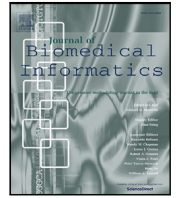




Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

A framework for longitudinal latent factor modelling of treatment response in clinical trials with applications to Psoriatic Arthritis and Rheumatoid Arthritis

Fabian Falck ^{a,c,1}, Xuan Zhu ^{b,1}, Sahra Ghalebikesabi ^a, Matthias Kormaksson ^b, Marc Vandemeulebroecke ^j, Cong Zhang ^d, Ruvie Martin ^b, Stephen Gardiner ^e, Chun Hei Kwok ^f, Dominique M. West ^g, Luis Santos ^c, Chengeng Tian ^d, Yu Pang ^d, Aimee Readie ^b, Gregory Ligozio ^b, Kunal K. Gandhi ^b, Thomas E. Nichols ^{e,h}, Ann-Marie Mallon ^c, Luke Kelly ⁱ, David Ohlssen ^b, George Nicholson ^{a,*}

^a Department of Statistics, University of Oxford, UK^b Novartis Pharmaceuticals Corporation, East Hanover, United States^c The Alan Turing Institute, London, UK^d China Novartis Institutes for Bio-medical Research CO., Shanghai, China^e Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Population Health, University of Oxford, UK^f Medical Research Council Harwell Institute, UK^g Radcliffe Department of Medicine, University of Oxford, UK^h Wellcome Centre for Integrative Neuroimaging, Nuffield Department of Clinical Neurosciences, University of Oxford, UKⁱ School of Mathematical Sciences, University College Cork, Ireland^j UCB Farchim SA, Bulle, Switzerland

ARTICLE INFO

Dataset link: <https://clinicalstudydatarequest.com/>

Keywords:

Longitudinal latent factor analysis
Linear mixed-effects model
Multilevel linear models
Probabilistic principal component analysis
Dimensionality reduction
Missing data
Clinical trials
Psoriatic Arthritis
Rheumatoid Arthritis

ABSTRACT

Objective: Clinical trials involve the collection of a wealth of data, comprising multiple diverse measurements performed at baseline and follow-up visits over the course of a trial. The most common primary analysis is restricted to a single, potentially composite endpoint at one time point. While such an analytical focus promotes simple and replicable conclusions, it does not necessarily fully capture the multi-faceted effects of a drug in a complex disease setting. Therefore, to complement existing approaches, we set out here to design a longitudinal multivariate analytical framework that accepts as input an entire clinical trial database, comprising all measurements, patients, and time points across multiple trials.

Methods: Our framework composes probabilistic principal component analysis with a longitudinal linear mixed effects model, thereby enabling clinical interpretation of multivariate results, while handling data missing at random, and incorporating covariates and covariance structure in a computationally efficient and principled way.

Results: We illustrate our approach by applying it to four phase III clinical trials of secukinumab in Psoriatic Arthritis (PsA) and Rheumatoid Arthritis (RA). We identify three clinically plausible latent factors that collectively explain 74.5% of empirical variation in the longitudinal patient database. We estimate longitudinal trajectories of these factors, thereby enabling joint characterisation of disease progression and drug effect. We perform benchmarking experiments demonstrating our method's competitive performance at estimating average treatment effects compared to existing statistical and machine learning methods, and showing that our modular approach leads to relatively computationally efficient model fitting.

Conclusion: Our multivariate longitudinal framework has the potential to illuminate the properties of existing composite endpoint methods, and to enable the development of novel clinical endpoints that provide enhanced and complementary perspectives on treatment response.

* Corresponding author.

E-mail address: george.nicholson@stats.ox.ac.uk (G. Nicholson).¹ Denotes joint first authors who have worked together in this publication and contributed equally.<https://doi.org/10.1016/j.jbi.2024.104641>

Received 2 December 2023; Received in revised form 10 March 2024; Accepted 11 April 2024

Available online 18 April 2024

1532-0464/© 2024 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Clinical trials gather a wealth of data comprising high-dimensional measurements from patients over multiple time points spanning months or even years. Most diseases are multi-faceted, and some trials are indeed designed to evaluate drug efficacy based on more than one endpoint. The dimensionality of a single cross-sectional observation on a patient can extend to the order of hundreds of measurements, including a broad range of clinical events, patient symptoms, physical function, molecular biomarkers, quality of life measurements, or surrogates of these events or symptoms [1]. To draw clinically relevant conclusions in this high-dimensional and longitudinal treatment-response space, it is common practice to calculate and analyse a cross-sectional composite endpoint, i.e. a function of multiple measurements taken at a pre-specified number of time points post-baseline [2]. These traditional composite endpoints are based on clinical domain knowledge of the disease and treatment effect, and are often derived heuristically; they provide a single, standardised, clinically motivated outcome with which to demonstrate and prove efficacy of a treatment in a conceptually simple and analytically attractive way [2]. For example, in *Psoriatic arthritis*, the Psoriatic Arthritis Disease Activity Score (PASDAS) summarises patient and physician global assessments, patient-reported measurements, peripheral joint counts, dactylitis, enthesitis, and acute-phase response into a single score.

However, by focusing on a single, heuristically derived cross-sectional endpoint, it is possible to overlook potentially important information about treatment effects and disease progression [1,2]. For example, longitudinal patterns in clinical trials contain additional information on the rate and nature of treatment effects [3] and can offer a better understanding of heterogeneity in populations, yielding insights into the determinants of progression and response [4]. Additionally, high-dimensional multivariate measurements may harbour clinically informative variation that is not captured by existing composite endpoints. There is the further benefit that correlation structure across multiple measurements and time points can be exploited to perform inference even in the presence of missing data, whereby a subset of measurements or survey responses are unavailable for a patient at one or several occasions [4].

In modelling longitudinal clinical trials data we face a combination of three main statistical challenges. First, the longitudinal dependence structure of observations, often gathered at irregular time points, requires the use of structured statistical models to address autocorrelation in a principled way [5–7]. Second, multivariate measurement data benefit from the use of specialised methods, such as sparse factor models, to enable effective joint interpretation [4]. Third, the data present with structured patterns of missingness. To address these points we combine probabilistic principal component methods with linear mixed-effects models, so as to account respectively for the correlation structure between measurements and the correlation structure over time, leading to efficient and interpretable inference in the presence of missing data.

In this work, we develop a framework for coherently combining information on ATEs across all time points, measurements, and patients in a pool of clinical trials. The novelty of our approach lies in its *composability*, whereby we split a full longitudinal factor model into two sub-models, and fit each sub-model by well-established methods and software (factor analysis in the first sub-model, and longitudinal mixed effects models in the second). The resulting algorithm is convenient, flexible, and fast, because within each of the sub-models we can select from a range of existing, optimised software tools according to our modelling requirements. This leads to simplified high-level code development, and stable, efficient inference at run-time as each sub-model fit involves optimising fewer parameters over less data compared to fitting the full model.

Statement of Significance

Summary	Description
Problem	Clinical trial efficacy is primarily assessed using a single one-dimensional endpoint, which could be usefully complemented by a multivariate clinical perspective.
What is already known	Clinical trials provide a wealth of diverse, longitudinally measured clinical variables. However, by focussing narrowly on a single measurement and time point, analyses may overlook clinically interesting longitudinal and multivariate dynamics.
What this paper adds	This paper introduces a longitudinal latent factor model capable of processing all measurements and time points of multiple clinical trials. Its interpretable nature facilitates the characterisation of disease progression and treatment effects, yielding clinically meaningful insights.

2. Related work

Longitudinal factor analysis has been extensively studied and applied in practice [8,9]. The literature subsumes works on item response theory [10–13] which focus on binary and ordinal survey data, on latent trait models [14–16] that consider binary response variables, and on latent growth curve models [17–19]. Within biomedical informatics, factor models are applied to disease prediction [20–22], gene selection [23,24], pattern recognition [25], disease subtyping [26], and individual treatment effect estimation [27]. We consider the latter where longitudinal latent factor models have been proven to be a valuable instrument, especially within clinical trial analysis.

Here we consider a subclass of latent factor models, namely longitudinal multivariate mixed-effect models [28–31] that incorporate two conceptual components: (i) how the disease state of a patient changes over time; and (ii) how these changes vary across patients. While most latent growth models focused on bivariate and ordinal outcome variables [32,33], there exists research modelling continuous outcomes [34] which we add to by considering both discrete and continuous outcome variables.

Our approach here involves two stages, and in the following paragraphs we discuss the related work to each of these in turn. Our first stage model, PPCA [35], is a probabilistic extension of principal component analysis [36] which enables estimation of latent factors in the presence of missing data [37]. PCA and PPCA are widely applied in biomedical informatics for identifying groups of genes with similar expression profiles [25,38], missing value estimation [39] and dimensionality reduction [40]. As such PCA is often used in conjunction with other approaches. Menaga et al. for instance, apply PPCA in conjunction with Deep Belief Networks [40]. Several prior works discuss the scenarios in which either factor analysis or PCA should be used [41–43] but, to the best of our knowledge, PPCA has not yet been leveraged *within* latent growth curve models.

In a second stage, we characterise the longitudinal variation in the latent factors found in the first stage by implementing a Linear Mixed-Effects (LME) model, as first formulated by [44], which is solved using ordinary or restricted maximum likelihood estimation [45,46]. Theoretical properties of LMEs and methods for efficient implementation have been discussed in detail [47–49], and extensive R packages facilitate their use by data practitioners [50,51]. Multiple generalisations of LMEs to more complex interaction modelling exist [52–54]. A straightforward extension to our Stage 2 model would be to incorporate

longitudinal smoothness assumptions on treatment effects, which can be achieved using regularised Generalised Additive Models [55–59]. An analogous approach to ours combined functional Principal Components Analysis (PCA) with LME modelling [60]. Other previous research, however, considered trading off model flexibility for computational complexity in related contexts [61], for example assuming conditional independence of the outcome variables given the latent variables [62, 63]. Please see Section A.11 for additional discussion of related work.

3. Introduction to case study

3.1. Clinical background

Rheumatoid Arthritis (RA), which is a systemic autoimmune disease characterised by symmetric synovitis, triggering cartilage damage and joint destruction, is one of the most prevalent chronic inflammatory diseases which primarily involves the joints and is complicated by numerous extra-articular manifestations [64–66]. Psoriatic Arthritis (PsA) is a chronic inflammatory and musculoskeletal condition. It comes with several comorbidities and is characterised by a variety of symptoms including arthritis, enthesitis, dactylitis, spondylitis, psoriasis and nail disease. It may adversely affect patients' health-related quality of life (HR-QOL) and cause significant disability [67–71].

Secukinumab (brand name Cosentyx), a human immunoglobulin G1- κ monoclonal antibody that directly inhibits interleukin 17A, has demonstrated long-term improvements in the signs and symptoms of patients with active PsA. Secukinumab has been an approved treatment for psoriasis, axial spondyloarthritis, PsA, and Enthesitis-Related Arthritis [65,72].

