

Using latent representations to link disjoint longitudinal data for mixed-effects regression

Clemens Schächter^{1,2,*}, Maren Hackenberg^{1,2}, Michelle Pfaffenlehner^{1,2}, Félix B. Tambe-Ndonfack³, Thorsten Schmidt³, Astrid Pechmann⁴, Janbernd Kirschner⁴, Jan Hasenauer^{5,6}, and Harald Binder^{1,2,7}

¹Institute of Medical Biometry and Statistics (IMBI), Medical Faculty and Medical Center – University of Freiburg, Germany

²Freiburg Center for Data Analysis, Modeling and AI – University of Freiburg, Germany

³Department of Mathematical Stochastics, Mathematical Institute – University of Freiburg, Germany

⁴Department of Neuropediatrics and Muscle Disorders, Medical Faculty and Medical Center – University of Freiburg, Germany

⁵Bonn Center for Mathematical Life Sciences – University of Bonn, Germany

⁶Life and Medical Sciences (LIMES) Institute – University of Bonn, Germany

⁷CIBSS, Centre for Integrative Biological Signalling Studies – University of Freiburg, Germany

*Corresponding author, contact: clemens.schaechter@uniklinik-freiburg.de

November 6, 2025

Abstract

Many rare diseases offer limited established treatment options, leading patients to switch therapies when new medications emerge. To analyze the impact of such treatment switches within the low sample size limitations of rare disease trials, it is important to use all available data sources. This, however, is complicated when the use of measurement instruments changes during the observation period, for example when instruments are adapted to specific age ranges. The resulting disjoint longitudinal data trajectories complicate the application of traditional modeling approaches like mixed-effects regression. We tackle this by mapping observations of each instrument to an aligned low-dimensional temporal trajectory, enabling longitudinal modeling across instruments. Specifically, we

employ a set of variational autoencoder architectures to embed item values into a shared latent space for each time point. Temporal disease dynamics and treatment switch effects are then captured through a mixed-effects regression model applied to latent representations. To enable statistical inference, we present a novel statistical testing approach that accounts for the joint parameter estimation of mixed-effects regression and variational autoencoders. The methodology is applied to quantify the impact of treatment switches for patients with spinal muscular atrophy. Here, our approach aligns motor performance items from different measurement instruments for mixed-effects regression and maps estimated effects back to the observed item level to quantify the treatment switch effect. Our approach allows for model selection as well as for assessing effects of treatment switching. The results highlight the potential of modeling in joint latent representations for addressing small data challenges.

Keywords: Measurement instrument alignment, Variational autoencoder, Mixed-effects regression, Treatment switches, Statistical inference, Small data, Spinal muscular atrophy

1 Introduction

Rare diseases often have a limited number of established treatments, leaving patients with few therapeutic options. Thus, when new medications emerge, patients often switch treatments. However, analyzing the impact of such treatment switches longitudinally is difficult since rare disease trials are characterized by small data challenges. [1, 2, 3]. These include small sample sizes, heterogeneous data collected across different sites and varying observation frequencies. Additional difficulties arise when the preferred instruments to measure clinical outcomes change during the observation window. Such changes can occur from an evolving gold standard or measurement instruments that are tailored to patient characteristics that may change over time, like age or disease severity. However, using data from all available sources is important to make the most of the limited study population. For example, in children affected by spinal muscular atrophy (SMA), symptom progression is monitored longitudinally using several specialized, age- and symptom-specific measurement instruments that evaluate different motor function skills [4, 5, 6, 7, 8, 9]. As patients age and their disease progresses, the preferred assessment instruments are adapted accordingly. Therefore, the longitudinal record often comprises only a subset of these instruments at a given time point.

Statistical approaches, such as mixed-effects regression, are widely used for modeling disease progression in longitudinal data [10]. They estimate both population-level effects (fixed effects) and subject-specific deviations (random effects) simultaneously. Applications span multiple diseases where disease progression is measured using a variety of motor function or cognitive assessments. They include neuromuscular disorders such as SMA [11, 12, 13, 14], as well as central nervous system disorders such as Parkinson’s disease [15, 16], and multiple sclerosis [17, 18]. Although use cases vary, we aim to model the impact of switching between medications using a mixed-effects regression.

However, there are two core issues with the application to disjoint data sources as encountered in SMA data. First, multidimensional measurements are usually aggregated into a single score to reduce the model complexity, which discards the information contained in individual test items [19]. Second, handling switches in measurement instruments is often circumvented by limiting analysis to time frames where a specific measurement instrument was applied. However, this does not provide generality for the whole observation window, as patients with an improving or worsening disease condition can transition to a different measure-

ment instrument to circumvent ceiling or floor effects at the top or bottom of the test scale [4]. This introduces a bias, while a reduction in sample size reduces statistical power, which already is an issue in rare disease settings to begin with. While results of different studies reporting treatment effects are usually aggregated using a meta-analysis [20, 21], aligning results from different instruments is challenging due to within-patient correlation across instruments and different measurement scales.

Motivated by these challenges in longitudinal data, we propose an approach that combines mixed-effects regression with an artificial neural network approach, namely variational autoencoders (VAEs) [22]. We leverage VAEs to incorporate multiple measurement instruments at the item level, to obtain a latent representation at each time point of longitudinally observed data, while linking the time points by a mixed-effects regression model. The aim is to facilitate modeling of treatment switch effects while providing an approach that allows for statistical inference on this joint low-dimensional latent representation. To enable statistical testing, we introduce a bootstrap knockoff variable approach which corrects for potential biases caused by fitting the mixed-effects model within a VAE architecture. As the mappings to and from the latent space are learnable and non-linear, the approach can also compensate for ceiling effects present in the datasets. Therefore, a continuous outcome linear regression in the latent space is sufficient. To reflect different aspects of the patient characteristics measured by different instruments and allow for a flexible embedding, we employ a latent space with several dimensions, and correspondingly consider mixed models with a multivariate outcome.

The problem of extracting similar information from different measurement instruments has been considered to some extent in the field of item response theory [23], which is typically based on factor analysis [24, 25]. This includes linear and non-linear factor models [26], and more general latent variable models [27, 28, 29]. Yet, these often rely on strong distributional assumptions and require anchor items present across different instruments [29], which limits applicability in settings with several clinical measurement instruments. The concept of domain adaptation, developed in the machine learning community, formalizes the challenge of linking data from a source and target domain under the assumption of a shared latent space. Applications have been particularly successful for image data [30, 31, 32], such as in MR imaging [33, 34] or microscopy [35]. There, deep learning approaches are employed to learn domain-invariant representations [36, 37, 38]. These approaches generally assume a rather large number of observations. Also, they typically do not consider time structure, and when they do, focus

on time series with a large number of time points. Therefore, while the general concept of domain adaptation might be useful for longitudinal rare disease registries, there is no readily applicable approach for small numbers of patients with limited observations. In particular, a combination of domain adaptation with statistical models for longitudinal models, such as mixed-effects regression is missing so far. While VAEs have been combined with mixed-effects models in the context of imaging data [39, 40] and intensive care unit data [41], this does not cover changes between measurement instruments. In our own work, we have previously combined VAEs with ordinary differential equations (ODEs) for modeling disease trajectories in a lower-dimensional representation [42, 43, 44], based on a neural differential equation framework that allows for simultaneously fitting neural networks and dynamic models [45, 46, 47]. While we have extended this to also allow for two different measurement instruments [48], the focus on ODEs as deterministic models did not allow for statistical inference so far, as needed for statistically testing the effect of treatment switches. This could be provided by combining mixed-effects regression within VAEs that align different measurement instruments.

The manuscript is structured as follows: First, we introduce the approach in Section 2, after a brief overview of VAEs and multivariate linear mixed-effects models. To enable statistical testing, we introduce a bootstrap knockoff variable approach which characterizes and corrects for potential biases caused by fitting the mixed-effects model on VAE obtained latent variables. This is accompanied by an approach for quantifying effects in the latent space by mapping them back to the item representation. In Section 3, we assess the approach using data for SMA patients from the SMARtCARE registry [49]. This registry reports observations collected using five different motor performance measurement instruments, which need to be integrated to analyze the impact of treatment switches. We compare the performance of the proposed approach to a method based on a meta-analysis over per-measurement instrument models. Additionally, we validate the approach by evaluating results when recovering artificially added treatment switches and illustrate how to perform model selection for the latent mixed-effects model. A discussion is provided in Section 4, including a perspective on the potential of combining latent representations with statistical modeling for jointly analyzing data in small data settings. Finally, details on the hyperparameters, dataset and model choice are given in Section 6.

2 Methods

In the following, we denote random variables as uppercase letters (e.g., Y, Z), their realizations as bold lowercase letters (e.g., $\mathbf{y}, \mathbf{z}$), their corresponding (conditional) probability distributions as uppercase P with appropriate subscripts (e.g., $P_Y, P_{Z|\mathbf{y}}$), and their associated probability density functions as lowercase p (e.g., $p_Y, p_{Z|\mathbf{y}}$). Matrices containing numerous observations are written in uppercase letters (e.g., $\mathbf{Y}, \mathbf{Z}$). Table 2 lists the notation used throughout the manuscript.

