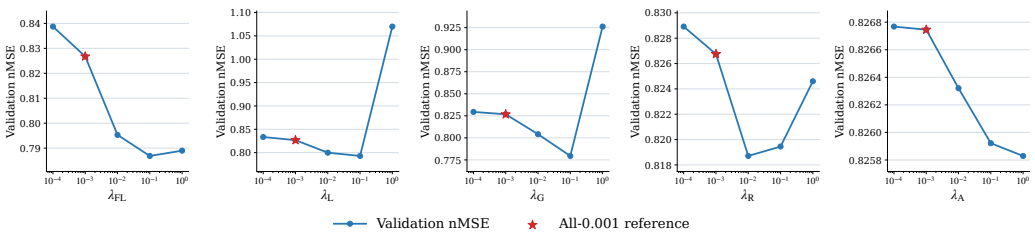


Figure 5: One-at-a-time sensitivity of validation nMSE; all other parameters are fixed at 0.001, with the shared reference marked by a red star.

Table 5: Visit-specific RMSE for PD H&Y (mean ± standard deviation over five matched participant splits). The best mean in each column is shown in bold.

| Method | T1 | T2 | T3 | T4 | T5 |
|---|---|---|---|---|---|
| cFSGL | 0.550 ± 0.023 | 0.510 ± 0.042 | 0.557 ± 0.029 | 0.535 ± 0.051 | 0.558 ± 0.037 |
| FRoTS | 0.543 ± 0.023 | 0.514 ± 0.040 | 0.550 ± 0.027 | 0.531 ± 0.042 | 0.573 ± 0.049 |
| MAGPP | 0.544 ± 0.021 | 0.522 ± 0.043 | 0.552 ± 0.030 | 0.543 ± 0.040 | 0.586 ± 0.053 |
| MLP-MSE | 0.556 ± 0.028 | 0.518 ± 0.040 | 0.560 ± 0.032 | 0.542 ± 0.046 | 0.565 ± 0.059 |
| MLP-L1 | 0.571 ± 0.026 | 0.532 ± 0.034 | 0.572 ± 0.026 | 0.542 ± 0.047 | 0.567 ± 0.036 |
| Rank-N | 0.577 ± 0.027 | 0.532 ± 0.034 | 0.583 ± 0.033 | 0.539 ± 0.050 | 0.568 ± 0.035 |
| RankSim | 0.556 ± 0.023 | 0.519 ± 0.038 | 0.558 ± 0.030 | 0.549 ± 0.041 | 0.558 ± 0.047 |
| ACCon | 0.562 ± 0.019 | 0.528 ± 0.038 | 0.563 ± 0.028 | 0.545 ± 0.052 | 0.569 ± 0.052 |
| DeMMO | **0.538 ± 0.022** | **0.504 ± 0.038** | **0.542 ± 0.025** | **0.524 ± 0.052** | **0.543 ± 0.045** |

Table 6: Visit-specific RMSE for PD MDS–UPDRS (mean ± standard deviation over five matched participant splits). The best mean in each column is shown in bold.

| Method | T1 | T2 | T3 | T4 | T5 |
|---|---|---|---|---|---|
| cFSGL | 11.959 ± 0.227 | 11.852 ± 0.077 | 12.299 ± 0.472 | 11.742 ± 0.527 | 11.488 ± 0.227 |
| FRoTS | 11.946 ± 0.328 | 11.700 ± 0.258 | 12.333 ± 0.500 | 11.674 ± 0.515 | 11.740 ± 0.489 |
| MAGPP | **11.925 ± 0.335** | 11.901 ± 0.368 | 12.517 ± 0.517 | 11.997 ± 0.485 | 11.722 ± 0.424 |
| MLP-MSE | 12.161 ± 0.213 | 11.847 ± 0.166 | 12.318 ± 0.636 | 11.865 ± 0.693 | 11.785 ± 0.375 |
| MLP-L1 | 12.424 ± 0.400 | 12.125 ± 0.426 | 12.256 ± 0.465 | 12.373 ± 0.665 | 12.087 ± 0.330 |
| Rank-N | 12.213 ± 0.213 | 12.162 ± 0.187 | 12.708 ± 0.931 | 12.165 ± 0.819 | 12.206 ± 0.454 |
| RankSim | 12.214 ± 0.338 | 11.953 ± 0.172 | 12.503 ± 0.863 | 11.877 ± 0.598 | 11.846 ± 0.371 |
| ACCon | 12.287 ± 0.258 | 12.009 ± 0.128 | 12.425 ± 0.719 | 12.285 ± 0.686 | 12.260 ± 0.949 |
| DeMMO | 12.073 ± 0.260 | **11.676 ± 0.055** | **12.002 ± 0.489** | **11.428 ± 0.524** | **11.339 ± 0.430** |

Tables 5–9 provide the numerical results underlying Figure 2.