The primary endpoint to assess drug efficacy in both PsA and RA clinical trials in current practice is a criterion called ACR20, developed by the American College of Rheumatology [73]. This criterion is a composite, binary measurement that assesses whether or not the patient experiences both: (i) a 20% improvement in the number of tender and in the number of swollen joints; and (ii) a 20% improvement in three of the following five components: patient global assessment, physician global assessment, functional ability measure (most often Health Assessment Questionnaire [HAQ]), visual analog pain scale, and high sensitivity C-reactive protein (CRP) or erythrocyte sedimentation rate. Note that the measurements used in ACR20 consist of 7 of the 12 key endpoints we consider in this work (see Table A.2). While ACR20 is a well-established composite measure for evaluating treatment effect in PsA and RA, it heuristically subsumes a rich set of multi-dimensional information into a single, binary measure. Additional key endpoints collected in PsA and RA clinical trials (such as other quality of life domains) are not included in ACR20 [74].

3.2. Data description

This case study analyses two Phase III Cosentyx PsA trials (FUTURE-2, FUTURE-5 [75–78]) and two Phase III Cosentyx RA trials (REASURE, NURTURE-1 [65,79–82]). An anonymised version of all trial data can be requested through a voluntary data-sharing process on <https://clinicalstudydatarequest.com/>. These four trials were chosen because of similar study design, treatment regimen, visit window for key endpoint assessments, and eligibility criteria. Table A.1 shows the number of patients stratified by treatment arm for each of these trials and the NCT unique identification number of each clinical study registered on ClinicalTrials.gov.

We jointly analysed twelve efficacy endpoints, which described key disease characteristics and symptoms for PsA and RA indications including objective measurements (e.g. laboratory result, DAS, joint stiffness and swelling), Quality of Life (e.g. HAQDI), and pain (see Table A.2). Note that the raw measurements analysed in the case study are either ordinal or continuous (see Table A.2), and each measurement is preprocessed using a rank-based inverse-normal transformation (see

Sections A.2 and A.10 for details). Collectively, they provide a holistic, partly overlapping description of a patient's condition. The patient's condition is what we aim to uncover and summarise as a latent variable in this work, providing a time-dependent, possibly richer description compared to ACR20. We note that the latent factor method is also applicable to a larger number of endpoints than illustrated here, while being able to handle missingness in the data.

Endpoints are partially observed at weeks 0 (baseline), 2, 4, 8, 12 and 16. We included measurement data up to week 16, with the later weeks omitted from this analysis due to possible treatment switches of subjects and the complex branching of study designs between the four clinical trials thereafter. In Figure A.6, a heatmap illustrates the data availability in each of the four trials by time point and measurement.

4. Methods

In this section we provide an overview of our longitudinal factor model, which we fit using a composable two-stage approach. We present the core building blocks of our model in a simplified and easily comprehensible way, referring to Section A for full technical details. Our starting point is clinical trial measurement input data for patient i , comprising a $P \times T$ matrix Y_i with rows corresponding to the P measurements and columns to the T time points (see the top panels of Fig. 1 for a single patient's data with $P = 12$ measurements and $T = 6$ time points). Our longitudinal factor approach has two key stages. The first stage receives as input the P -dimensional measurements $y_{i,:t}$ which can be represented by a $P \times K$ matrix A ($K < P$) and latent variables $z_{i,:t}$ as:

$$y_{i,:t} = Az_{i,:t} + \epsilon_{i,t} \quad (1)$$

where the vector $\epsilon_{i,t}$ denotes white Gaussian noise $\epsilon_{i,t} \sim \mathcal{N}(0, I\sigma_\epsilon^2)$. We will illustrate this linear combination in Section 4.1. The latent matrix A is orthogonal and *sparse* (i.e. a matrix with most values close to zero). This improves the interpretability of the factor loadings as each factor describes a different aspect of the measurements, and is enforced to use as few measurements as possible.

The second stage of our approach models the longitudinal component of the latent variables:

$$z_{i,k,:} = \begin{cases} \mu_{k,\text{placebo}} + \xi_{i,k} & \text{if } i \text{ is in placebo arm} \\ \mu_{k,\text{active}} + \xi_{i,k} & \text{if } i \text{ is in active arm} \end{cases} \quad (2)$$

where $\mu_{k,\text{placebo}}$ and $\mu_{k,\text{active}}$ are T -vectors representing the average longitudinal trajectory of the k th latent variable in the placebo and active treatment arms. The vector $\xi_{i,k}$ denotes Gaussian noise that is correlated across T time points: $\xi_{i,k} \sim \mathcal{N}(0, \Sigma)$ for some $T \times T$ covariance matrix Σ that is learnt as part of the model fitting. We next describe and illustrate stages 1 and 2 further.

4.1. Stage 1: PPCA factor model

In Stage 1 we perform the dimension reduction in (1) from $P \times T$ dimensional Y_i to $K \times T$ dimensional Z_i using a sparse PPCA model fit [35,83]. This is illustrated in Fig. 1, where raw data gathered from twelve selected clinical endpoints between weeks 0–16 are displayed at the top, serving as inputs to the method. These measurements are weighted by the *loadings* matrix A (illustrated as a heatmap in the centre of Fig. 1 presents the matrix A^T). These weights can be interpreted as taking a linear combination of the data measurements to describe a latent state of a patient in the form of the latent variables $z_{i,:t}$. To see this, note that as A is orthogonal, we have $A^T A = I$, so that pre-multiplying (1) by A^T yields the latents $z_{i,:t}$ as noisy linear combination of the observed measurements: $z_{i,:t} = A^T y_{i,:t} - A^T \epsilon_{i,t}$. The weighted measurement combinations are referred to as *scores* Z , and this patient's estimated scores for the three factors are shown in the right-hand panels of Fig. 1. The error bands on the scores represent posterior uncertainty around the true scores for this patient, which

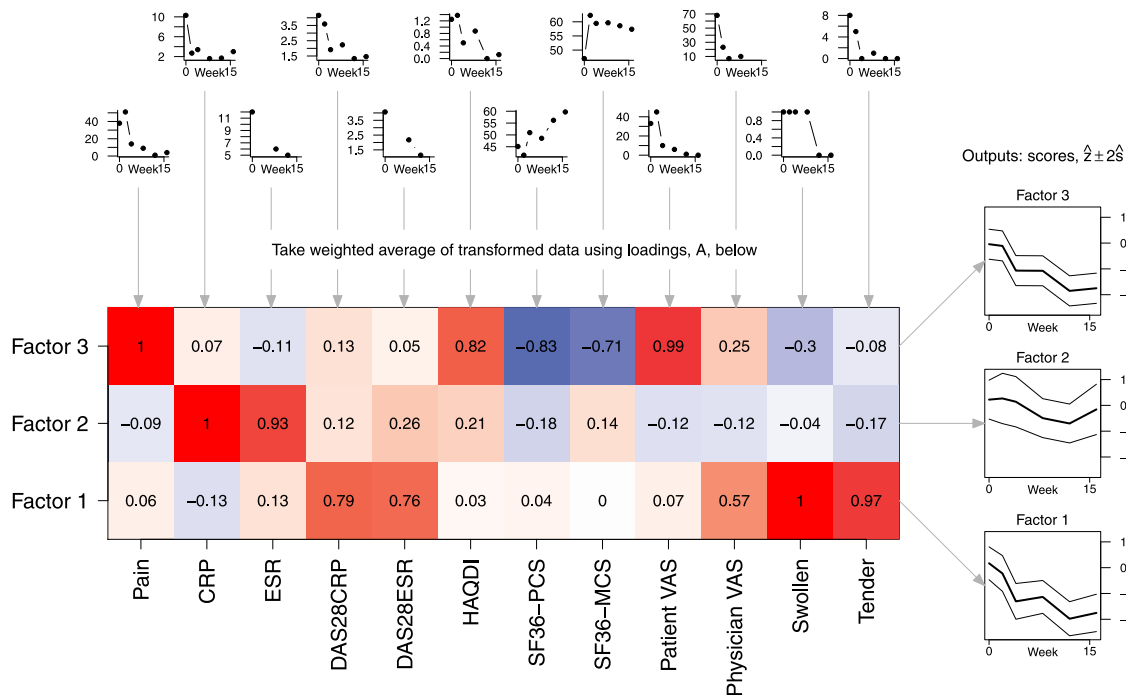


Fig. 1. Illustration of the PPCA factor model in Stage 1. The raw data measurements gathered from a particular patient between 0 and 16 weeks are shown at the top, for 12 clinical endpoints. The heatmap displays the *loadings* matrix A . In the heatmap, the weight (loading) values are given and also represented by colour (blue to red ranging from -1 to 1). The output of the factor analysis are the *scores* \hat{Z} , which are combinations of the data measurements weighted by the loadings, plotted for the three factors on the right. The scores are augmented by error bands \hat{S} representing posterior uncertainties around the true scores of this patient (approximate posterior credible intervals of ± 2 SEs are shown). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

can be quantified by sharing information about the typical size of measurement errors across all patients.

The interpretation of each factor can be determined by inspecting which raw measurements are its main constituents, as shown in the heatmap of loadings in the centre of Fig. 1. An important feature of the heatmap is that it shows how the loadings have been transformed to be *sparse*, by which we mean that several of a factor's weighting coefficients are close to zero, allowing us to interpret the factor as a disease domain consisting of a relatively small subset of measurements (Section A.4). Here, the heatmap depicts the relative weights of outcome measurements in three disease domains corresponding to the three factors. The colour intensity represents the absolute weight of each measurement in each disease domain, with red and blue representing positive and negative weights, respectively.

For example, Factor 1 has large positive coefficients in Tender, Swollen, Physician VAS, DAS28ESR and DAS28CRP, indicating that these measurements tend to co-vary with one another, and we can broadly associate this factor with the signs and symptoms in the joints. Similarly, we can relate Factor 2 to Inflammatory markers (CRP, ESR), and Factor 3 to Pain, Physical function, and Quality of Life (QoL). In the last factor, the positively weighted endpoints (e.g., Pain) move in the same direction as the factor, while the negatively weighted endpoints (e.g., SF36-PCS) move in the opposite direction.

In terms of choosing a number of factors, there are a variety of objective quantitative approaches available, aimed at selecting the true number of underlying factors [84,85] while monitoring the cumulative variance explained by a particular choice of K . Additionally, having a choice of K that leads to scientific interpretability in factor loadings can be highly desirable, and this is based upon context-specific subjective judgment. Here we apply Horn's parallel analysis method [86,87] which is the most frequently applied objective method for choosing K [88,89]. This yields $K = 3$, thereby cumulatively explaining 74.5% of the variance in our dataset. Further, this choice yields factors whose loadings align with scientifically intuitive concepts, as reflected in our assigned labels above. While we detail our primary findings with

$K = 3$, we ensure the robustness of our results by conducting a sensitivity analysis where K varies across $\{2, 3, 4, 5, 8, 10\}$. By comparing the estimated factor scores and loadings across different values of K in Supplementary File 1, we verify that the directionality of each raw measurement's treatment effect, as implied by the major loadings and longitudinal trajectories of the factors, is consistent across different choices of K . We provide general recommendations in Section A.10 aimed at ensuring robustness and replicability of findings in the context of clinical trials data.