2.1 Variational autoencoders

In this study, we use a set of variational autoencoder (VAE) [22] architectures to map data from the observation spaces of multiple measurement instruments into a shared low-dimensional representation of patient characteristics and reconstruct them back to the original space. A variational autoencoder is a generative deep learning approach, i.e., a model that allows for sampling, that learns latent representations of high-dimensional data based on the principle of variational inference [50]. Therefore, a VAE provides tractable approximations to probability distributions without requiring restrictive distributional assumptions. In a first step, we introduce the classical VAE framework on a dataset containing i.i.d. observations from a single measurement instrument and time point, while we outline our approach to align multiple instruments and link them longitudinally via a mixed model in subsection 2.3.

Consider i.i.d. observations $\{\mathbf{y}_i\}_{i \in \mathcal{I}}, \mathbf{y}_i \in \mathbb{R}^n$. A VAE introduces a latent variable $Z \in \mathbb{R}^d$ with $d < n$ in a generative model $P_{Y,Z}$ to capture the underlying characteristics of the data. Specifically, the generative process draws a sample $\mathbf{z}_i$ from a prior distribution P_Z , typically a multivariate standard normal $P_Z = \mathcal{N}_d(\mathbf{0}, \mathbf{I}_d)$, and then generates an observation $\mathbf{y}_i$ from the conditional sampling distribution $P_{Y|\mathbf{z}_i}$. Accordingly, the density of the joint distribution factorizes as $p_{Y,Z}(\mathbf{y}_i, \mathbf{z}_i) = p_{Y|\mathbf{z}_i}(\mathbf{y}_i)p_Z(\mathbf{z}_i)$.

Within the VAE framework, the marginal likelihood of an observation, $p_Y(\mathbf{y})$, is obtained by

$$p_Y(\mathbf{y}_i) = \int_{\mathbb{R}^d} p_{Y|\mathbf{z}}(\mathbf{y}_i)p_Z(\mathbf{z})d\mathbf{z}.$$

Therefore, evaluation of the marginal likelihood $p_Y(\mathbf{y}_i)$ involves integrating over the latent space, making the required computations analytically intractable. This intractability extends to the posterior $P_{Z|\mathbf{y}_i}$, because Bayes' rule demands dividing

by $p_Y(\mathbf{y}_i)$. As a result, any inference or model evaluation that relies on the exact posterior or marginal likelihood becomes unfeasible.

To circumvent this problem, VAEs employ a variational distribution $Q_{Z|Y_i} \approx P_{Z|Y_i}$, providing a tractable approximation to the true posterior. This variational distribution is chosen as a multivariate normal distribution with independent components, whose parameters are the function output of an encoder neural network $\text{enc}_\phi : \mathbb{R}^n \rightarrow \mathbb{R}^d \times \mathbb{R}_{>0}^d$ parameterized by learnable neural network parameters ϕ :

$$Q_{Z|Y_i} = \mathcal{N}_d(\boldsymbol{\mu}_i, \text{diag}(\boldsymbol{\sigma}_i^2)); (\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i) = \text{enc}_\phi(\mathbf{y}_i).$$

To enable gradient-based optimization, latent values are obtained by sampling through the reparametrization trick, with decomposition $\mathbf{z}_i = \boldsymbol{\mu}_i + \boldsymbol{\sigma}_i \odot \boldsymbol{\varepsilon}_i$, such that the gradient can be obtained from the non-random part, while $\boldsymbol{\varepsilon}_i$ is auxiliary noise drawn from a multivariate standard normal distribution $\mathcal{N}_d(\mathbf{0}, \mathbf{I}_d)$, and $\odot$ denotes the Hadamard product between two vectors. A decoder neural network $\text{dec}_\theta : \mathbb{R}^d \rightarrow \Psi$, parameterized by θ , maps the latent values back to the data space by modeling the parameters $\psi \in \Psi$ of the conditional sampling distribution $P_{Y|\mathbf{z}_i}$. The parameters of the encoder and decoder neural networks are simultaneously optimized by estimating and minimizing the evidence lower bound (ELBO) function

$$\mathcal{L}_{\text{VAE}}(\phi, \theta) = -\mathbb{E}_{Z \sim Q_{Z|Y_i}} [\log p_{Y|Z}(\mathbf{y}_i)] + \beta \text{KL} [Q_{Z|Y_i} \parallel P_Z],$$

where $\text{KL} [\cdot \parallel \cdot]$ denotes the Kullback-Leibler (KL) divergence between the variational distribution $Q_{Z|Y_i}$ and the prior P_Z and the parameter $\beta > 0$ balances the KL-divergence with the reconstruction error term [51]. Optimization of the ELBO function is performed over mini-batches, i.e., randomly selected groups of observations, using stochastic gradient descent methods.

For patients $i \in \mathcal{I}$ that have longitudinal measurements, we denote their data as $\mathbf{Y}_i = (\mathbf{y}_{i,t})_{t \in \mathcal{T}_i}^\top \in \mathbb{R}^{m_i \times n}$. The individual observations are collected at m_i visit times $\mathcal{T}_i = \{t_{i,1}, \dots, t_{i,m_i} : 0 = t_{i,1} < \dots < t_{i,m_i}\}$. We optimize the network weights simultaneously on a dataset containing all time points and patients, treating them as i.i.d. observations. This results in an $m_i \times d$ -dimensional representation $\mathbf{Z}_i = (\mathbf{z}_{i,t})_{t \in \mathcal{T}_i}^\top \in \mathbb{R}^{m_i \times d}$ for each patient $i \in \mathcal{I}$, as a basis for subsequent modeling by linear mixed-effects regression.

2.2 Multivariate longitudinal mixed-effects regression

To incorporate time structure into the latent representation, we utilize multivariate mixed-effects regression models [52, 53]. They represent a patient trajectory as a realization of a random variable $Z_i \in \mathbb{R}^{m_i \times d}$, which is a linear combination of fixed effects, random effects, and residual errors:

$$Z_i = \mathbf{X}_i \mathbf{B} + \mathbf{T}_i U_i + E_i.$$

Fixed effects $\mathbf{B} \in \mathbb{R}^{p \times d}$ with design matrix $\mathbf{X}_i \in \mathbb{R}^{m_i \times p}$ represent the influence of p patient covariates with consistent effects across all patients. Such covariates could include age, sex, or administered medications. Interaction terms with age can model a change in influence over time.

Random effects $U_i \in \mathbb{R}^{q \times d}$ with design matrix $\mathbf{T}_i \in \mathbb{R}^{m_i \times q}$ allow for individual-specific deviations from the fixed group effects. Common choices include a random intercept and slope to model patient-specific trajectories. In this study, we assume that the random effects follow a normal distribution $\text{vec}(U_i) \sim \mathcal{N}_{qd}(\mathbf{0}, \Phi)$ with shared covariance matrix Φ across patients, where $\text{vec} : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^{nm}$ denotes the vectorization function, which stacks the columns of a matrix into a vector. Residual errors $E_i \in \mathbb{R}^{m_i \times d}$ account for variability not explained by fixed and random effects. Residuals are assumed to be independent across time given the random effects and across patients, $\text{Cov}(\text{vec}(U_i), \text{vec}(E_i)) = 0$. Furthermore, we assume that they follow a normal distribution $\text{vec}(E_i) \sim \mathcal{N}_{m_i d}(\mathbf{0}, \Sigma \otimes \mathbf{I}_{m_i})$, where $\otimes$ denotes the Kronecker product between two matrices.

The marginal distribution of the response variable Z_i can be described as

$$\text{vec}(Z_i) \sim \mathcal{N}_{m_i d} \left(\text{vec}(\mathbf{X}_i \mathbf{B}), \underbrace{(\mathbf{I}_d \otimes \mathbf{T}_i) \Phi (\mathbf{I}_d \otimes \mathbf{T}_i)^\top + \Sigma \otimes \mathbf{I}_{m_i}}_{=:\mathbf{V}_i^{-1}} \right)$$

Given observed patient trajectories $\mathbf{Z}_i$ of the random variable Z_i , the variance parameters Φ and Σ are estimated using maximum likelihood (ML) or restricted maximum likelihood (REML) estimation [54], where the log-likelihoods can be

calculated as

$$\begin{aligned}\mathcal{L}_{\text{ML}}(\mathbf{B}, \Phi, \Sigma) &= \frac{1}{2} \sum_{i \in \mathcal{I}} \left(\log \det(\mathbf{V}_i) - \text{vec}(\mathbf{Z}_i - \mathbf{X}_i \mathbf{B})^\top \mathbf{V}_i \text{vec}(\mathbf{Z}_i - \mathbf{X}_i \mathbf{B}) \right. \\ &\quad \left. - m_i d \log(2\pi) \right) \\ \mathcal{L}_{\text{REML}}(\Phi, \Sigma) &= \frac{1}{2} \sum_{i \in \mathcal{I}} \left(\log \det(\mathbf{V}_i) - \log \det \left((\mathbf{I}_d \otimes \mathbf{X}_i)^\top \mathbf{V}_i (\mathbf{I}_d \otimes \mathbf{X}_i) \right) \right. \\ &\quad \left. - \text{vec}(\mathbf{Z}_i - \mathbf{X}_i \hat{\mathbf{B}})^\top \mathbf{V}_i \text{vec}(\mathbf{Z}_i - \mathbf{X}_i \hat{\mathbf{B}}) - (m_i - p) d \log(2\pi) \right).\end{aligned}$$