**Outcome-specific observations.** For PD H&Y, DeMMO is best or tied for best at all five visits, with its clearest advantages at T2–T4. For PD MDS–UPDRS, it achieves the lowest mean RMSE at T2–T5, while MAGPP performs best at T1. DeMMO is also consistently competitive for MS EDSS: it is best at T1–T3 and T5 and remains close to the best result at T4. The PFF task exhibits a different pattern. cFSGL is strongest at T1 and RankSim at T2, whereas DeMMO becomes best from T3 to T5. This later-visit advantage is particularly relevant because the PFF sample size decreases sharply across follow-up visits, suggesting that temporal and cross-objective information sharing provides useful regularisation when outcome-specific data are sparse. For COPD $FEV_1$, DeMMO achieves the lowest mean RMSE at all five visits, demonstrating that its advantage also extends to pulmonary-function prediction.

**Patterns across visits and methods.** Across the five outcomes, DeMMO achieves the lowest mean RMSE in 21 of the 25 outcome–visit comparisons. Its performance is not uniformly dominant,

Table 7: Visit-specific RMSE for MS EDSS (mean ± standard deviation over five matched participant splits). The best mean in each column is shown in bold.

| Method | T1 | T2 | T3 | T4 | T5 |
|---|---|---|---|---|---|
| cFSGL | 0.826 ± 0.053 | 0.845 ± 0.038 | 0.998 ± 0.041 | 1.023 ± 0.028 | 1.108 ± 0.053 |
| FRoTS | 0.832 ± 0.047 | 0.841 ± 0.043 | 0.997 ± 0.043 | 1.022 ± 0.039 | 1.132 ± 0.061 |
| MAGPP | 0.831 ± 0.047 | 0.856 ± 0.044 | 1.014 ± 0.045 | 1.035 ± 0.045 | 1.153 ± 0.060 |
| MLP-MSE | 0.843 ± 0.054 | 0.851 ± 0.053 | 1.018 ± 0.036 | 1.001 ± 0.021 | 1.167 ± 0.075 |
| MLP-L1 | 0.861 ± 0.052 | 0.869 ± 0.056 | 1.028 ± 0.039 | 1.038 ± 0.052 | 1.137 ± 0.058 |
| Rank-N | 0.845 ± 0.058 | 0.892 ± 0.046 | 1.050 ± 0.033 | 1.064 ± 0.037 | 1.159 ± 0.051 |
| RankSim | 0.837 ± 0.054 | 0.854 ± 0.048 | 0.996 ± 0.031 | **0.997 ± 0.013** | 1.165 ± 0.047 |
| ACCon | 0.844 ± 0.061 | 0.857 ± 0.047 | 1.037 ± 0.034 | 1.020 ± 0.012 | 1.176 ± 0.078 |
| DeMMO | **0.818 ± 0.052** | **0.823 ± 0.039** | **0.984 ± 0.038** | 1.005 ± 0.032 | **1.098 ± 0.062** |

Table 8: Visit-specific RMSE for PFF SPPB (mean ± standard deviation over five matched participant splits). The best mean in each column is shown in bold.

| Method | T1 | T2 | T3 | T4 | T5 |
|---|---|---|---|---|---|
| cFSGL | **1.961 ± 0.026** | 2.251 ± 0.119 | 2.293 ± 0.101 | 2.299 ± 0.157 | 2.301 ± 0.106 |
| FRoTS | 1.967 ± 0.052 | 2.228 ± 0.142 | 2.264 ± 0.075 | 2.489 ± 0.077 | 2.389 ± 0.140 |
| MAGPP | 1.996 ± 0.051 | 2.225 ± 0.165 | 2.299 ± 0.079 | 2.539 ± 0.077 | 2.433 ± 0.144 |
| MLP-MSE | 1.979 ± 0.060 | 2.215 ± 0.116 | 2.379 ± 0.091 | 2.484 ± 0.145 | 2.471 ± 0.122 |
| MLP-L1 | 2.005 ± 0.038 | 2.217 ± 0.099 | 2.379 ± 0.150 | 2.494 ± 0.172 | 2.507 ± 0.132 |
| Rank-N | 2.028 ± 0.069 | 2.261 ± 0.170 | 2.322 ± 0.098 | 2.479 ± 0.086 | 2.517 ± 0.082 |
| RankSim | 1.996 ± 0.060 | **2.213 ± 0.095** | 2.332 ± 0.105 | 2.470 ± 0.117 | 2.449 ± 0.092 |
| ACCon | 2.021 ± 0.059 | 2.224 ± 0.142 | 2.320 ± 0.053 | 2.544 ± 0.115 | 2.553 ± 0.156 |
| DeMMO | 2.019 ± 0.036 | 2.221 ± 0.121 | **2.228 ± 0.109** | **2.236 ± 0.191** | **2.265 ± 0.104** |