4.2. Stage 2: longitudinal model

The second stage of our approach takes all the patients' scores for a particular factor k and then compares the average longitudinal trend across different treatment arms, under the model presented in (2). This is illustrated in Fig. 2. To capture the average longitudinal behaviour across a treatment arm, we must take account of the fact that a patient's consecutive observations in a clinical trial are not independent, but rather are correlated over time. We employ a multilevel linear model to share information across all time points within a patient in a way that models the dependence structure of observations across time ([44,49, 58,90]; see Section A.6 for details). The Stage 2 analysis outputs two important types of estimates, *treatment arm means* and *average treatment effects* (ATEs).

At each time point, the treatment arm mean is defined as the expected value of a variable (e.g. an endpoint, a composite endpoint, or a factor score) for the patients in that particular treatment arm (see, e.g., Fig. 3(b)). The treatment arm mean is a clinically useful summary, as it can be interpreted as the average response over time for a randomly selected patient from the corresponding arm. Note that we estimate a treatment arm mean for the Placebo arm. In contrast, the ATE is defined as the difference between the active treatment arm mean and the placebo treatment arm mean (see, e.g., Fig. 3(c)).

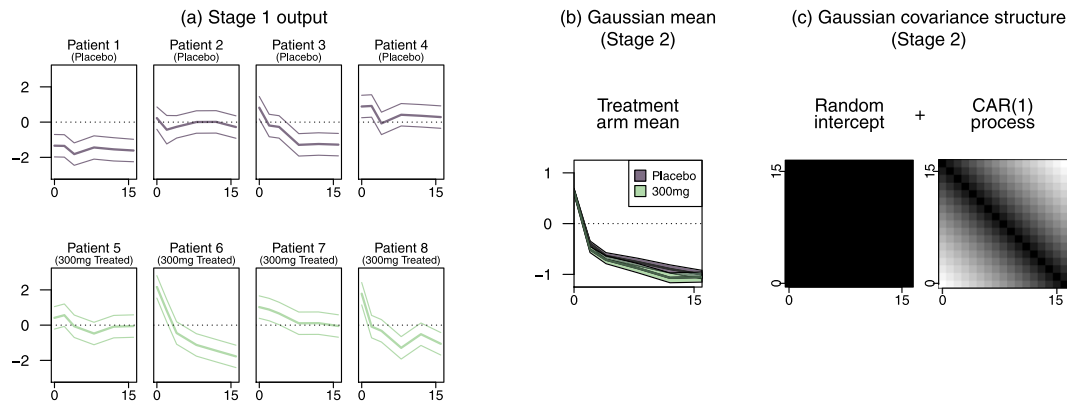


Fig. 2. Illustration of the longitudinal model in Stage 2. (a) Estimated scores $\hat{Z}_{i,k}$ and standard errors, outputted from Stage 1, serve as inputs for Stage 2 (estimates are shown with ± 2 SE error bands). They are shown for eight example patients comprising a small subset of the two treatment arms, “Placebo” and “300mg Treated”. (b) Treatment arm means, $\mu_{k,\text{placebo}}$ and $\mu_{k,\text{active}}$ are estimated in Stage 2 (example estimates are shown with ± 2 SE error bands). (c) Patient i 's latent $\hat{Z}_{i,k}$ varies at random around the corresponding treatment arm mean (see (2)) according to autocorrelated noise vector $\xi_{i,k} \sim \mathcal{N}(\mathbf{0}, \Sigma)$, where Σ is the sum of a random intercept and a CAR(1) process. The gray scale indicates a larger (positive) value the darker the heatmap is. All panels illustrate Factor 1 as, for simplicity, only a single factor is shown in this figure.

5. Case study

5.1. Single-trial analysis

We begin our analysis with a single PsA trial. In Fig. 3 we plot the estimated treatment arm trajectories for PsA study FUTURE-5. Since each of the underlying measurements has been scaled across indications (PsA and RA), studies, and time points to have zero mean and unit standard deviation, the resulting scores in Fig. 3(b) and (c) tend also to be centred around zero and varying on unit scale. The loadings matrix is displayed in Fig. 3(a) for reference, which is identical to the loadings previously discussed in Fig. 1.

The three panels in Fig. 3(b) display the means for each of the three treatment arms – Placebo, 150mg and 300mg – with error bands of ± 2 standard errors representing approximate 95% confidence intervals. We observe that the *treatment arm mean trajectories* are rather smooth and tend to decrease over time. The decreasing mean curves in Factor 3 [Pain, Physical function, QoL] can be explained by a decrease in each of the positively weighted components of the factors (e.g. Tender, ESR, Pain), and increases in the negatively weighted ones (e.g. SF36-PCS) which contribute to it. We further notice non-overlapping confidence intervals in Fig. 3(b) between the Placebo and the active treatment arms, suggestive of a significant ATE.

We can quantify these ATEs directly by estimating the difference between Placebo and each of the active arms in turn. The three panels in Fig. 3(c) show these estimated differences relative to Placebo (ATEs). At baseline, the confidence intervals in Fig. 3(c) overlap with zero for each of the active treatment arms. This is consistent with the randomised study design under which there is no expected difference between treatment arms at baseline. (Confidence intervals for ATEs that did not contain zero would correspond to rejection of the null hypothesis of no treatment effect at a particular time point.)

The shapes of the trajectories are clinically informative. In Fig. 3(c), the ATE for Factor 2 [Inflammatory markers] decreases abruptly between week 0 and week 2 and then remains relatively stable up to week 16. In comparison, Factor 1 [Joints] and Factor 3 [Pain, Physical function, QoL] decrease, i.e. improve, across the whole time period, beginning with a steep and ending with a shallow negative gradient. Here, this may clinically indicate a rapid improvement of inflammatory markers (disease domain corresponding to Factor 2 [Inflammatory markers]), while joint signs and symptoms (Factor 1 [Joints]) and Pain, Physical Function and Quality of Life (Factor 3 [Pain, Physical function, QoL]) are more gradually affected by treatment.

For comparison, Figure A.7 displays model outputs for RA study REASSURE. The loadings matrix for RA in Figure A.7(a) is remarkably

similar to that for the PsA trial in Fig. 3(a), even though these loadings were calculated by applying PPCA to non-overlapping data sets from separate trials. Figure A.8 shows that the similarity extends to all four trials considered. This indicates that the empirical correlation structure amongst measurements is qualitatively similar and stable between these trials and indications. The observation that latent factors are shared by different indications leads us to compare indication-specific ATEs, pooling trials by indication, in Section 5.3. A further point of note between Figs. 3 and A.7 is the relative width of confidence intervals – generally narrower in FUTURE-5 than in REASSURE – consistent with the larger sample size of 996 in FUTURE-5 vs. 637 in REASSURE.

Within our latent longitudinal factor framework, there is the additional option of introducing a degree of longitudinal smoothness on ATEs, which can be estimated or specified. This has the effect of increasing information sharing across time points, thereby further increasing precision on meta-analysed ATEs, a point we return to in the Discussion.

5.2. Meta-analysis of multiple trials

Combining information across several clinical trials is highly attractive in principle, as it allows us to increase the total sample size and estimate treatment effects more precisely. However, an important consideration is that different trials may recruit patients with systematically different characteristics. For example, in PsA trials FUTURE-2 and FUTURE-5, we see that measurements have different average baseline values, and that the placebo arms behave quite differently in the two trials.

In cases where there is evident heterogeneity across trials in the treatment arm means, it still may be desirable and reasonable to estimate other effects which are common across trials with a shared component in our model, for instance the ATEs (i.e. the effect of active treatment relative to the Placebo arm). This can be achieved within our linear modelling framework by adjusting for the structures that are heterogeneous across trials (e.g. different baseline enrolment, handled via subject-specific random intercept in Eq. A.8), while information is pooled on structures that are reasonably homogeneous across trials (e.g., as might be the case with an ATE common across trials, handled via appropriate specification of the fixed effects in Eq. A.7).

An application of this approach to our two RA trials is shown in Fig. 4 (a similar meta-analysis for the two PsA trials is shown in Figure A.9). First, we compare the ATEs when they are estimated separately for the two RA trials. That is, we choose the subset of Y of the subjects for each RA trial in turn, perform Stages 1 and 2 of our method, and plot the estimated ATEs outputted by Stage 2, with Fig. 4(a) and (b)

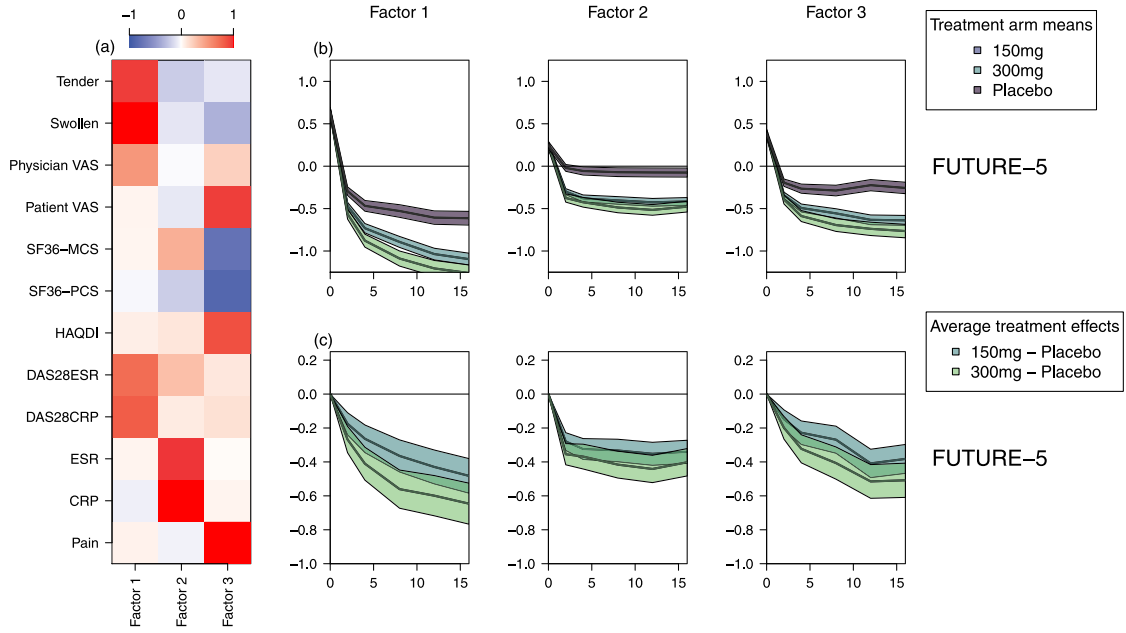


Fig. 3. Analysis of a single PsA trial (FUTURE-5). (a) Loadings matrix calculated based on data from trial FUTURE-5 only. (b) Treatment arm means for each of three factor scores. (c) ATEs (active arms relative to Placebo) for each of three factor scores. In panels (b) and (c) the estimates are outputted by Stage 2, and approximate 95% confidence intervals are shown, based upon ± 2 standard errors.

corresponding to trials REASSURE and NURTURE-1 respectively. The loadings for the two trials are shown in Figure A.8(d–e). For both trials, the confidence intervals are generally overlapping between treatment arms, with no evidence of heterogeneous ATEs between arms. This motivates a meta-analysis model that estimates the ATE jointly between the two trials. In the meta-analysis model, we subset Y to contain both RA trials and proceed as before, performing both Stages 1 and 2 of our method, plotting estimated ATEs in Fig. 4(c). The loadings matrix for the meta analysis of two trials is shown in Figure A.8(f). It is not necessary to make further assumptions of homogeneity across trials, because we adjust for any systematic differences in the baseline characteristics and Placebo arms. We observe that the estimated ATEs in the meta-analysis in Fig. 4(c) are compatible with the ones estimated by trial in panels Fig. 4(a–b), yet exhibit narrower confidence bands due to a higher sample size, as is expected and desired in such a meta-analysis.