Optimization algorithms like Newton-Raphson or quasi-Newton methods can be used to maximize the likelihood functions [55]. With an estimate of the covariance components and observed response $\mathbf{Z}_i$, the realized best linear unbiased estimator (BLUE) $\hat{\mathbf{B}}$ of the fixed effects, is given by

$$\text{vec}(\hat{\mathbf{B}}) = \left(\sum_{i \in \mathcal{I}} (\mathbf{I}_d \otimes \mathbf{X}_i)^\top \hat{\mathbf{V}}_i (\mathbf{I}_d \otimes \mathbf{X}_i) \right)^{-1} \left(\sum_{i \in \mathcal{I}} (\mathbf{I}_d \otimes \mathbf{X}_i)^\top \hat{\mathbf{V}}_i \text{vec}(\mathbf{Z}_i) \right).$$

The best linear unbiased predictor (BLUP) of the random effects U_i can be obtained through

$$\text{vec}(\hat{\mathbf{U}}_i) = \hat{\Phi} (\mathbf{I}_d \otimes \mathbf{T}_i)^\top \hat{\mathbf{V}}_i \text{vec}(\mathbf{Z}_i - \mathbf{X}_i \hat{\mathbf{B}}).$$

2.3 Joining multiple measurement instruments in a latent representation

For incorporating data from multiple measurement instruments $l \in \mathcal{L} \subset \mathbb{N}$, where each instrument comprises n_l test items, we use one encoder and decoder network per instrument, embedding data into a joint latent representation. Here observations are linked by multivariate mixed-effects regression, as schematically shown in Figure 1. We assume a patient $i \in \mathcal{I}$ is assessed using a subset of different measurement instruments at each time point. Denote by $\mathcal{T}_{i,l} \subseteq \mathcal{T}_i$ the observations in which measurement instrument $l \in \mathcal{L}$ was utilized. At the data level we obtain one trajectory per patient and observed instrument,

$$\mathbf{Y}_{i,l} = (\mathbf{y}_{i,t,l})_{t \in \mathcal{T}_{i,l}}^\top \in \mathbb{R}^{m_{i,l} \times n_l}, \quad l \in \mathcal{L},$$

where $m_{i,l}$ denotes the overall number of observations for patient i with measurement instrument l . Measurement instrument specific encoder networks enc_{ϕ_l} are

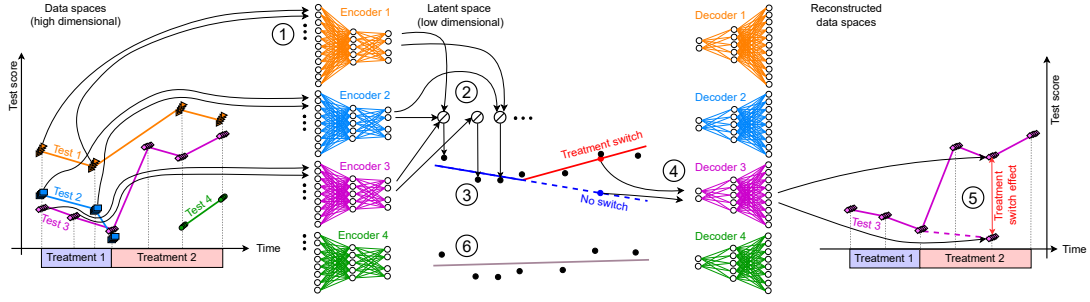


Figure 1: Schematic illustration of the proposed approach for multiple measurement instruments: 1) The items of each measurement instrument are encoded by a dedicated encoder network and corresponding latent values are drawn. 2) The latent values are averaged across instruments at each time step, to obtain a joint latent trajectory. 3) The averaged latent values serve as outcome variable for a multivariate mixed-effects regression, which provides BLUE and BLUP estimators. 4) Predictions from the mixed-effects model serve as input to a dedicated decoder network for each measurement instrument, for reconstruction at the original item level. 5) Treatment switch effects can be quantified per item. 6) A likelihood-ratio test provides statistical inference.

used to obtain the parameters of the variational distributions $Z_l \sim Q_{Z_l|y_{i,t,l}}$ from which we draw d -dimensional realizations

$$(\mathbf{z}_{i,t,l})_{t \in \mathcal{T}_{i,l}}^\top \in \mathbb{R}^{m_{i,l} \times d}, \quad l \in \mathcal{L}.$$

Denote with $\mathcal{L}_{i,t} \subseteq \mathcal{L}$ the index set of all measurement instruments used for patient i at observation time t . To obtain a unified latent trajectory for a patient, we average latent variables over all available values at the respective time points, yielding

$$\mathbf{Z}_i = (\mathbf{z}_{i,t})_{t \in \mathcal{T}_i}^\top = \left(\sum_{l \in \mathcal{L}_{i,t}} \frac{\mathbf{z}_{i,t,l}}{|\mathcal{L}_{i,t}|} \right)_{t \in \mathcal{T}_i}^\top \in \mathbb{R}^{m_i \times d}.$$

We then model the longitudinal structure in the latent space using a multivariate mixed-effects model of dimension d . The latent mixed-effects model in particular can incorporate effects of treatment switches. Let t_{switch_i} denote the time of a switch from one treatment to another. We collect the time differences relative to the switch

$$\Delta \mathbf{t}_i = (t_{i,1} - t_{\text{switch}_i}, t_{i,2} - t_{\text{switch}_i}, \dots, t_{i,m_i} - t_{\text{switch}_i})^\top.$$

To capture group-level trends following a treatment switch, the time since switch

$\max(0, \Delta t_i)$ is included as a fixed effect. To model individual deviations from group trends, there are random effects included for the time since $\max(0, \Delta t_i)$ and the time up to a treatment switch $\min(0, \Delta t_i)$. When multiple treatments are involved their influence can be separated into treatment specific covariates. Interaction terms with other covariates (e.g., age) can also be included in the mixed-regression model. All decoder networks receive as input the prediction of the mixed model

$$\hat{\mathbf{Z}}_i = \mathbf{X}_i \hat{\mathbf{B}} + \mathbf{T}_i \hat{\mathbf{U}}_i$$

where $\hat{\mathbf{B}}$ is the BLUE and $\hat{\mathbf{U}}_i$ is the BLUP, given the latent outcome $\mathbf{Z}_i$.

To estimate the parameters of the encoders and decoders as well as the latent mixed-effects model, we use an iterative approach: To obtain the encoder and decoder parameters, (ϕ_l, θ_l) for $l \in \mathcal{L}$, we freeze the mixed-effect model parameters and minimize as a combination of per-measurement instrument ELBO loss functions and two alignment terms:

$$\begin{aligned} \mathcal{L}_{\text{MMVAE}}((\phi_l, \theta_l)_{l \in \mathcal{L}}) = & \frac{1}{|\mathcal{I}| \sum_{i \in \mathcal{I}} |\mathcal{T}_i|} \left(\gamma \sum_{i \in \mathcal{I}} \sum_{t \in \mathcal{T}_i} \|\hat{\mathbf{z}}_{i,t} - \mathbf{z}_{i,t}\|_2^2 - \eta \mathcal{L}_{\text{ML}}(\mathbf{B}, \Phi, \Sigma) \right) + \\ & \frac{1}{|\mathcal{I}| \sum_{i \in \mathcal{I}} \sum_{l \in \mathcal{L}} |\mathcal{T}_{i,l}|} \sum_{i \in \mathcal{I}} \sum_{l \in \mathcal{L}} \sum_{t \in \mathcal{T}_{i,l}} \left(-\mathbb{E}_{\mathbf{Z}_l \sim Q_{\mathbf{Z}_l | \mathbf{y}_l}} [\log p_{Y_l | \hat{\mathbf{z}}_{i,t}}(\mathbf{y}_{i,t,l})] + \right. \\ & \left. \beta \text{KL} [Q_{\mathbf{Z}_l | \mathbf{y}_{i,t,l}} || P_{\mathbf{Z}}] \right) \end{aligned}$$

The first alignment term, weighted by $\eta \geq 0$ encourages agreement with the training criterion (ML/REML) of the frozen mixed model, while the second alignment term weighted by $\gamma \geq 0$, encourages agreement between the encoder output and the latent mixed model prediction.

Because the joint latent, serving as response to the mixed model, is the average of measurement specific latent variables $\mathbf{z}_{i,t} = |\mathcal{L}_{i,t}|^{-1} \sum_{l \in \mathcal{L}_{i,t}} \mathbf{z}_{i,t,l}$, the encoders are implicitly driven to place their outputs on a common scale. Instrument-specific shifts or a rescaling inflates the disagreement penalty $\|\hat{\mathbf{z}}_{i,t} - \mathbf{z}_{i,t}\|_2^2$, as well as reduce the mixed-effects likelihood, and as mixed model predictions are passed to the decoder ultimately the reconstruction loss and are therefore corrected during training.