Table 9: Visit-specific RMSE for COPD $FEV_1$ (mean ± standard deviation over five matched participant splits). The best mean in each column is shown in bold.

| Method | T1 | T2 | T3 | T4 | T5 |
|---|---|---|---|---|---|
| cFSGL | 18.927 ± 0.135 | 20.601 ± 0.421 | 19.472 ± 0.809 | 20.009 ± 0.756 | 20.835 ± 0.334 |
| FRoTS | 19.383 ± 0.274 | 20.828 ± 0.706 | 19.197 ± 0.901 | 20.331 ± 0.884 | 21.238 ± 0.257 |
| MAGPP | 19.424 ± 0.171 | 20.885 ± 0.720 | 19.337 ± 0.980 | 20.558 ± 1.078 | 21.562 ± 0.176 |
| MLP-MSE | 19.231 ± 0.739 | 20.899 ± 0.684 | 20.476 ± 0.971 | 20.320 ± 0.735 | 20.851 ± 0.589 |
| MLP-L1 | 19.241 ± 0.393 | 21.582 ± 0.892 | 20.104 ± 0.986 | 20.795 ± 1.192 | 21.144 ± 0.490 |
| Rank-N | 19.289 ± 0.527 | 21.263 ± 0.592 | 20.013 ± 0.566 | 20.538 ± 0.715 | 21.078 ± 0.321 |
| RankSim | 19.187 ± 0.605 | 20.921 ± 0.478 | 19.765 ± 0.723 | 20.192 ± 0.826 | 20.658 ± 0.603 |
| ACCon | 19.491 ± 0.661 | 21.229 ± 0.738 | 20.004 ± 0.659 | 20.706 ± 0.854 | 20.896 ± 0.518 |
| DeMMO | **18.814 ± 0.152** | **20.237 ± 0.437** | **19.072 ± 0.887** | **19.847 ± 0.654** | **20.478 ± 0.521** |

however: other methods lead at PD MDS–UPDRS T1, MS EDSS T4, and PFF T1–T2. Errors for MS EDSS and PFF generally increase at later visits, consistent with their declining sample sizes and potentially greater follow-up heterogeneity, whereas both PD outcomes show non-monotonic trajectories. Deep regression and ranking methods rarely lead these visit-specific comparisons and often exhibit larger variability at later visits. Overall, the tables support selective rather than uniform information sharing: DeMMO preserves outcome-specific behaviour while gaining most where longitudinal observations become limited.

# G Potential Application: Interpretable Feedback for Longitudinal Mobility Monitoring

One potential application of DeMMO is an interpretable software system for longitudinal mobility monitoring. Such software could process repeated wearable measurements, use DeMMO to estimate a participant's clinical status at successive visits, and flag a possible decline when the estimated severity increases or crosses a predefined monitoring boundary. The important question is then not only *whether* the software detects a change, but also *how* it explains that change to the user. The following retrospective analysis illustrates this potential workflow. It does not constitute a deployed product or a validated clinical alert system.

**From DeMMO discovery to software feedback.** DeMMO assigns a visit-specific coefficient to every DMO and uses stability selection to identify DMOs that are repeatedly selected across resampled datasets and hyperparameter settings. For MS EDSS, the two most stable DMOs are walking-speed P90 and cadence P90 in walking bouts longer than 30 seconds, with selection probabilities of $0.878$ and $0.874$, respectively. A walking bout is a continuous period during which a participant is detected as walking without a substantial interruption. Walking-speed P90 and cadence P90 are the 90th percentiles of walking speed and cadence, respectively, calculated from such bouts and measured in m/s and steps/min. They reflect relatively high levels of speed and cadence achieved during sustained real-world walking, rather than average walking behaviour. Their Pearson and Spearman correlations are $0.766$ and $0.778$, indicating strongly related but non-identical information.