Importantly, the meta-analysis model is going beyond the simple averaging of measurements at a time point across multiple trials. In particular: (i) information is shared across the multiple measurements comprising a factor in Stage 1; (ii) information is shared across multiple sequential measurements of each patient while accounting for the correlation structure across time points in the residual in Stage 2; and (iii) adjustments can be flexibly made for heterogeneous response curves across sexes and studies, by including suitable covariates in the linear predictor in Stage 2 (see Section A.6 for more details). By sharing information suitably across multiple trials, endpoints and time points, our analysis framework shows the potential to identify and characterise ATEs more comprehensively and precisely.

5.3. Comparison of treatment effects across indications PsA and RA

Fig. 5(a) compares estimated treatment arm means for the Placebo and 150mg arms for PsA and RA patients. The estimates in Fig. 5 are taken directly from Fig. 4(c) (RA) together with the treatment effects in Figure A.9(c) (PsA) for purposes of comparison across indications. There are clear differences in the baseline average for PsA compared to RA patients. This is to be expected, and reflects the fundamentally different clinical characteristics of PsA and RA patients.

Despite the very different baseline levels between indications, the ATE in the 150mg arm (relative to Placebo) does not appear to differ substantially between indications in Fig. 5(a). We explore this feature in more detail in Fig. 5(b), where we plot the ATE directly. It is remarkable that the confidence bands for the treatment effects for Factor 1 [Joints] and Factor 2 [Inflammatory markers] in Fig. 5(b) are overlapping across PsA and RA indications at all time points, pointing to strong qualitative and quantitative similarities between the average treatment response in PsA and RA. This shows the suitability of our longitudinal framework to represent shared effects (here ATEs shared across indications) while maintaining the flexibility of capturing heterogeneous effects (here baseline measurement means differing across indications).

6. Benchmarking experiments

Here we perform a simulation study investigating how well our proposed model estimates the ATE in randomised clinical trials. We benchmark our method's performance on this task against a variety of other statistical and machine learning tools. Whilst our method outputs multivariate longitudinal ATEs, in the simulation experiments we focus in on performance at estimating a single dimension of this output, namely the ATE for measurement variable 1 at week 16, defined as

$$\text{ATE} := \mathbb{E}[y_{i,1,16}^1 - y_{i,1,16}^0], \quad (3)$$

where superscripts y^1 and y^0 denote the potential outcomes for an individual receiving active treatment and placebo respectively, and $y_{i,1,16}$ is the first element of the multivariate measurement captured on individual i at week 16.

6.1. Simulation model

A range of analytical and methods are benchmarked on simulated data sets. The simulation model is designed to mimic the properties of real clinical trials data. Simulations proceed by first generating latent variables Z according to (2) with average trajectories shown in Figure A.4; these curves are specified within the E_{\max} dose-response

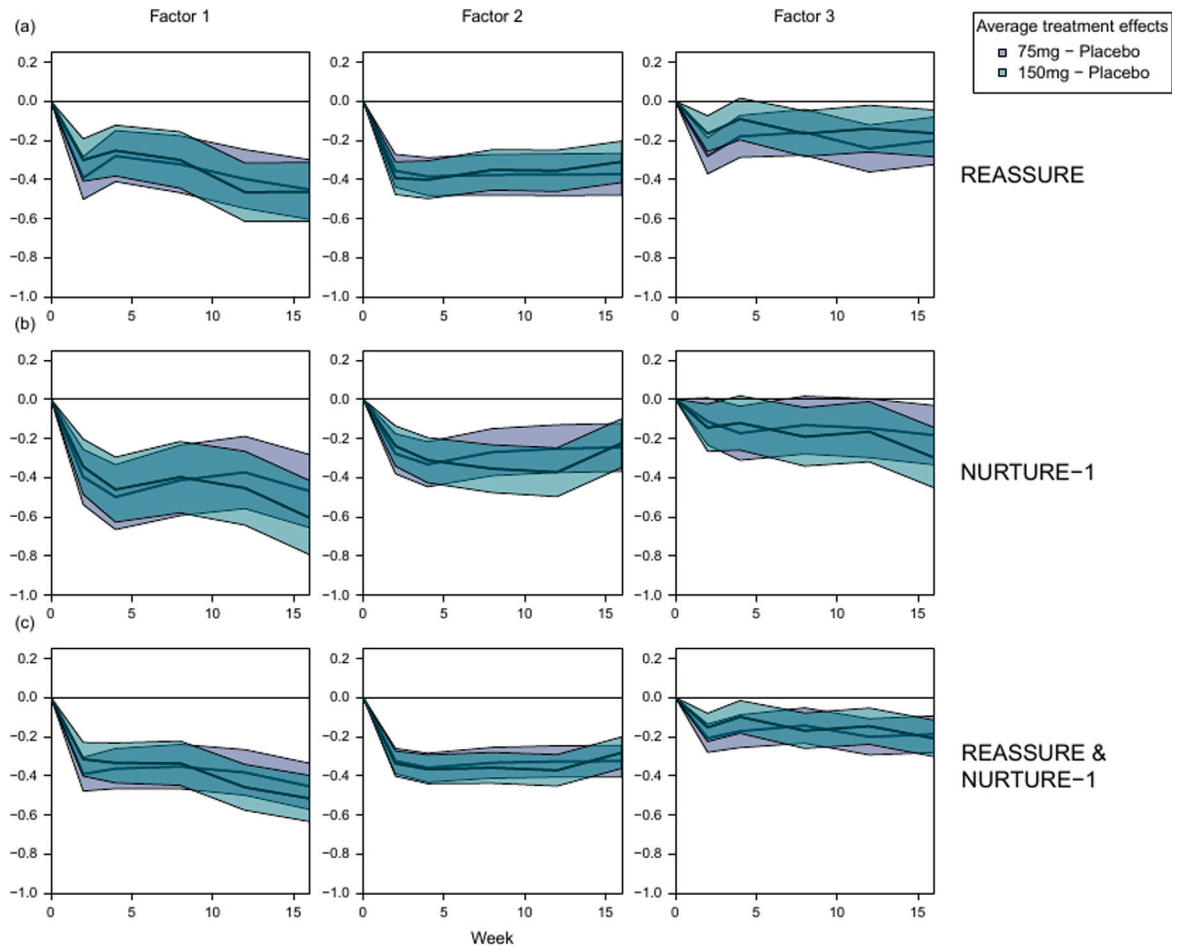


Fig. 4. Meta-analysis of two RA trials (REASSURE and NURTURE-1). (a) ATEs for REASSURE. (b) ATEs for NURTURE-1. (c) ATEs under a meta-analysis model with treatment effects pooled across trials, while adjusting for trial-specific baseline characteristics and Placebo arms. In each panel, error bands refer to ± 2 standard errors representing approximate 95% confidence intervals for each of the three factors.

framework [91] so as to resemble the trajectories estimated in our case study, such as those illustrated in Fig. 2(b). The simulation model generates observed measurements \mathbf{Y} according to (1), conditional on the generated latent trajectories \mathbf{Z} . Full details are presented in Section A.8, and default parameter settings Θ_1 are shown in Table A.3. In our benchmarking experiments we vary parameter settings by perturbing a single parameter away from its default setting at a time (keeping all other parameters fixed), leading to a total of seven distinct parameter settings, $\Theta_{1:7}$, explored across Tables 1, 2, A.5 and A.6 — see sub-table headings for details.

Depending on the particular analytical method being used, the inputted data subset has two separate binary properties: it is either multivariate or univariate; and it is either multiple-time-point (longitudinal) or single-time-point (cross-sectional). This results in four distinct data subsets, which we label $D_{1:4}$ and illustrate in Fig. 6.

6.2. Benchmarked statistical and machine learning methods

Default linear models. For each of the four data subsets $D_{1:4}$ we can fit a linear model that includes longitudinal and/or factor components as appropriate. We use LM_j to denote the default linear model for data subset D_j . The linear models that analyse multivariate measurements (LM_2 and LM_4) incorporate a PPCA factor model in Stage 1, while linear models that analyse multiple time points (LM_3 and LM_4) have longitudinal linear mixed models in Stage 2. Further details of benchmarked methods are provided in Section A.9.

Benchmarked univariate time series models for dataset D_3 . In the default linear model LM_3 we use a CAR(1) process to model the

temporally correlated, stationary residuals. As alternatives, we benchmark a variety of linear models, similar to LM_3 in all regards except with the CAR(1) component replaced with discrete-time autoregressive moving average (ARMA) models on the autocorrelated, stationary residuals [92]. Additionally, we benchmark an exponential smoothing (ES) model as applied to D_3 .

Benchmarked multivariate time series models for dataset D_4 . We investigate the performance of vector autoregressive models (VARs) [93] as well as a variety of modern machine learning methods for longitudinal data: recurrent neural networks (RNNs) [94], gated recurrent unit networks (GRUs) [95] and long short-term memory networks (LSTMs) [96]. LM_4 is the method we introduce in the current paper and applied in the case studies. LM_4 has a separate unconstrained parameter for each treatment arm mean at each time point. We additionally benchmark a constrained parameterisation of this model, denoted $LM_4^{(\text{lin})}$, that assumes linear treatment-arm trajectories.