After fitting the encoder and decoder parameters for given mixed-effect model parameters for a given amount of iterations with a gradient descent method like

Adam [56], a latent representation for each trajectory $\mathbf{Z}_i, i \in \mathcal{I}$ is drawn. Afterwards, the mixed-effects model is fitted using these latent representations as response, with ML or REML estimation. Such alternating updates of the VAE and mixed-effects model parameters are performed until the loss function $\mathcal{L}_{\text{MMVAE}}$ no longer steeply decreases and decreases saturate.

With the final parameter estimates, the impact of treatment switches can be quantified on the item level by decoding from the latent trajectory with and without treatment switch. These trajectories can be obtained by calculating the BLUE and BLUP, modifying the design matrices $\mathbf{X}_i, \mathbf{T}_i$ by setting the covariates and interaction terms relating to time elapsed after treatment switches to zero and extending the covariates relating to time before treatment switches. The latent prediction of both models can be mapped back to the data level using the decoder neural networks and analyzed at the item level for differences.

2.4 Model selection

Statistical testing based on the mixed-effects model can be useful for further assessing the effect of treatment switches, as well as for potential model selection, e.g., to decide which covariates to include for adjustment. One would normally assess the significance of a covariate by a likelihood-ratio test [57] with test-statistic $\Lambda = 2(\mathcal{L}_{\text{ML full}} - \mathcal{L}_{\text{ML red}})$. Here, $\mathcal{L}_{\text{ML full}}$ denotes the likelihood of a model containing the covariate of interest and $\mathcal{L}_{\text{ML red}}$ corresponds to the likelihood of a nested model without it. Under Wilks' theorem and standard regularity Λ is asymptotically χ_{rd}^2 for a fixed effects block, with r being the number of tested covariates and d the dimension of the latent space.

The classical statistical model selection approach cannot be applied here as the final latent responses $\mathbf{Z}_i$ depend on encoder parameters $\phi_l, l \in \mathcal{L}$ that were jointly optimized with the mixed-effects coefficients. The same data is therefore used first to pick a representation and then to test a hypothesis, which can induce a post-selection bias [58, 59]. In this setting the χ_{rd}^2 reference distribution can be invalid. To correct for the intractable selection event ("the VAEs picked a particular latent representation"), we combine knockoff negative-control covariates [60, 61] with bootstrap [62, 58]. The resulting procedure allows for inference for the joint VAE–mixed-effects estimator.

Let $\mathbf{w}_i \in \mathbb{R}^k$ denote the realization of a random variable W_i , drawn independently of $\{(\mathbf{Z}_i, \mathbf{X}_i, \mathbf{T}_i)\}_{i \in \mathcal{I}}$ with $\mathbb{E}[W_i] = \mathbf{0}$, $\text{Cov}(W_i) = \mathbf{I}_k$. Let $\mathbf{W}_i = (\mathbf{w}_{i,1} \cdots \mathbf{w}_{i,m_i})^\top$ be concatenated variables that are either equal for each visit for patient-level, or

re-drawn to represent visit-level knockoff variables. We augment the fixed effect design matrix by $\widetilde{\mathbf{X}}_i = [\mathbf{X}_i \mid \mathbf{W}_i]$ and test

$$H_0 : \beta_W = \mathbf{0} \quad \text{vs.} \quad H_1 : \beta_W \neq \mathbf{0},$$

where $\beta_W \in \mathbb{R}^{K \times d}$ denotes the coefficient block of $\mathbf{W}_i$. Because $\mathbf{W}_i \perp\!\!\!\perp \mathbf{Z}_i \mid \mathbf{X}_i, \mathbf{T}_i$ by construction, H_0 is true with probability 1 and $\mathbf{W}_i$ acts as a negative-control predictor in the sense of the work by [60], yielding a reference statistic that captures what one should observe under the null hypothesis.

To empirically recreate the intractable selection event, we construct an empirical null distribution for the likelihood ratio test F_Λ^0 via bootstrap. For bootstrap samples $b \in \mathcal{B}$, we randomly reinitialize the VAE parameters and regenerate knockoff variables $\mathbf{W}_{i,b}$. Afterwards we re-optimize for $\{(\phi_{l,b}, \theta_{l,b})\}_{l \in \mathcal{L}}$ to obtain latent responses $\mathbf{Z}_{i,b}, b \in \mathcal{B}$, with which we calculate the likelihood ratio statistic Λ_b from the fitted full and reduced models. The empirical cumulative distribution function

$$\hat{F}_\Lambda^{0,\mathcal{B}}(x) = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \mathbb{1}\{\Lambda_b \leq x\}$$

converges in probability to F_Λ^0 as $|\mathcal{B}| \rightarrow \infty$, while resampling the knockoff together with the data (via different network weights) preserves the exchangeability property that underlies false discovery rate control in the original knockoff filter [61]. Consequently, the p -value

$$\hat{p} = \frac{1 + \sum_{b \in \mathcal{B}} \mathbb{1}\{\Lambda_b \geq \Lambda_{\text{obs}}\}}{|\mathcal{B}| + 1}$$

is an unbiased estimate of $\mathbb{P}_0(\Lambda \geq \Lambda_{\text{obs}})$.

3 Results

We use the proposed approach on data from the SMArtCARE registry [49], which tracks the disease course of patients with spinal muscular atrophy (SMA). SMA is a genetic disorder characterized by progressively declining motor function with symptom onset at birth or within early childhood. The severity of symptom progression is influenced by patient characteristics like genetic markers (e.g., count of the *SMN2* gene) or age at symptom onset and varies significantly across individuals [63]. Recently, there were advances in treating SMA [64], leading patients

to switch to emerging treatments when they became available to them. Treatment assignment within the SMARtCARE registry is not randomized, yet mainly depends on the availability of a treatment in a medical center. To monitor the impact of treatment switches on disease development, longitudinal motor function assessments are conducted during semi-regular visits. Here, movement ability is assessed using five specialized measurement instruments, the CHOP-INTEND [7], HINE-2 [8], HFMSE [6], RULM [5] and ALSFRS-R [9] motor function tests. (See Table 1 for details). In practice, infants and weak non-sitters are evaluated with CHOP-INTEND and motor milestone attainment in early childhood is captured with HINE-2. When patients age, assessments transition to HFMSE to assess gross motor ability and RULM for upper-limb performance, with ALSFRS-R complementing motor scales in adult cohorts. Instrument changes often occur to use stage-appropriate instruments and to avoid flooring and ceiling effects.

Our approach integrates these five different motor function measurement instruments for a cohort of 522 patients, with a median of 17 individual observations per patient and mean of 2.1 out of 5 measurement instruments used per visit. The latent mixed-effects regression model incorporates as covariates the time elapsed since treatment switch, age, a genetic marker, specifically the *SMN2* count, age at symptom onset, an indicator whether patients are currently asymptomatic for example because they were diagnosed via newborn screening, the current ventilation status, the current scoliosis status, sex of the patient and if a family member is affected by SMA. We also include interactions with age as fixed effects. Additionally, three random-effect terms were used to allow for individual deviations from group-level trends as random effects: a random intercept, and random slopes for the trajectory before and after the treatment switch.

3.1 Quantifying the impact of treatment switches

A key question in the context of SMA treatment is the impact of treatment switches on the individual and the population level. To assess this effect in the light of different measurement approaches, we use the proposed approach to compare two sets of predicted disease trajectories: one incorporating the treatment switch and a scenario in which the switch did not occur. We decoded the mixed-effects model predictions from both scenarios for one year after the treatment change and compared their reconstructed motor function test item scores. The difference in total sum scores then served as our estimate of the treatment switch effect.

As seen in Table 3, there is a positive predicted improvement between 1.8%

and 5.4% of the maximal score across the measurement instruments. Results are not deterministic as the encoder and decoder model distributions introduce additional stochasticity in the sampling and fitting process. However, assessed over ten random initializations there was no occurrence in which the aggregated treatment effect switched from positive to negative in one of the assessed instruments.

To assess sensitivity, we simulated an additional artificial treatment switch by adding a sum score of +2 points, distributed to randomly selected test items for each year after the switch (see chapter 6 for details). However, this perturbation ignores ceiling effects as scores cannot exceed the sum of the maximal individual item scores, within-observation structure, and cross-instrument alignment. We therefore expect the approach to detect the switch but understate its magnitude. Consistent with this expectation, the approach recovered an average treatment switch effect between 0.78 and 1.97 on the measurement instruments, indicating correct directionality and no overestimation of the effect size.

We also found that our approach reduces ceiling effects compared with a data-level mixed model without additional parameters to account for ceiling or floor effects (see Figure 2). Because the linear latent mixed-model predictions are passed through a decoder neural network which models the data distribution, samples are restricted to consist of plausible reconstructions. Non-linearity is further introduced by neural network transformations.