In principle, a complete explanation could report the contribution of all 24 DMOs used by DeMMO. However, such an explanation would be lengthy, include several correlated mobility measures, and place a substantial cognitive burden on users who may not have computational or clinical expertise. Listing every contributing variable could therefore obscure the main reason for an alert rather than make it more understandable. We instead use stability selection to prioritise a small number of repeatedly identified DMOs and present the two strongest as a concise user-facing explanation. This is a deliberate trade-off between completeness and comprehensibility: the full DeMMO model still use all 24 DMOs in the background, while the explanation layer summarises the most stable evidence in a form that a user can readily interpret. The two-DMO explanation should consequently be understood as a compact summary, not a claim that the remaining DMOs are irrelevant.

In a potential software pipeline, DeMMO would first indicate that the participant's estimated status had worsened across visits. The software could then examine whether these two stable DMOs had also declined or entered ranges associated with greater impairment, providing a concise explanation for the alert. We use a deliberately simple, class-balanced decision tree to examine whether the two DMOs can support such an explanation. The analysis follows the main experiment's participant-level $70\%/10\%/20\%$ training/validation/test protocol with seeds 42–46; all visits from one participant remain in the same split. Maximum depth ($1, 2, 3, 4,$ or $5$) and minimum leaf size ($5, 10, 20,$ or $30$) are selected by validation balanced accuracy. The selected tree is then refitted on the combined training and validation participants and evaluated once on the held-out test participants.

**Binary classification of severe disability.** We first consider a clinically focused binary task separating severe disability (EDSS $\geq 6.0$) from non-severe disability (EDSS $< 6.0$). Higher EDSS indicates greater impairment, and EDSS 6 marks the stage at which a walking aid is required for ambulation (Kurtzke, 1983). The analytical sample contains 2,083 participant–visit observations from 578 participants, comprising 1,157 non-severe and 926 severe observations.

The two-DMO tree achieves balanced accuracy of $0.789 \pm 0.031$ and macro-F1 of $0.785 \pm 0.032$. It correctly identifies $84.2\%$ of severe observations and $73.7\%$ of non-severe observations. Among observations predicted as severe, $73.1\%$ are truly severe. The principal error is that $26.4\%$ of non-severe observations are flagged as severe, whereas $15.8\%$ of severe observations are missed. The always-most-common-class baseline, which assigns every observation to the larger non-severe class without using either DMO, has balanced accuracy of $0.5$ and cannot identify any severe observations.

**Potential feedback after a detected decline.** The leading divisions of the binary trees are reasonably consistent. Three of the five fitted trees first divide walking-speed P90 at $0.945$ m/s; the other two first divide cadence P90 at $97.5$ or $99.5$ steps/min. Within this cohort, lower values of these

Table 10: Binary MS EDSS classification into non-severe ($< 6.0$) and severe ($\geq 6.0$) disability using walking-speed P90 and cadence P90. Values are mean $\pm$ standard deviation over five participant-level test splits.

| Method | Class | Balanced accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| *Overall class-balanced performance* | | | | | |
| Always most common class | Macro average | 0.500 $\pm$ 0.000 | 0.271 $\pm$ 0.008 | 0.500 $\pm$ 0.000 | 0.351 $\pm$ 0.007 |
| Two-DMO decision tree | Macro average | 0.789 $\pm$ 0.031 | 0.790 $\pm$ 0.034 | 0.789 $\pm$ 0.031 | 0.785 $\pm$ 0.032 |
| *Class-specific performance* | | | | | |
| Always most common class | Non-severe | – | 0.542 $\pm$ 0.016 | 1.000 $\pm$ 0.000 | 0.703 $\pm$ 0.014 |
| | Severe | – | 0.000 $\pm$ 0.000 | 0.000 $\pm$ 0.000 | 0.000 $\pm$ 0.000 |
| Two-DMO decision tree | Non-severe | – | 0.850 $\pm$ 0.041 | 0.737 $\pm$ 0.047 | 0.788 $\pm$ 0.032 |
| | Severe | – | 0.731 $\pm$ 0.037 | 0.842 $\pm$ 0.057 | 0.782 $\pm$ 0.037 |
| *Row-normalised confusion matrix for the two-DMO decision tree* | | | | | |
| True class | | | | Predicted non-severe | Predicted severe |
| Non-severe | | | | 0.737 $\pm$ 0.047 | 0.264 $\pm$ 0.047 |
| Severe | | | | 0.158 $\pm$ 0.057 | 0.842 $\pm$ 0.057 |

Table 11: Three-class MS EDSS classification using walking-speed P90 and cadence P90. Values are mean $\pm$ standard deviation over five participant-level test splits; the final panel reports the row-normalised confusion matrix.