6.3. Verifying valid inference under default linear models

Here we verify the inferential validity of the four default linear models $LM_{1:4}$. These models allow for statistical inference incorporating formal hypothesis testing and confidence intervals. Furthermore, each comprises a multivariate Gaussian likelihood, correctly specified with respect to the simulation model, and satisfying the required regularity conditions for optimal asymptotic efficiency [97] under maximum likelihood (ML) estimation. Hence these models are ideal candidates for comparing performance across data subsets. Table 1 shows results for

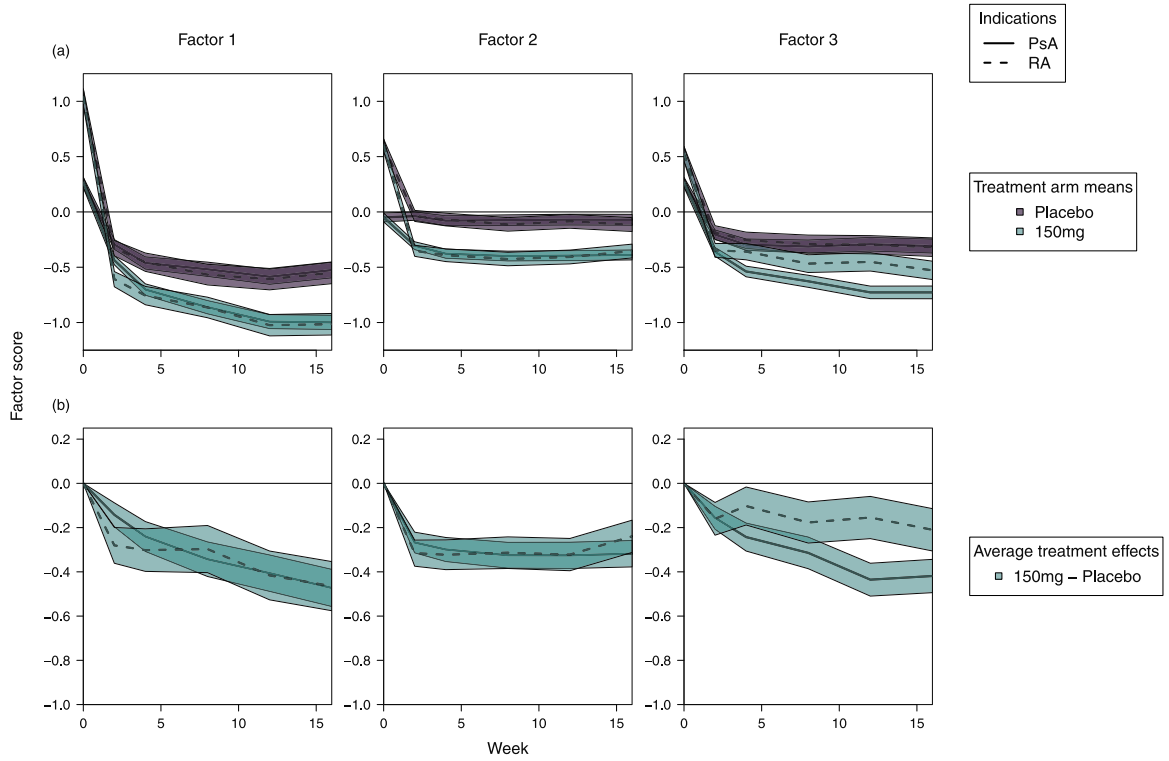


Fig. 5. Comparison of treatment effects across PsA (FUTURE-2 and FUTURE-5) and RA (REASSURE and NURTURE-1). The top panels show the treatment arm means (Placebo and 150mg) compared between indications. In contrast, the bottom panels compare the ATEs (150mg - Placebo) between RA and PsA.

each of the seven distinct parameter settings, $\Theta_{1:7}$, labelled top of each sub-table. The small proportion of blue numbers in Table 1 supports the inferential validity of $LM_{1:4}$, confirming Type I error control at 5% and estimated confidence interval (CI) coverage of 95%, while also presenting no evidence of statistical bias.

As a further check of the validity of our proposed method, we perform an ablation analysis on our case-study datasets to demonstrate that the methodology does not detect treatment effects when none are present. We permuted each dataset, randomly relabelling individuals' treatment arms without replacement. In the resulting comparison of active versus placebo arms, we confirm that our proposed method generates results compatible with zero effect of treatment relative to placebo in the permuted datasets (see Supplementary File 2, panels labelled (c)).

6.4. Benchmarking results

Here we benchmark the wide range of statistical and machine learning methods introduced in Section 6.2, using MSE as the performance measure. Results are presented in Table 2 for parameter settings $\Theta_{1:2}$, with corresponding results for $\Theta_{3:7}$ shown across Tables A.5 and A.6.

Models within our proposed longitudinal factor framework (LM_4 and $LM_4^{(lin)}$) perform competitively, showing the smallest MSE throughout, with the exception of Table A.6(a). The linearly constrained implementation of our framework ($LM_4^{(lin)}$), outperforms the unconstrained parameterisation LM_4 in terms of MSE, while exhibiting significant negative bias. This is a manifestation of the well-known bias-variance trade-off (see, e.g., section 6.4.4 of [98]) where a biased estimator can outperform an unbiased one. In the current context, $LM_4^{(lin)}$ introduces bias via a linear approximation to the underlying curved treatment arm trajectories, but reduces variance compared to LM_4 via increased information sharing across time points, resulting in a net reduction in MSE.

The MSEs of the various ARMA models are generally very similar to the MSE of LM_3 (which has a CAR(1) model on the residuals). As

intuition would suggest, models fitted to the full multivariate longitudinal dataset, D_4 , have more information at their disposal, and exhibit typically smaller estimated MSEs than models fitted to data subsets $D_{1:3}$. For example, under the default parameter settings in Table 1(a) and 2(a), models involving a MV component outperform the other models according to all benchmarking measures, e.g. $MSE_{LM_2} = 0.0089$ and $MSE_{LM_4} = 0.0071$ versus $MSE_{LM_1} = 0.012$ and $MSE_{LM_3} = 0.011$.

Increasing the magnitude of the effect size (Tables 1(b) and 2(b)) does not have much of an impact on MSE, whilst power increases for $LM_{1:4}$; the intuition here is simply that increasing the size of an effect can make it easier to detect without affecting the properties of the estimator (here bias and standard deviation (SD)). Tables 1(c) and A.5(a) explore the effect of stronger temporal autocorrelation (larger ρ) in the latent factors. Compared to Table 1(a) and 2(a), we observe a modest performance boost of the longitudinal models LM_3 and LM_4 for increased ρ , but not for the other models, consistent with the intuition of longitudinal models pooling information across time points. Finally, Tables 1(e-g) and A.6(a-c) indicate that increasing each of the three noise variance parameters individually leads to a decrease in performance of all models.

6.5. Computational efficiency of methods

Details of the computational complexity and run times of the various methods are shown in Table A.4. Compared to other multivariate longitudinal informatics approaches our models are much faster to fit. For example, LM_4 (0.9 s) exhibits an approximate 10-fold speed-up compared to VAR (8.6 s), and an average 77-fold average speed-up relative to RNN (69.5 s), LSTM (70.1 s) and GRU (69.6 s), despite these latter three methods making use of extensive parallel processing on graphics processing units (GPUs).

7. Discussion

Clinical trial data sets contain a range of information comprising many diverse measurements repeatedly conducted at each of multiple

Table 1

Simulation results for validating and benchmarking linear models. Each sub-table presents results from simulations based on a different set of parameters in the data generating model. Sub-table (a) corresponds to default parameter settings Θ_1 , listed in Table A.3, while sub-tables (b–g) show results under settings $\Theta_2, \dots, \Theta_7$, each perturbing a single parameter away from its default value (other parameters held constant) labelled top of each sub-table; e.g. in Θ_2 in (b), $\mu_{\text{active},16}$ is changed from its default setting -1.2 to -1.4 . The first four columns detail which model is fitted, with “MV” and “Longit” denoting whether the model has a factor and/or longitudinal component respectively. Columns “Bias”, “Type I error” and “Coverage” check whether the estimated bias, type I error, and 95% CI coverage with respect to the true ATE, defined in (3), are compatible with 0 (bias), 0.05 (type I error) or 0.95 (coverage). If an estimate deviates by more than $1.96 \times \text{SE}$ from 0 (bias), 0.05 (type I error) or 0.95 (coverage) it is coloured blue (the number of blue entries is $3/84 = 3.6\%$, consistent with the intended 5% false positive rate). The final two columns benchmark the performance of the four models (best performance in **bold**), with “MSE” showing mean squared error and “Power” showing the statistical power to reject the null hypothesis of zero ATE at 5% significance level. For each estimated performance measure, standard errors (SEs) are shown subscripted in parenthesis. Each estimate shown is based upon 1000 simulated data sets.