3.2 Latent model selection

We evaluated the statistical significance of effects in the mixed-effects model, such as of treatment switches, with likelihood ratio (LR) tests with critical values obtained from the proposed knockoff variable bootstrap procedure. The empirical distribution of the test statistic was constructed with $|\mathcal{B}| = 1000$ bootstrap samples.

First, we compared the empirical null distributions with the theoretical χ^2 -distributions that ignore the influence of the VAEs (see Figure 3). For a one-dimensional latent representation, the empirical knockoff distribution follows the theoretical distribution closely. However, for the three-dimensional representation that we want to use, there is a considerable difference, where the theoretical distribution would lead to an anti-conservative test. This is seen for fixed effects as well as for random effects, where deviations are even larger.

Subsequently, we conducted LR tests for the 11 fixed effects covariates in the

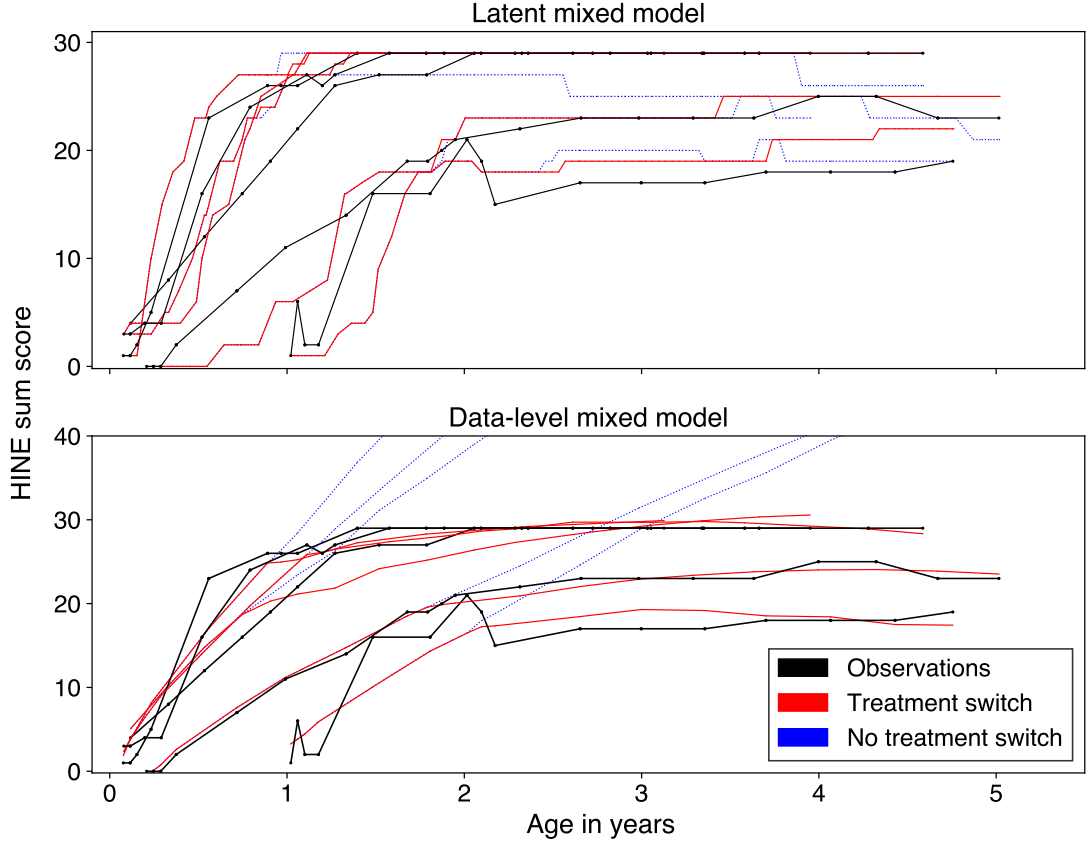


Figure 2: Comparison of total HINE-2 sum scores for five patient trajectories. Observed data trajectories are displayed in black, the respective mixed model prediction with treatment switch in red, and mixed model predictions for a hypothetical trajectory without treatment switch in blue. The reconstructed trajectories of the latent mixed model are displayed in the upper subplot and for a standard data-level mixed model in the lower subplot.

mixed-effects model. We indicate with rd the degrees of freedom the theoretical χ^2 -test statistic would follow, consisting of the amount of differences in included fixed effect (r) multiplied with the latent dimension (d). The effects for ventilation status ($\Lambda_{\text{obs}} = 182.8, \text{SD} = 28.0, rd = 6$), and for treatment switches ($\Lambda_{\text{obs}} = 335.2, \text{SD} = 59.8, rd = 12$), and disease onset ($\Lambda_{\text{obs}} = 237.6, \text{SD} = 29.7, rd = 12$) and surgery for scoliosis ($\Lambda_{\text{obs}} = 85.3, \text{SD} = 14.36, rd = 6$) showed high LR-values, above critical values from the knockoff null distribution. Other covariates, such as sex ($\Lambda_{\text{obs}} = 17.3, \text{SD} = 5.4, rd = 6$) and family history ($\Lambda_{\text{obs}} = 11.2, \text{SD} = 4.3, rd = 6$), did not display significant effects.

As a further check, we generated data using the generative nature of the VAEs based on the fitted latent trajectories, but without the treatment switch effect, and refitted the mixed-effects model to this newly generated dataset. The likelihood ra-

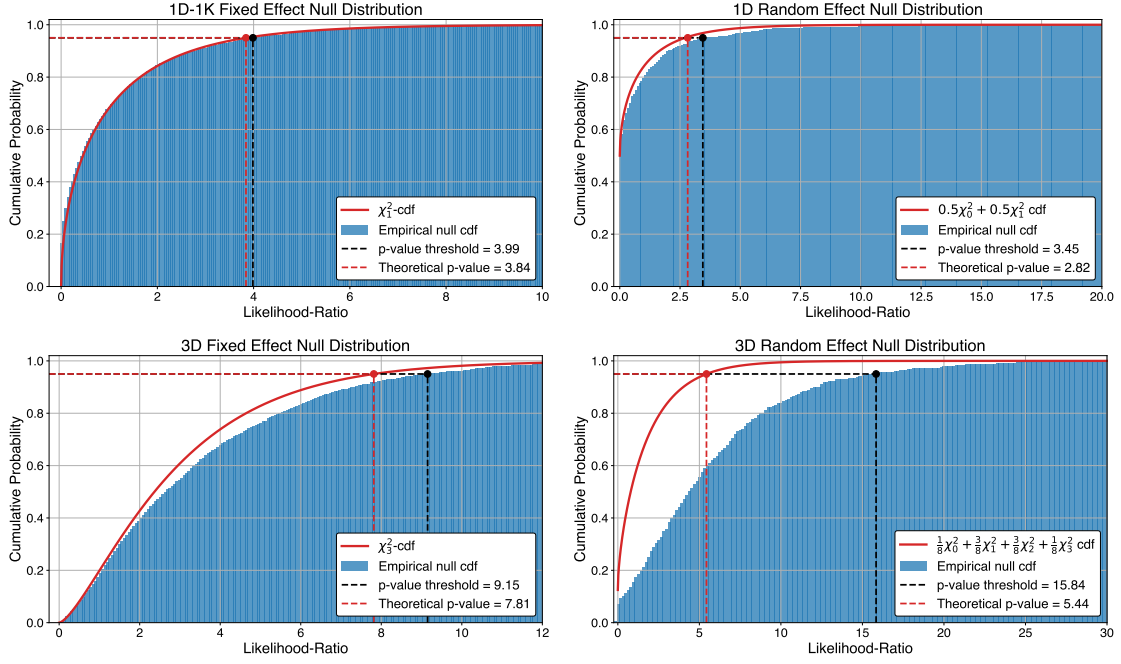


Figure 3: Empirical cumulative distribution functions of the null distributions of the likelihood ratio test statistic from artificially added knockoff variables for fixed (left column) and random effects (right column) in the mixed-effects regression in latent representations of dimension $d = 1$ (top row) and $d = 3$ (bottom row). The red line represents the theoretical chi-squared distribution that does not take into account the interdependent VAE and mixed model training procedure.

tion for including the treatment switch parameters dropped to 18.2, which no longer exceeded the empirical significance threshold (23.6 for $rd = 12$, meaning $d = 3$ latent dimensions with $r = 4$ differences in covariates, consisting of switches to two SMA medications and their interactions with age).

3.3 Comparison with naive meta-analysis

As a comparison to the proposed approach, we fit separate linear mixed-effects models for each motor instrument at the original data level, using the same fixed and random effects as in the latent model and taking instrument-specific sum scores as outcomes. Additionally, we added an intercept to the fixed effects as test sum scores are not centered around zero compared to latent variables in our latent mixed model. Because many instruments are not observed with sufficient pre- and post-switch visits for each patient, the effective sample is smaller than for the latent approach. Whereas the latent model could use a sample size of 522 patients, the data-level models were restricted to 308 patients for the RULM measurement instrument, 318 for HINE-2, 239 for HFMSE, and to 158 for CHOP-

INTEND. For the ALSFRS-R test, too few patients (61) were available to fit a mixed model using the same set of covariates.