| Method | Class | Balanced accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| *Overall class-balanced performance* | | | | | |
| Always most common class | Macro average | 0.333 $\pm$ 0.000 | 0.153 $\pm$ 0.005 | 0.333 $\pm$ 0.000 | 0.209 $\pm$ 0.005 |
| Two-DMO decision tree | Macro average | 0.593 $\pm$ 0.023 | 0.543 $\pm$ 0.016 | 0.593 $\pm$ 0.023 | 0.533 $\pm$ 0.030 |
| *Class-specific performance* | | | | | |
| Always most common class | Mild | – | 0.000 $\pm$ 0.000 | 0.000 $\pm$ 0.000 | 0.000 $\pm$ 0.000 |
| | Moderate | – | 0.000 $\pm$ 0.000 | 0.000 $\pm$ 0.000 | 0.000 $\pm$ 0.000 |
| | Severe | – | 0.458 $\pm$ 0.016 | 1.000 $\pm$ 0.000 | 0.628 $\pm$ 0.015 |
| Two-DMO decision tree | Mild | – | 0.367 $\pm$ 0.045 | 0.708 $\pm$ 0.082 | 0.481 $\pm$ 0.041 |
| | Moderate | – | 0.516 $\pm$ 0.034 | 0.264 $\pm$ 0.075 | 0.344 $\pm$ 0.060 |
| | Severe | – | 0.745 $\pm$ 0.051 | 0.807 $\pm$ 0.054 | 0.774 $\pm$ 0.041 |
| *Row-normalised confusion matrix for the two-DMO decision tree* | | | | | |
| True class | | | Predicted mild | Predicted moderate | Predicted severe |
| Mild | | | 0.708 $\pm$ 0.082 | 0.230 $\pm$ 0.085 | 0.063 $\pm$ 0.051 |
| Moderate | | | 0.431 $\pm$ 0.095 | 0.264 $\pm$ 0.075 | 0.305 $\pm$ 0.063 |
| Severe | | | 0.056 $\pm$ 0.043 | 0.137 $\pm$ 0.054 | 0.807 $\pm$ 0.054 |

sustained-walking measures are more frequently associated with severe EDSS disability. It is worth noting that these values are descriptive model splits, not proposed clinical cut-offs.

This result illustrates the type of feedback that the selected DMOs might support. Rather than returning only an unexplained message such as "your mobility may be declining", a future system could identify the measurements that contributed to the alert. For example, it might report: "Your sustained walking-speed P90 and cadence P90 have decreased since the previous monitoring period and have moved into ranges that were associated with greater mobility impairment in the reference data. You may wish to pay closer attention to your sustained walking performance and discuss a persistent decline with a relevant professional." Such feedback directly connects the software output to observable changes in daily-life mobility. Importantly, it explains the basis of the software's judgement rather than the biological cause of the person's condition.

**Extension to three severity classes.** We next examine whether the same two selected DMOs can address the more difficult task of distinguishing mild (EDSS 0–3.0), moderate (3.5–5.5), and severe ($\geq 6.0$) disability. The corresponding class counts are 317, 840, and 926 observations. The model and evaluation protocol remain unchanged.

For this harder task, the two-DMO tree achieves balanced accuracy of $0.593 \pm 0.023$, macro-precision of $0.543 \pm 0.016$, and macro-F1 of $0.533 \pm 0.030$. It correctly identifies $70.8\%$ of mild and $80.7\%$ of severe observations, but only $26.4\%$ of moderate observations. The row-normalised confusion matrix shows that $43.1\%$ of moderate observations are predicted as mild and $30.5\%$ as severe, whereas direct confusion between the two extreme classes is less frequent ($6.3\%$ and $5.6\%$). Thus, the two

selected DMOs retain useful information in a finer-grained setting, but they do not reliably resolve the intermediate disability group.

**Scope and limitations.** These thresholds and classifications are associations learned from Mobilise-D and should be viewed as hypothesis-generating observations, not diagnostic cut-offs. The current experiments classify disability at an observed visit; they do not yet implement the complete software pipeline or prospectively detect individual deterioration. A practical system would require explicit rules for defining within-participant change, repeated-measure validation, personalised reference ranges, and careful design of user-facing feedback. It would also need to account for factors such as age, body size, walking aids, comorbidities, and sensor wear conditions. The experiments therefore provide an initial illustration of how DeMMO-derived patterns could make future longitudinal mobility-monitoring software more transparent and informative.