(a) Θ_1 : Default parameters (see Table A.3)								
Method	Input	MV	Longit	Bias (SE)	Type I error (SE)	Coverage (SE)	Power (SE)	MSE (SE)
LM ₁	D ₁			0.0012 (0.0035)	0.036 (0.0059)	0.96 (0.0059)	0.26 (0.014)	0.012 (0.00054)
LM ₂	D ₂	✓		0.0052 (0.0030)	0.045 (0.0066)	0.95 (0.0068)	0.32 (0.015)	0.0089 (0.00037)
LM ₃	D ₃		✓	−0.00051 (0.0033)	0.051 (0.0070)	0.95 (0.0070)	0.28 (0.014)	0.011 (0.00048)
LM ₄	D ₄	✓	✓	0.0052 (0.0027)	0.045 (0.0066)	0.96 (0.0065)	0.40 (0.015)	0.0071 (0.00030)
(b) Θ_2 : Increased magnitude of (negative) effect size $\mu_{\text{active},16} = -1.4$								
Method	Input	MV	Longit	Bias (SE)	Type I error (SE)	Coverage (SE)	Power (SE)	MSE (SE)
LM ₁	D ₁			0.00028 (0.0036)	0.049 (0.0068)	0.95 (0.0068)	0.73 (0.014)	0.013 (0.00059)
LM ₂	D ₂	✓		0.00061 (0.0030)	0.050 (0.0069)	0.95 (0.0068)	0.87 (0.011)	0.0087 (0.00042)
LM ₃	D ₃		✓	0.0026 (0.0034)	0.047 (0.0067)	0.95 (0.0067)	0.78 (0.013)	0.011 (0.00053)
LM ₄	D ₄	✓	✓	0.0025 (0.0026)	0.046 (0.0066)	0.95 (0.0069)	0.94 (0.0076)	0.0070 (0.00034)
(c) Θ_3 : Increased $\rho = 0.95$								
Method	Input	MV	Longit	Bias (SE)	Type I error (SE)	Coverage (SE)	Power (SE)	MSE (SE)
LM ₁	D ₁			0.0028 (0.0036)	0.047 (0.0067)	0.95 (0.0067)	0.25 (0.014)	0.013 (0.00061)
LM ₂	D ₂	✓		0.0049 (0.0031)	0.056 (0.0073)	0.95 (0.0072)	0.32 (0.015)	0.0094 (0.00042)
LM ₃	D ₃		✓	0.0020 (0.0032)	0.055 (0.0072)	0.94 (0.0072)	0.32 (0.015)	0.010 (0.00050)
LM ₄	D ₄	✓	✓	0.0034 (0.0024)	0.074 (0.0083)	0.92 (0.0084)	0.55 (0.016)	0.0056 (0.00027)
(d) Θ_4 : Decreased $\rho = 0.25$								
Method	Input	MV	Longit	Bias (SE)	Type I error (SE)	Coverage (SE)	Power (SE)	MSE (SE)
LM ₁	D ₁			0.0017 (0.0035)	0.047 (0.0067)	0.95 (0.0067)	0.26 (0.014)	0.012 (0.00053)
LM ₂	D ₂	✓		0.0020 (0.0029)	0.041 (0.0063)	0.96 (0.0063)	0.34 (0.015)	0.0084 (0.00039)
LM ₃	D ₃		✓	0.0043 (0.0034)	0.048 (0.0068)	0.95 (0.0068)	0.27 (0.014)	0.011 (0.00048)
LM ₄	D ₄	✓	✓	0.0052 (0.0027)	0.048 (0.0068)	0.95 (0.0069)	0.38 (0.015)	0.0071 (0.00031)
(e) Θ_5 : Increased $\sigma_{\text{AR}} = 1$								
Method	Input	MV	Longit	Bias (SE)	Type I error (SE)	Coverage (SE)	Power (SE)	MSE (SE)
LM ₁	D ₁			−0.0017 (0.0050)	0.045 (0.0066)	0.95 (0.0066)	0.17 (0.012)	0.025 (0.0011)
LM ₂	D ₂	✓		0.0016 (0.0046)	0.057 (0.0073)	0.94 (0.0073)	0.17 (0.012)	0.021 (0.00089)
LM ₃	D ₃		✓	−0.0025 (0.0049)	0.050 (0.0069)	0.95 (0.0069)	0.18 (0.012)	0.024 (0.0011)
LM ₄	D ₄	✓	✓	0.00079 (0.0045)	0.055 (0.0072)	0.94 (0.0072)	0.19 (0.012)	0.020 (0.00086)
(f) Θ_6 : Increased $\sigma_{\text{RI}} = 1$								
Method	Input	MV	Longit	Bias (SE)	Type I error (SE)	Coverage (SE)	Power (SE)	MSE (SE)
LM ₁	D ₁			0.0068 (0.0048)	0.050 (0.0069)	0.95 (0.0069)	0.13 (0.011)	0.023 (0.0011)
LM ₂	D ₂	✓		0.0049 (0.0044)	0.049 (0.0068)	0.95 (0.0067)	0.16 (0.012)	0.020 (0.00087)
LM ₃	D ₃		✓	0.0037 (0.0037)	0.055 (0.0072)	0.94 (0.0072)	0.23 (0.013)	0.013 (0.00060)
LM ₄	D ₄	✓	✓	0.0040 (0.0029)	0.055 (0.0072)	0.94 (0.0073)	0.34 (0.015)	0.0085 (0.00040)
(g) Θ_7 : Increased $\sigma_Y = 1$								
Method	Input	MV	Longit	Bias (SE)	Type I error (SE)	Coverage (SE)	Power (SE)	MSE (SE)
LM ₁	D ₁			−0.0022 (0.0052)	0.045 (0.0066)	0.95 (0.0066)	0.14 (0.011)	0.027 (0.0014)
LM ₂	D ₂	✓		0.0031 (0.0037)	0.048 (0.0068)	0.95 (0.0070)	0.24 (0.013)	0.013 (0.00065)
LM ₃	D ₃		✓	−0.0035 (0.0051)	0.048 (0.0068)	0.95 (0.0068)	0.14 (0.011)	0.026 (0.0013)
LM ₄	D ₄	✓	✓	0.0024 (0.0035)	0.055 (0.0072)	0.94 (0.0077)	0.27 (0.014)	0.012 (0.00060)

time points. The most common approach to dimension reduction is to summarise key measurements via a composite endpoint, such as ACR20, which is carefully crafted according to clinical considerations and domain knowledge. Here we have adopted a complementary approach, capturing and representing the full measurement space in an entirely data-driven yet interpretable way. Our sparse implementation of PPCA reduces the whole set of measurements down to a small and interpretable set of factors (referred to here as disease domains) that explain 74.5% of the variation in the data. Comparison of treatment effects in data-driven disease domains with existing composite endpoints can motivate the design of new composite endpoints and/or the

refinement of existing ones, so as to capture treatment response more effectively.

There are several benefits of examining treatment effects in a low-dimensional space that captures most of the variation in the data. First, it is easier in practice to deal with a small number of derived variables than to focus on the entire set of measurements one by one. In our case study, this equates to considering a three-dimensional response space – Factor 1 [*Joints*], Factor 2 [*Inflammatory markers*] and Factor 3 [*Pain, Physical function, QoL*] – instead of a 12-dimensional one. Second, because the dimensionality reduction is data driven, we can be reassured that our analyses capture the key variation in the

Table 2

Benchmarking results. Each sub-table presents results from simulations based on a different set of parameters in the data generating model. Sub-table (a) corresponds to default parameter settings θ_1 , listed in Table A.3, while setting θ_2 in sub-table (b) perturbs a single parameter away from its default value (other parameters held constant). Further benchmarking results, corresponding to parameter settings $\theta_3, \dots, \theta_7$ in Table 1, are shown in Tables A.5 and A.6. The first four columns detail which model is fitted, with “MV” and “Longit” denoting whether the model has a multivariate and/or longitudinal component respectively. The final three columns show the bias, standard deviation (SD) and mean squared error (MSE) of the estimator for the true underlying ATE as defined in (3). The smallest MSE is displayed in **bold**. Significant bias is designated by blue text, showing when an estimate deviates by more than $1.96 \times \text{SE}$ from 0. Biased methods can lead to smaller MSE if the corresponding variance reduction is sufficient (see text for discussion of the bias–variance trade-off). For each estimated performance measure, standard errors (SEs) are shown subscripted in parenthesis. Each estimate shown is based upon 1000 simulated data sets.

(a) θ_1 : Default parameters (see Table A.3)						
Method	Input	MV	Longit	Bias (SE)	SD (SE)	MSE (SE)
LM ₁	D ₁			0.0012 (0.0035)	0.11 (0.00054)	0.012 (0.00054)
LM ₂	D ₂	✓		0.0052 (0.0030)	0.094 (0.00037)	0.0089 (0.00037)
LM ₃	D ₃		✓	−0.00051 (0.0033)	0.11 (0.00048)	0.011 (0.00048)
ARMA(1,0)	D ₃		✓	−0.00051 (0.0033)	0.11 (0.00048)	0.011 (0.00048)
ARMA(0,1)	D ₃		✓	−0.00059 (0.0033)	0.11 (0.00048)	0.011 (0.00048)
ARMA(1,1)	D ₃		✓	−0.00019 (0.0033)	0.11 (0.00048)	0.011 (0.00048)
ARMA(2,0)	D ₃		✓	−0.00034 (0.0033)	0.11 (0.00048)	0.011 (0.00048)
ARMA(0,2)	D ₃		✓	−0.00052 (0.0033)	0.11 (0.00048)	0.011 (0.00048)
ES	D ₃		✓	−0.0037 (0.0041)	0.13 (0.00069)	0.017 (0.00069)
LM ₄	D ₄	✓	✓	0.0052 (0.0027)	0.084 (0.00030)	0.0071 (0.00030)
LM ₄ ^(lin)	D ₄	✓	✓	−0.0073 (0.0024)	0.076 (0.00025)	0.0059 (0.00025)
VAR	D ₄	✓	✓	0.022 (0.0028)	0.087 (0.00036)	0.0081 (0.00036)
RNN	D ₄	✓	✓	0.0016 (0.0036)	0.11 (0.00056)	0.013 (0.00056)
LSTM	D ₄	✓	✓	0.00064 (0.0036)	0.11 (0.00055)	0.013 (0.00055)
GRU	D ₄	✓	✓	0.00034 (0.0035)	0.11 (0.00054)	0.013 (0.00054)
(b) θ_2 : Increased magnitude of (negative) effect size $\mu_{\text{active},16} = -1.4$						
Method	Input	MV	Longit	Bias (SE)	SD (SE)	MSE (SE)
LM ₁	D ₁			0.00028 (0.0036)	0.11 (0.00059)	0.013 (0.00059)
LM ₂	D ₂	✓		0.00061 (0.0030)	0.093 (0.00042)	0.0087 (0.00042)
LM ₃	D ₃		✓	0.0026 (0.0034)	0.11 (0.00053)	0.011 (0.00053)
ARMA(1,0)	D ₃		✓	0.0026 (0.0034)	0.11 (0.00053)	0.011 (0.00053)
ARMA(0,1)	D ₃		✓	0.0027 (0.0034)	0.11 (0.00053)	0.011 (0.00053)
ARMA(1,1)	D ₃		✓	0.0025 (0.0034)	0.11 (0.00054)	0.011 (0.00054)
ARMA(2,0)	D ₃		✓	0.0025 (0.0034)	0.11 (0.00053)	0.011 (0.00053)
ARMA(0,2)	D ₃		✓	0.0026 (0.0034)	0.11 (0.00053)	0.011 (0.00053)
ES	D ₃		✓	0.0065 (0.0041)	0.13 (0.00075)	0.017 (0.00075)
LM ₄	D ₄	✓	✓	0.0025 (0.0026)	0.083 (0.00034)	0.0070 (0.00034)
LM ₄ ^(lin)	D ₄	✓	✓	−0.015 (0.0025)	0.078 (0.00030)	0.0062 (0.00030)
VAR	D ₄	✓	✓	0.042 (0.0027)	0.086 (0.00042)	0.0091 (0.00042)
RNN	D ₄	✓	✓	0.010 (0.0035)	0.11 (0.00057)	0.012 (0.00057)
LSTM	D ₄	✓	✓	0.010 (0.0034)	0.11 (0.00056)	0.012 (0.00056)
GRU	D ₄	✓	✓	0.011 (0.0034)	0.11 (0.00054)	0.012 (0.00054)

data. Third, by pooling information across multiple related endpoints, factors provide more precise estimates of treatment effects, essentially by aggregating shared biologically relevant signals while averaging out independent, irrelevant measurement noise.

When analysing data that have been collected longitudinally it is vital to account for the fact that an individual’s measurements tend to be correlated over time. Incorrectly assuming independence across time points tends to result in artificially tight error bounds on ATEs. We incorporate this correlation structure via a CAR(1) process available in the nlme software package [50]. It may additionally be reasonable, in some contexts, to impose a degree of smoothness in ATEs over time. The greater the degree of smoothness, the greater the amount of information which is pooled across time points, in turn leading to greater precision on ATE estimates.