To pool treatment switch effects across instruments while accounting for overlapping patients, we performed a correlated-effects fixed-effect meta-analysis via generalized least squares, using a patient-level bootstrap to estimate the full cross-instrument covariance and a bootstrap calibration for inference. For the combined treatment effects, the global likelihood-ratio statistic was 214.06, $rd = 3$ with a bootstrap-calibrated p-value of $p < 10^{-3}$ (no exceedance out of 1000 null replicates). Therefore, the meta-analysis provides strong evidence for a non-zero treatment switch effect, consistent in magnitude with the latent mixed-effects regression model.

When randomly sampled to a third of the original sample size, the latent approach could still reliably detect the treatment switch ($\Lambda_{\text{obs}} = 102.5$), while the sample size was now often too small to get a stable fit for the data-level approach, particularly for the CHOP-INTEND measurement instrument.

4 Discussion

In clinical studies of rare diseases, it is essential to account for changing measurement strategies, e.g., to assess treatment effects. As patient trajectories get longer and more observations become available, it becomes more likely that different measurement instruments are employed to monitor outcomes. This could be due to a changing state of the art or that measurement instruments are conditioned on patient characteristics. To exploit all historical data collected under different settings, novel methods are necessary. Here, we proposed an approach that combines VAEs with a latent longitudinal mixed-effects regression model to capture disease progression and treatment switches in SMA data collected from multiple measurement instruments. By using encoder neural networks to map items from these instruments onto a unified latent representation, we handle systematic changes in measurement instruments during the observation period. To enable statistical inference, we introduced a knockoff variable bootstrap testing approach which allows us to correct for potential biases arising during model fitting.

This provides unique capabilities, which we demonstrated when assessing effects of treatment switches in patients with SMA. By mapping to a latent embedding via artificial neural networks, we utilize the item-level details of observations. The learnable network mappings also handles ceiling effects and aligns measure-

ment instruments of different scales. This allowed us to integrate observations of different measurement instruments, e.g., tailored to different age ranges, to leverage a larger effective sample size compared to traditional methods that analyze each instrument separately. In other settings, power gains are particularly relevant for evaluating treatments in rare disease registries, where data is scarce. Potential bias in subsequent statistical inference, incurred by the flexibility of neural networks, could be mitigated by the proposed knockoff variable approach. At the same time, this approach allows for the detection of treatment switch effects in the SMA application, which might not have been identified otherwise.

Still there are several aspects that need to be carefully considered when utilizing the proposed approach. Although the projection to a low-dimensional space reduces the number of required parameters, the latent dimensionality is still restricted to avoid over-parametrization of the latent model, which can limit the modeled complexity of the underlying dynamics. On the other hand, choosing the latent dimension too large can increase the effects of overfitting, i.e., a larger deviation of the skewed test statistic from the theoretical distribution. Thus, the number of latent dimensions needs to be carefully chosen, ideally depending on the dataset and based on domain knowledge. One potential approach could be to estimate the intrinsic dimension (ID) [65] in a higher-dimensional latent space and then reduce the latent dimensionality to match the estimated ID. Further, specification of the mixed-effects model for the latent representation should incorporate biomedical knowledge, supported by the proposed statistical testing approach for model selection where necessary. This might even be an opportunity to gain further insight into the joint structure underlying several measurement instruments.

As datasets grow in complexity but not necessarily in size, which is typical for many rare diseases, there will be an increasing need for approaches that integrate deep learning methods and classical statistical theory. Such hybrid approaches can facilitate modeling on small but heterogeneous, high-dimensional and multi-modal datasets and provide answers for research questions that were previously challenging to address given the limited and complex data. As the explicit incorporation of statistical inference within deep learning frameworks is still uncommon, we therefore see this as a promising avenue of further research.

References

1. Mitani AA and Haneuse S. Small data challenges of studying rare diseases. *JAMA Netw Open* 2020; 3:e201965. DOI: 10.1001/jamanetworkopen.2020.1965
2. Whicher D, Philbin S, and Aronson N. An overview of the impact of rare disease characteristics on research methodology. *Orphanet J Rare Dis* 2018; 13:14. DOI: 10.1186/s13023-017-0755-5
3. Gaasterland CMW et al. The patient's view on rare disease trial design – a qualitative study. *Orphanet J Rare Dis* 2019; 14:31. DOI: 10.1186/s13023-019-1002-z
4. Gaasterland CMW et al. The patient's view on rare disease trial design – a qualitative study. *Orphanet J Rare Dis* 2019; 14:31. DOI: 10.1186/s13023-019-1002-z
5. Mazzone ES et al. Revised upper limb module for spinal muscular atrophy: Development of a new module. *Muscle Nerve* 2017; 55:869–74. DOI: 10.1002/mus.25430
6. O'Hagen JM et al. An expanded version of the Hammersmith functional motor scale for SMA II and III patients. *Neuromuscul Disord* 2007; 17:693–7. DOI: 10.1016/j.nmd.2007.05.009
7. Glanzman AM et al. The children's hospital of Philadelphia infant test of neuromuscular disorders (CHOP INTEND): test development and reliability. *Neuromuscul Disord* 2010; 20:155–61. DOI: 10.1016/j.nmd.2009.11.014
8. Haataja L et al. Optimality score for the neurologic examination of the infant at 12 and 18 months of age. *J Pediatr* 1999; 135:153–61. DOI: 10.1016/S0022-3476(99)70016-8
9. Cedarbaum JM et al. The ALSFRS-R: A revised ALS functional rating scale that incorporates assessments of respiratory function. *J Neurol Sci* 1999; 169:13–21. DOI: 10.1016/S0022-510X(99)00210-5
10. Laird NM and Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982; 38:963–74. DOI: 10.2307/2529876
11. Cavaloiu B, Simina IE, Vilciu C, Trăilă IA, and Puiu M. Nusinersen improves motor function in type 2 and 3 spinal muscular atrophy patients across time. *Biomedicines* 2024; 12:1782. DOI: 10.3390/biomedicines12081782

12. Jacqmin P et al. Mathematical disease progression modeling in type 2/3 spinal muscular atrophy. *Muscle Nerve* 2018; 58:528–35. DOI: 10.1002/mus.26178
13. Duong T et al. Nusinersen treatment in adults with spinal muscular atrophy. *Neurol Clin Pract* 2021; 11:e317–e327. DOI: 10.1212/CPJ.0000000000001033
14. Oskoui M et al. Two-year efficacy and safety of risdiplam in patients with type 2 or non-ambulant type 3 spinal muscular atrophy (SMA). *J Neurol* 2023; 270:2531–46. DOI: 10.1007/s00415-023-11560-1
15. Hanff AM et al. Sex-specific progression of Parkinson's disease: A longitudinal mixed-models analysis. *J Parkinsons Dis* 2025; 15:805–18. DOI: 10.1177/1877718X251339201
16. Miller SA, Mayol M, Moore ES, Heron A, Nicholas V, and Ragano B. Rate of progression in activity and participation outcomes in exercisers with Parkinson's disease: A five-year prospective longitudinal study. *Parkinsons Dis* 2019; 2019:5679187. DOI: 10.1155/2019/5679187
17. Bralić A et al. Correspondence of MRI and nTMS with EDSS in multiple sclerosis: longitudinal follow-up study. *Ann Clin Transl Neurol* 2025; 12:1240–55. DOI: 10.1002/acn3.70041
18. Lin TI and Lee JC. A robust approach to t linear mixed models applied to multiple sclerosis data. *Stat Med* 2006; 25:1397–412. DOI: 10.1002/sim.2384
19. McNeish D and Wolf MG. Thinking twice about sum scores. *Behav Res Methods* 2020; 52:2287–305. DOI: 10.3758/s13428-020-01398-0
20. Pascual-Morena C et al. Efficacy of risdiplam in spinal muscular atrophy: A systematic review and meta-analysis. *Pharmacotherapy* 2024; 44:97–105. DOI: 10.1002/phar.2866
21. Hagenacker T et al. Effectiveness of nusinersen in adolescents and adults with spinal muscular atrophy: systematic review and meta-analysis. *Neurol Ther* 2024; 13:1483–504. DOI: 10.1007/s40120-024-00653-2
22. Kingma DP and Welling M. Auto-encoding variational bayes. *arXiv preprint* 2013. DOI: 10.48550/arXiv.1312.6114
23. Kolen MJ and Brennan RL. Test equating, acaling, and linking: methods and practices. New York, NY, USA: Springer, 2014. DOI: 10.1007/978-1-4939-0317-7