We have developed an efficient two-stage approach to model fitting. We harness two existing software packages, each of which has

been optimised to perform a particular task: pcaMethods contains an efficient implementation of PPCA, and nlme a highly optimised and flexible implementation of linear models with structured residual covariance. The relationship between packages is represented in a universal machine language (UML) class diagram in Figure A.2 [99], and the interaction between user and software throughout the analysis pipeline is visualised in a UML use case diagram in Figure A.3. There are several advantages to this two-stage approach, which come under the mantle of a property we refer to as *composability* [100]. First, by performing inference in two stages, we simplify the analysis greatly, reducing total run times and memory requirements (Table A.4). Second, we are able to take advantage of software packages that are established, robust, and efficient, obviating the need for us to develop and optimise our own software for these core building blocks. Third, it is relatively straightforward to exchange the Stage 1 or Stage 2 software for other packages (e.g. mgcv for Stage 2), in the case that we wish to fit qualitatively different models. Finally, as we discuss in Section A.7, our two-stage optimisation is methodologically justifiable as a version of empirical Bayes inference under the full hierarchical model that relates raw measurement data to ATEs via latent scores.

In conclusion, we have presented a methodological framework for longitudinal multivariate analysis of clinical trial data. Our analyses pool information coherently across measurements, patients and time points (and potentially multiple trials and indications) to obtain precise and interpretable estimates of ATEs. We encountered several interesting statistical challenges, arising from the highly structured nature of clinical trials data. For example, data typically present with correlation across measurements, correlation across time points within patients, informative baseline levels, and missing data. Performing useful and well calibrated inference in this context requires careful specification of flexible and interpretable statistical models, alongside computational methods capable of fitting such models effectively and efficiently.

Funding statement

This paper is the output from the Novartis funded alliance with Oxford Big Data Institute, UK. Novartis funded the design of the study and collection, analysis and interpretation of data, and in writing the manuscript. Academic advisors from Oxford Big Data Institute (BDI) and Novartis personnel designed the project. The Oxford BDI – Cosen-tyx Collaboration group conducted the data analyses, and all authors had access to the data and, to the best of their knowledge, confirm the accuracy and completeness of the analyses. G.N. acknowledges funding from the NIHR Biomedical Research Centre, Oxford, UK (grant no. NIHR203311).

Data consent statement

The four datasets (REASSURE, NURTURE-1, FUTURE-2, and FUTURE-5) used in this paper were collected from completed, anonymised clinical trials. Since the datasets were anonymised, no further EC/IRB (Ethics Committees/Institutional Review Board) approvals are required.

CRedit authorship contribution statement

Fabian Falck: Investigation, Formal analysis, Data curation, Conceptualization, Methodology, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Xuan Zhu:** Data curation, Investigation, Resources, Validation, Visualization, Writing – original draft, Writing – review & editing. **Sahra Ghalebikesabi:** Writing – original draft, Investigation. **Matthias Kormaksson:** Conceptualization, Data curation, Investigation, Methodology, Project administration, Software, Supervision, Writing – review & editing. **Marc Vandemeulebroecke:** Conceptualization, Methodology, Project administration, Supervision, Writing – original draft,

- D_1 : Univariate measurement, single time point
- D_2 : Multivariate measurement, single time point
- D_3 : Univariate measurement, multiple time points
- D_4 : Multivariate measurement, multiple time points

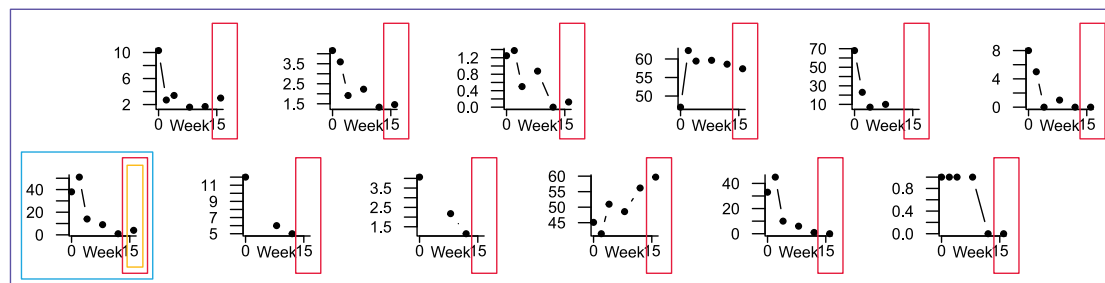


Fig. 6. Data subsets corresponding to the four models benchmarked in our simulation experiments. Each sub-plot corresponds to a longitudinal measurement, all from a single, arbitrarily chosen patient. The orange coloured model (D_1) inputs only the week 16 time point from measurement 1, the red model (D_2) inputs week 16 data from all 12 measurements, the light blue model (D_3) inputs all time points from measurement 1, and the purple model (D_4) inputs all time points from all 12 measurements. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Writing – review & editing. **Cong Zhang**: Data curation, Investigation. **Ruvie Martin**: Project administration. **Stephen Gardiner**: Data curation, Writing – review & editing. **Chun Hei Kwok**: Data curation. **Dominique M. West**: Data curation. **Luis Santos**: Data curation. **Chengeng Tian**: Project administration. **Yu Pang**: Project administration. **Aimee Readie**: Project administration. **Gregory Ligozio**: Project administration. **Kunal K. Gandhi**: Project administration. **Thomas E. Nichols**: Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing. **Ann-Marie Mallon**: Data curation, Funding acquisition, Project administration. **Luke Kelly**: Supervision, Project administration. **David Ohlssen**: Supervision, Project administration. **George Nicholson**: Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Fabian Falck, Sahra Ghalebikesabi, Stephen Gardiner, Luis Santos, Thomas E. Nichols, Ann-Marie Mallon, Luke Kelly and George Nicholson report that financial support was provided by Novartis Pharmaceuticals Corporation, East Hanover, United States. Marc Vandemeulebroecke reports a relationship with Novartis Pharma AG, Basel, Switzerland that included, during the period of this research, employment and equity or stocks. Cong Zhang, Chengeng Tian and Yu Pang report a relationship with China Novartis Institutes for Bio-medical Research CO., Shanghai, China that included, during the period of this research, employment and equity or stocks. Xuan Zhu, Matthias Kormaksson, Ruvie Martin, Aimee Readie, Gregory Ligozio, Kunal K. Gandhi and David Ohlssen report a relationship with Novartis Pharmaceuticals Corporation, East Hanover, United States that included, during the period of this research, employment and equity or stocks.

Data availability

The data of all four clinical trials analysed in this work can be requested in anonymised form through a voluntary data-sharing process on <https://clinicalstudydatarequest.com/>. We refer to Section 3.2 for further details.

Acknowledgments

We thank Thomas Dumortier and Hanno Richards for valuable comments and their pharmacokinetic-pharmacodynamic and clinical insight.

Software availability

The R source code implementing our framework and reproducing our simulation experiments is provided open-source at <https://github.com/georgenicholson/latent-factors>.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.jbi.2024.104641>.

References

- [1] U S Department of Health and Human Services, Multiple endpoints in clinical trials guidance for industry, 2017, Available from: <https://www.fda.gov/files/drugs/published/Multiple-Endpoints-in-Clinical-Trials-Guidance-for-Industry.pdf>.
- [2] S.J. Pocock, N.L. Geller, A.A. Tsiatis, The analysis of multiple endpoints in clinical trials, *Biometrics* 43 (3) (1987) 487–498, Available from: <http://www.jstor.org/stable/2531989>.
- [3] X. An, Q. Yang, P.M. Bentler, A latent factor linear mixed model for high-dimensional longitudinal data analysis, *Stat. Med.* 32 (24) (2013) 4229–4239.
- [4] D. Hedeker, R.D. Gibbons, *Longitudinal Data Analysis*, vol. 451, John Wiley & Sons, 2006.
- [5] P.J. Diggle, P. Heagerty, K.Y. Liang, S.L. Zeger, *Analysis of Longitudinal Data*, second ed., Oxford University Press, Great Clarendon Street, Oxford OX2 6DP, 2002.
- [6] J.D. Singer, J.B. Willett, J.B. Willett, et al., *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*, Oxford University Press, 2003.
- [7] G.M. Fitzmaurice, C. Ravichandran, A primer in longitudinal data analysis, *Circulation* 118 (19) (2008) 2005–2010.
- [8] G. Box, G. Jenkins, G. Reinsel, *Time Series Analysis, Forecasting and Control*, Prentice Hall, Englewood Cliffs. NJ, 1994.
- [9] M. Vandemeulebroecke, B. Bornkamp, T. Krahne, J. Mielke, A. Monsch, P. Quarg, A longitudinal item response theory model to characterize cognition over time in elderly subjects, *CPT: Pharm. Syst. Pharmacol.* 6 (9) (2017) 635–641.
- [10] G. Rasch, *Probabilistic Models for Some Intelligence and Attainment Tests*, ERIC, 1993.
- [11] R.D. Hays, L.S. Morales, S.P. Reise, Item response theory and health outcomes measurement in the 21st century, *Med. Care* 38 (9 Suppl) (2000) I128.
- [12] S.E. Embretson, S.P. Reise, *Item Response Theory*, Psychology Press, 2013.
- [13] A. Barbieri, J. Peyhardi, T. Conroy, S. Gourgu, C. Laverne, C. Mollevi, Item response models for the longitudinal analysis of health-related quality of life in cancer clinical trials, *BMC Med. Res. Methodol.* 17 (1) (2017) 1–13.
- [14] R. Darrell Bock, M. Lieberman, Fitting a response model for dichotomously scored items, *Psychometrika* 35 (2) (1970) 179–197.
- [15] B. Muthén, Contributions to factor analysis of dichotomous variables, *Psychometrika* 43 (4) (1978) 551–560.
- [16] I. Moustaki, M. Knott, Generalized latent trait models, *Psychometrika* 65 (3) (2000) 391–411.