24. Wirth RJ and Edwards MC. Item factor analysis: current approaches and future directions. *Psychol Methods* 2007; 12:58–79. DOI: 10.1037/1082-989X.12.1.58
25. Skrondal A and Rabe-Hesketh S. Generalized latent variable modeling: multilevel, longitudinal, and structural equation models. Boca Raton, FL, USA: Chapman and Hall/CRC, 2004. DOI: 10.1201/9780203489437
26. Bauer DJ and Hussong AM. Psychometric approaches for developing commensurate measures across independent studies: traditional and new models. *Psychol Methods* 2009; 14:101–25. DOI: 10.1037/a0015583
27. Kolen MJ. Population invariance in equating and linking: Concept and history. *J Educ Meas* 2004; 41:3–14. DOI: 10.1111/j.1745-3984.2004.tb01155.x
28. Buuren S van, Eyres S, Tennant A, and Hopman-Rock M. Improving comparability of existing data by response conversion. *J Off Stat* 2005; 21:53–72. DOI: 10.1111/1468-0068.00319
29. Heuvel ER van den, Griffith LE, Sohel N, Fortier I, Muniz-Terrera G, and Raina P. Latent variable models for harmonization of test scores: a case study on memory. *Biometrical Journal* 2020; 62:34–52. DOI: 10.1002/bimj.201800146
30. Csurka G. A comprehensive survey on domain adaptation for visual applications. *Domain adaptation in computer vision applications*. Ed. by Csurka G. Springer, 2017 :1–35. DOI: 10.1007/978-3-319-58347-1_1
31. Ganin Y and Lempitsky V. Unsupervised domain adaptation by backpropagation. *ICML 2015*. Vol. 37. 2015 :1180–9
32. Guan H and Liu M. Domain adaptation for medical image analysis: A survey. *IEEE Trans Biomed Eng* 2022; 69:1173–85. DOI: 10.1109/TBME.2021.3117407
33. Götz M et al. DALSA: Domain adaptation for supervised learning from sparsely annotated MR images. *IEEE Trans Med Imaging* 2016; 35:184–96. DOI: 10.1109/TMI.2015.2463078
34. Ghafoorian M et al. Transfer learning for domain adaptation in MRI: application in brain lesion segmentation. *Medical image computing and computer assisted intervention – MICCAI 2017, LNCS 10435*. 2017 :516–24. DOI: 10.1007/978-3-319-66179-7_59

35. Becker C, Christoudias CM, and Fua P. Domain adaptation for microscopy imaging. *IEEE Trans Med Imaging* 2015; 34:1125–39. DOI: 10.1109/TMI.2014.2376872
36. Chen M, Xu Z(E), Weinberger KQ, and Sha F. Marginalized denoising autoencoders for domain adaptation. *ICML 2012*. 2012 :767–74. DOI: 10.1145/2339530.2339636
37. Long M, Cao Y, Wang J, and Jordan MI. Learning transferable features with deep adaptation networks. *ICML 2015*. 2015 :1180–9. DOI: 10.1145/2783258.2788606
38. Tzeng E, Hoffman J, Saenko K, and Darrell T. Adversarial discriminative domain adaptation. *CVPR 2017*. 2017 :2962–71. DOI: 10.1109/CVPR.2017.316
39. Couronné R, Vernhet P, and Durrleman S. Longitudinal self-supervision to disentangle inter-patient variability from disease progression. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021 (Lecture Notes in Computer Science 12902)*. 2021 :231–41. DOI: 10.1007/978-3-030-87196-3_22
40. Sauty B and Durrleman S. Progression models for imaging data with longitudinal variational auto encoders. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. 2022 :3–13. DOI: 10.1007/978-3-031-16431-6_1
41. Ong P, Haußmann M, Lönnroth O, and Lähdesmäki H. Latent mixed-effect models for high-dimensional longitudinal data. *arXiv preprint* 2024. DOI: 10.48550/arXiv.2409.11008
42. Köber G et al. Individualizing deep dynamic models for psychological resilience data. *Sci Rep* 2022; 12:8061. DOI: 10.1038/s41598-022-11650-6
43. Hackenberg M, Pechmann A, Kreutz C, Kirschner J, and Binder H. A statistical approach to latent dynamic modeling with differential equations. *arXiv preprint* 2023. DOI: 10.48550/arXiv.2311.16286
44. Hackenberg M et al. Deep dynamic modeling with just two time points: can we still allow for individual trajectories? *Biometrical Journal* 2022; 64:1426–45. DOI: 10.1002/bimj.202000366
45. Rubanova Y, Chen RTQ, and Duvenaud DK. Latent ODEs for irregularly-sampled time series. *arXiv preprint* 2019. DOI: 10.48550/arXiv.1907.03907

46. Fortuin V, Baranchuk D, Rätsch G, and Mandt S. GP-VAE: deep probabilistic time series imputation. *AISTATS 2020*. Vol. 108. 2020 :1651–61. DOI: 10.1109/CVPR.2017.316
47. Alaa A, Mayer E, and Barahona M. ICE-NODE: integration of clinical embeddings with neural ordinary differential equations. *MLHC 2022*. 2022 :537–64. DOI: 10.5555/3511112.3511162
48. Hackenberg M, Pfaffenlehner M, Behrens M, Pechmann A, Kirschner J, and Binder H. Investigating a domain adaptation approach for integrating different measurement instruments in a longitudinal clinical registry. *arXiv preprint 2023*. DOI: 10.48550/arXiv.2312.00616
49. Pechmann A et al. SMARtCARE - a platform to collect real-life outcome data of patients with spinal muscular atrophy. *Orphanet J Rare Dis* 2019; 14:18. DOI: 10.1186/s13023-019-0998-4
50. Blei DM, Kucukelbir A, and McAuliffe JD. Variational inference: a review for statisticians. *J Am Stat Assoc* 2017; 112:859–77. DOI: 10.1080/01621459.2017.1285773
51. Higgins I et al. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. *ICLR 2017*. 2017. Available from: <https://openreview.net/forum?id=Sy2fzU9gl>
52. Chipperfield JO and Steel DG. Multivariate random effect models with complete and incomplete data. *Journal of Multivariate Analysis* 2012; 109:146–55. DOI: <https://doi.org/10.1016/j.jmva.2012.02.014>
53. Amemiya Y. On multivariate mixed model analysis. *NIPS 2004*. 2004 :777–84. DOI: 10.1214/lnms/1215463787
54. Pinheiro JC and Bates DM. *Mixed-Effects Models in S and S-PLUS: Methods and Practices*. New York, NY, USA: Springer, 2000. DOI: 10.1007/b98882
55. Latra IN, Linuwih S, Purhadi, and Suhartono. Estimation for multivariate linear mixed models. *Int J Basic Appl Sci* 2010; 10:79–86
56. Kingma DP and Ba J. Adam: a method for stochastic optimization. *arXiv preprint* 2014. DOI: 10.48550/arXiv.1412.6980
57. Wilks SS. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann Math Statist* 1938; 9:60–2. DOI: 10.1214/aoms/1177732360

58. Taylor J and Tibshirani RJ. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences* 2015; 112:7629–34. DOI: 10.1073/pnas.1507583112
59. Lee JD, Sun DL, Sun Y, and Taylor JE. Exact post-selection inference, with application to the lasso. *Ann Statist* 2016; 44:907–27. DOI: 10.1214/15aas1371
60. Barber RF and Candès EJ. Controlling the false discovery rate via knockoffs. *Ann Statist* 2015; 43:2055–85. DOI: 10.1214/15-AOS1337
61. Nguyen TB, Chevalier JA, Thirion B, and Arlot S. Aggregation of multiple knockoffs. *ICML 2020, PMLR 119*. 2020 :7283–93
62. Efron B. Bootstrap methods: another look at the jackknife. *Ann Statist* 1979; 7:1–26. DOI: 10.1214/aos/1176344552
63. Kolb SJ and Kissel JT. Spinal muscular atrophy: a timely review. *Arch Neurol* 2011; 68:979–84. DOI: 10.1001/archneurol.2011.74
64. Yeo CJJ, Tizzano EF, and Darras BT. Challenges and opportunities in spinal muscular atrophy therapeutics. *Lancet Neurol* 2024; 23:205–18. DOI: 10.1016/S1474-4422(23)00419-2
65. Levina E and Bickel PJ. Maximum likelihood estimation of intrinsic dimension. *NIPS 2004*. 2004 :777–84
66. Nazábal A, Olmos PM, Ghahramani Z, and Valera I. Handling incomplete heterogeneous data using VAEs. *Pattern Recogn* 2020; 107:107501. DOI: 10.1016/j.patcog.2020.107501
67. Liu DC and Nocedal J. On the limited memory method for large scale optimization. *Math Program* 1989; 45:503–28. DOI: 10.1007/BF01589116

5 Funding

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 499552394 – SFB 1597 Small Data.

6 Data preprocessing and Hyperparameters

In collaboration with our clinical partners from the SMArtCARE registry, we restricted our analysis to patients who underwent a treatment switch. To define this

cohort, we filtered these patients based on the following criteria. First, patients who received less than six months of treatment with a medication before or after the switch were excluded. For patients who switched medication more than twice, we included their data only up to the second treatment switch. Further, we required that patients had motor-function evaluations at a minimum of four visits, including at least two visits before and after the switch. Lastly, we excluded switches to medications with fewer than 10 patients, which left the dataset with switches to two different SMA medications.

This resulted in 522 patients with an average of 17.3 hospital visits and a mean of 2.1 out of 5 tests used to evaluate motor function at each visit. The median period under observation period was 5.7 years, while the median patient age was 6.7 years.

We binarized each test item using thermometer encoding, since physicians grade motor function on ordinal categorical scales comprising between two and six levels per test item. We also included an indicator variable for each motor function item that was set to one if a test item could not be performed by a patient. To capture the data distribution, we chose an ordinal logistic model as decoder distribution [66].

Missing values within an individual assessment were imputed by zero-filling, reflecting the judgment of our clinical partners that such items were typically skipped due to patient exhaustion. However, observations with more than a quarter of missing values were excluded from analysis, as some child patients occasionally refuse to perform a larger set of tasks due to bad mood despite being physically able to perform them.

Some covariates were split up into multiple fixed effects; these included the *SMN2* count, which was one-hot encoded into two categories ($SMN2 \leq 2$ and $SMN2 \geq 3$). Age at symptom onset was divided into a continuous variable representing the age at symptom onset in years for symptomatic patients and an indicator variable showing whether a patient was still presymptomatic. Years since treatment switch was divided into two separate variables each indicating the years since switching to medication A and B. We standardized the covariates except time since medication switch and included interaction terms with age for all fixed effects (except age itself). We included an intercept for data-level mixed models, but not in the latent model as latent variables are centered around zero to follow the standard normal prior distribution.

We chose dense neural networks consisting of two hidden layers with 250 and 100 units and $\tanh$ activation function within encoder and decoder network

architectures. We trained neural networks with the Adam optimizer [56] with standard hyperparameters and mixed models using the L-BFGS algorithm with $\eta = 0.15$ [67]. In each epoch we updated the VAE parameters 100 times on the whole dataset before refitting the mixed model parameters. We trained for 20 epochs, after which we observed no further improvement in total loss and reconstruction quality for more than two epochs. The latent space dimensionality was set to $d = 3$, chosen by domain knowledge to be expressive enough for three aspects of SMA. First, signal forwarding, meaning the short-term efficiency of neuromuscular transmission, second motor neuron integrity, reflecting the slowly varying capacity of the motor unit pool, and lastly the development of motor function. In the loss function, we down-weighted the KL divergence loss term by 0.5 because of the modest data dimensionality which led to an impact of reconstruction quality for higher KL-weighting. The weighted KL term converged to contribute roughly 10% to the loss. The alignment terms were up-weighted by 5 and converged to contribute about 2.5% each to the final loss, placing primary emphasis on reconstruction quality. Model performance remained stable when these coefficients were halved or doubled. Increasing the alignment terms tended to increase skewness of the empirical test-statistic distribution, whereas decreasing them led to larger deviations between encoder outputs and latent mixed-model predictions. We chose a diagonal covariance structure for Φ, Σ . We found that alternative covariance structures did not contribute to a better reconstruction quality of test items, while conditional residuals were approximately uncorrelated.

To simulate an artificial treatment effect, we first computed the incremental score to be added for each observation; no extra score was added to measurements before the treatment switch, after the switch an artificial improvement of one item every six months per test was applied. This increment was implemented by randomly selecting test items for each measurement instrument that had not yet reached their maximum value and increasing their score by one for each subsequent observation. If an item was already at its full score in one of these subsequent observations, the additional points were allocated to another randomly chosen item if possible.

7 Tables

Full Name	Abbr.	Test Description	Details	
Revised Upper Limb Module[5]	RULM	Measures upper limb function and strength across tasks like reaching, lifting, and hand movements.	Target Median Age Items Max Score Inputs Obs.	Children 11.1 20 + 1 37 + 6 63 3625
Hammersmith Functional Motor Scale Expanded[6]	HFMSE	Assesses gross motor skills, including activities like sitting, rolling, standing, and transitional movements.	Target Median Age Items Max Score Inputs Obs.	Children 9.6 33 + 1 66 + 6 104 3279
Children's Hospital of Philadelphia Infant Test of Neuromuscular Disorders[7]	CHOP-INTEND	Assesses neuromuscular function, concentrating on spontaneous and prompted movements.	Target Median Age Items Max Score Inputs Obs.	Infants 3.0 16 64 68 1833
Hammersmith Infant Neurological Examination, Section 2[8]	HINE-2	Monitors early development, especially the attainment of motor function milestones (sitting, standing, walking).	Target Median Age Items Max Score Inputs Obs.	Infants 4.4 8 + 3 26 + 3 40 4071
ALS Functional Rating Scale Revised[9]	ALSFRS-R	Covers speech, swallowing, and fine motor tasks; designed for ALS but also used for SMA.	Target Median Age Items Max Score Inputs Obs.	Adults 34.9 12 + 1 48 + 4 65 732

Table 1: Measurement instruments to assess a patient's motor function in SMART-CARE. Some tests contain additional items that are not used to determine the official test score. However, we still include these items to increase the amount of information.

Symbol	Meaning
$i \in \mathcal{I}$	Patient index
$t \in \mathcal{T}_i, t \in \mathcal{T}_{i,l}$	Observation times
$l \in \mathcal{L}, l \in \mathcal{L}_{i,t}$	Measurement instrument index
$b \in \mathcal{B}$	Bootstrap samples
n, n_l	Data dimension
$m_i, m_{i,l}$	Number of observations
d	Latent dimension
p, q	Number of fixed and random effect coefficients
rd	Difference in parameters between nested models
$Y, \mathbf{y}_i / Z, \mathbf{z}_i$	Single data/latent observation
$Y_i, \mathbf{Y}_i = (\mathbf{y}_{i,t})_{t \in \mathcal{T}_i}^\top$	Data trajectory
$Z_i, \mathbf{Z}_i = (\mathbf{z}_{i,t})_{t \in \mathcal{T}_i}^\top$	(Latent) response
$Y_{i,l}, \mathbf{Y}_{i,l} = (\mathbf{y}_{i,t,l})_{t \in \mathcal{T}_{i,l}}^\top$	Instrument specific data trajectory
$Z_{i,l}, \mathbf{Z}_{i,l} = (\mathbf{z}_{i,t,l})_{t \in \mathcal{T}_{i,l}}^\top$	Instrument specific latent trajectory
$\hat{\mathbf{Z}}_i = (\hat{\mathbf{z}}_{i,t})_{t \in \mathcal{T}_i}^\top$	Latent mixed model estimation
$\mathbf{X}_i, \mathbf{T}_i$	Design matrices for fixed and random effects
$\mathbf{B}, \hat{\mathbf{B}}$	Fixed effect coefficient matrix
$U_i, \mathbf{U}_i, \hat{\mathbf{U}}_i$	Random effects
$E_i, \mathbf{E}_i$	Residual error
$\Phi, \Sigma, \mathbf{V}_i$	Random effects covariance, residual covariance, precision
$\mathbf{I}_d$	Identity matrix of dimension d
$P_{Y,Z}, p_{Y,Z}(\mathbf{y}, \mathbf{z})$	Joint generative distribution of (Y, Z)
$P_Z, p_Z(\mathbf{z})$	Prior on latent variable
$P_{Y \mathbf{z}}, p_{Y \mathbf{z}}(\mathbf{y})$	Conditional sampling distribution of Y given latent $\mathbf{z}$
$P_Y, p_Y(\mathbf{y})$	Marginal distribution of Y
$P_{Z \mathbf{y}}, p_{Z \mathbf{y}}(\mathbf{z})$	True posterior of Z given $\mathbf{y}$
$Q_{Z \mathbf{y}}, q_{Z \mathbf{y}}(\mathbf{z})$	Variational posterior of Z given $\mathbf{y}$
$\text{enc}_\phi, \text{enc}_{\phi_l}, \text{dec}_\theta, \text{dec}_{\theta_l}$	Encoder/decoder neural network
$\mathcal{L}_{\text{VAE}}$	negative ELBO loss for VAE
$\mathcal{L}_{\text{ML}}, \mathcal{L}_{\text{REML}}$	(Restricted) ML log-likelihood
$\mathcal{L}_{\text{MMVAE}}$	Joint loss for multi-measurement VAE + mixed model
β, η, γ	Weighting of KL and alignment terms in $\mathcal{L}_{\text{MMVAE}}$

Table 2: Notation used throughout the manuscript.

Instrument	Real			Artificial +2	
	Mean	SD	Percentage	Mean	SD
HINE-2	0.87	0.20	3.00	2.84 (+1.97)	0.57
RULM	1.47	0.12	5.09	2.62 (+1.15)	0.41
CHOP	0.91	0.31	3.17	2.64 (+1.73)	0.46
HFMSE	1.56	0.38	5.37	2.72 (+1.16)	0.59
ALSFRS-R	0.52	0.30	1.81	1.30 (+0.78)	0.60

Table 3: Mean predicted item-level differences for five motor function instruments (averaged over all patients and ten random seeds) between predictions with and without treatment switch one year after the switch took place. Results are based on a three-dimensional latent mixed effects trajectory. Positive values indicate improved motor function compared to no switch. The SD column lists the standard deviation of patient-level mean values across the ten random seeds. The percentage column converts absolute differences to a percentage of the instrument’s maximum sum-score to enable cross-instrument comparability. The left block refers to the switch detected in the unaltered data; the right to an artificial switch of +2 items per year. Here, changes in mean values w.r.t. unaltered trajectories are indicated in brackets.