- [17] J.J. McArdle, D. Epstein, Latent growth curves within developmental structural equation models, *Child Dev.* (1987) 110–133.
- [18] K.J. Preacher, A.L. Wichman, R.C. MacCallum, N.E. Briggs, Latent Growth Curve Modeling, vol. 157, Sage, 2008.
- [19] T.E. Duncan, S.C. Duncan, L.A. Strycker, An Introduction to Latent Variable Growth Curve Modeling: Concepts, Issues, and Applications, Routledge, 2013.
- [20] C.M. Carvalho, J. Chang, J.E. Lucas, J.R. Nevins, Q. Wang, M. West, High-dimensional sparse factor modeling: applications in gene expression genomics, *J. Am. Stat. Assoc.* 103 (484) (2008) 1438–1456.
- [21] P. Broët, S. Richardson, F. Radvanyi, Bayesian hierarchical model for identifying changes in gene expression from microarray experiments, *J. Comput. Biol.* 9 (4) (2002) 671–683.
- [22] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, et al., Predicting the clinical status of human breast cancer by using gene expression profiles, *Proc. Natl. Acad. Sci.* 98 (20) (2001) 11462–11467.
- [23] K.E. Lee, N. Sha, E.R. Dougherty, M. Vannucci, B.K. Mallick, Gene selection: a Bayesian variable selection approach, *Bioinformatics* 19 (1) (2003) 90–97.
- [24] K. Bae, B.K. Mallick, Gene selection using a two-level hierarchical Bayesian model, *Bioinformatics* 20 (18) (2004) 3423–3430.
- [25] J. Joo, S.A. Williamson, A.I. Vazquez, J.R. Fernandez, M.S. Bray, Advanced dietary patterns analysis using sparse latent factor models in young adults, *J. Nutr.* 148 (12) (2018) 1984–1992.
- [26] Q. Liu, B. Cheng, Y. Jin, P. Hu, Bayesian tensor factorization-drive breast cancer subtyping by integrating multi-omics data, *J. Biomed. Inform.* 125 (2022) 103958.
- [27] J. Chu, Z. Sun, W. Dong, J. Shi, Z. Huang, On learning disentangled representations for individual treatment effect estimation, *J. Biomed. Inform.* 124 (2021) 103940.
- [28] A.S. Bryk, S.W. Raudenbush, Application of hierarchical linear models to assessing change, *Psychol. Bull.* 101 (1) (1987) 147.
- [29] D.R. Rogosa, J.B. Willett, Understanding correlates of change by modeling individual differences in growth, *Psychometrika* 50 (2) (1985) 203–228.
- [30] W.N. Venables, B.D. Ripley, Random and mixed effects, in: *Modern Applied Statistics with S*, Springer, 2002, pp. 271–300.
- [31] M. Davidian, D.M. Giltinan, Nonlinear models for repeated measurement data: an overview and update, *J. Agric. Biol. Environ. Stat.* 8 (4) (2003) 387–419.
- [32] D. Todem, K. Kim, E. Lesaffre, Latent-variable models for longitudinal data with bivariate ordinal outcomes, *Stat. Med.* 26 (5) (2007) 1034–1054.
- [33] C.M. Laffont, M. Vandemeulebroecke, D. Concordet, Multivariate analysis of longitudinal ordinal data with mixed effects models, with application to clinical outcomes in osteoarthritis, *J. Amer. Statist. Assoc.* 109 (507) (2014) 955–966.
- [34] S. Bianconcini, K.A. Bollen, The latent variable-autoregressive latent trajectory model: A general framework for longitudinal data analysis, *Struct. Equation Model.: Multidiscip. J.* 25 (5) (2018) 791–808.
- [35] M.E. Tipping, C.M. Bishop, Probabilistic principal component analysis, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 61 (1999) 611–622, Available from: <https://www.jstor.org/stable/2680726>.
- [36] K. Pearson, LIII. on lines and planes of closest fit to systems of points in space, *Lond. Edinb. Dublin Philos. Mag. J. Sci.* 2 (11) (1901) 559–572.
- [37] A. Sportisse, C. Boyer, J. Josses, Estimation and imputation in probabilistic principal component analysis with missing not at random data, *Adv. Neural Inf. Process. Syst.* (2020) 33.
- [38] L.E. Peterson, Partitioning large-sample microarray-based gene expression profiles using principal components analysis, *Comput. Methods Programs Biomed.* 70 (2) (2003) 107–119.
- [39] S. Oba, Sato. Ma, I. Takemasa, M. Monden, Matsubara. Ki, S. Ishii, A Bayesian missing value estimation method for gene expression profile data, *Bioinformatics* 19 (16) (2003) 2088–2096.
- [40] D. Menaga, S. Revathi, Probabilistic principal component analysis (PPCA) based dimensionality reduction and deep learning for cancer classification, in: *Intelligent Computing and Applications: Proceedings of ICICA 2019*, Springer, 2021, pp. 353–368.
- [41] M. Alavi, D.C. Visentin, D.K. Thapa, G.E. Hunt, R. Watson, M. Cleary, Exploratory factor analysis and principal component analysis in clinical studies: Which one should you use, *J. Adv. Nurs.* 76 (8) (2020) 1886–1889.
- [42] Santos. RdO, B.M. Gorgulho, Castro. MAd, R.M. Fisberg, D.M. Marchioni, V.T. Baltar, Principal component analysis and factor analysis: Differences and similarities in nutritional epidemiology application, *Rev. Bras. Epidemiol.* (2019) 22.
- [43] A. Bédard, J. Garcia-Aymerich, M. Sanchez, N.Le. Moual, F. Clavel-Chapelon, M.C. Boutron-Ruault, et al., Confirmatory factor analysis compared with principal component analysis to derive dietary patterns: a longitudinal study in adult women, *J. Nutr.* 145 (7) (2015) 1559–1568.
- [44] N.M. Laird, J.H. Ware, Random-effects models for longitudinal data, *Biometrics* (1982) 963–974.
- [45] M.J. Lindstrom, D.M. Bates, Nonlinear mixed effects models for repeated measures data, *Biometrics* (1990) 673–687.
- [46] M.J. Lindstrom, D.M. Bates, Newton–Raphson and EM algorithms for linear mixed-effects models for repeated-measures data, *J. Am. Stat. Assoc.* 83 (404) (1988) 1014–1022.
- [47] E. Vonesh, V.M. Chinchilli, Linear and Nonlinear Models for the Analysis of Repeated Measurements, CRC Press, 1996.
- [48] S.R. Searle, G. Casella, C.E. McCulloch, Variance Components, John Wiley & Sons, 2009.
- [49] J. Pinheiro, D. Bates, Mixed-Effects Models in S and S-PLUS, Springer science & business media, 2006.
- [50] J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, R Core Team, Nlme: Linear and nonlinear mixed effects models, 2022.
- [51] C. Proust-Lima, V. Philipps, B. Liquet, Estimation of extended mixed models using latent classes and latent processes: the R package lcmm, 2015, arXiv preprint [arXiv:150300890](https://arxiv.org/abs/150300890).
- [52] S.L. Zeger, M.R. Karim, Generalized linear models with random effects; a Gibbs sampling approach, *J. Am. Stat. Assoc.* 86 (413) (1991) 79–86.
- [53] D.M. Bates, D.G. Watts, Nonlinear regression analysis and its applications, 519.536, B3; 1988.
- [54] D.B. Dunson, Bayesian latent variable models for clustered mixed outcomes, *J. R. Stat. Soc.: Ser. B (Stat. Methodol.)* 62 (2) (2000) 355–366.
- [55] T.J. Hastie, Generalized additive models, in: *Statistical Models in S*, Routledge, 2017, pp. 249–307.
- [56] S.N. Wood, Stable and efficient multiple smoothing parameter estimation for generalized additive models, *J. Am. Stat. Assoc.* 99 (467) (2004) 673–686.
- [57] S.N. Wood, Mgc: GAMs and generalized ridge regression for R, *R news* 1 (2) (2001) 20–25.
- [58] S. Wood, Generalized Additive Models: An Introduction with R, Chapman and Hall/CRC, 2017, Available from: <https://www.taylorfrancis.com/books/generlized-additive-models-simon-wood/10.1201/9781315370279>.
- [59] G. Marra, S.N. Wood, Coverage properties of confidence intervals for generalized additive model components, *Scand. J. Stat.* 39 (1) (2012) 53–74.
- [60] J.A. Aston, J.M. Chiou, J.P. Evans, Linguistic pitch analysis using functional principal component mixed effect models, *J. R. Stat. Soc.: Ser. C (Appl. Stat.)* 59 (2) (2010) 297–317.
- [61] A. Pickles, T. Croudace, Latent mixture models for multivariate and longitudinal outcomes, *Stat. Methods Med. Res.* 19 (3) (2010) 271–289.
- [62] L.C. Liu, D. Hedeker, A mixed-effects regression model for longitudinal multivariate ordinal data, *Biometrics* 62 (1) (2006) 261–268.
- [63] S. Fieuws, G. Verbeke, Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles, *Biometrics* 62 (2) (2006) 424–431.
- [64] I.B. McInnes, G. Schett, The pathogenesis of rheumatoid arthritis, *N. Engl. J. Med.* 365 (23) (2011) 2205–2219, Available from: <http://dx.doi.org/10.1056/NEJMr1004965>.
- [65] H. Tahir, A. Deodhar, M. Genovese, T. Takeuchi, J. Aelion, F. Van den Bosch, et al., Secukinumab in active rheumatoid arthritis after Anti-TNFalpha therapy: A randomized, double-blind placebo-controlled phase 3 study, *Rheumatol. Ther.* 4 (2) (2017) 475–488.
- [66] J.S. Smolen, D. Aletaha, I.B. McInnes, Rheumatoid arthritis, *Lancet (London, England)* 388 (10055) (2016) 2023–2038.
- [67] H.A. Blair, Secukinumab: A review in psoriatic arthritis, *Drugs* (2021) 1–12.
- [68] S. Hackett, L. Coates, et al., Psoriatic arthritis: an up to date overview, *Indian J. Rheumatol.* 15 (5) (2020) 45.
- [69] A. Ogdie, L.C. Coates, D.D. Gladman, Treatment guidelines in psoriatic arthritis, *Rheumatology* 59 (1) (2020) i37–i46.
- [70] A. Toussi, N. Maverakis, S.T. Le, S. Sarkar, S.K. Raychaudhuri, S.P. Raychaudhuri, Updated therapies for the management of psoriatic arthritis, *Clin. Immunol.* (2020) 108536.
- [71] L. Garcia-Montoya, H. Marzo-Ortega, The role of secukinumab in the treatment of psoriatic arthritis and ankylosing spondylitis, *Ther. Adv. Musculoskelet. Dis.* 10 (9) (2018) 169–180, Available from: <http://dx.doi.org/10.1177/1759720X18787766>.
- [72] A.B. Gottlieb, P.J. Mease, B. Kirkham, P. Nash, Balsa.B.A.C. Combe, J. Rech, et al., Secukinumab efficacy in psoriatic arthritis: Machine learning and meta-analysis of four phase 3 trials, *J. Clin. Rheumatol.* (2020).
- [73] D.T. Felson, J.J. Anderson, M. Boers, et al., The American college of rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. The committee on outcome measures in rheumatoid arthritis clinical trials, *Arthritis Rheum.* 36 (6) (1993) 729–740.
- [74] G.L. Hickey, P. Philipson, A. Jorgensen, R. Kolamunnage-Dona, Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues, *BMC Med. Res. Methodol.* 16 (1) (2016).
- [75] Novartis Pharmaceuticals, A Phase III, Randomized, Double-Blind, Placebo Controlled Multi-Center Study of Subcutaneous Secukinumab (150 Mg and 300 Mg) in Prefilled Syringe To Demonstrate Efficacy (Including Inhibition of Structural Damage), Safety, and Tolerability Up To 2 Years in Subjects with Active Psoriatic Arthritis (FUTURE 5), [clinicaltrials.gov](https://clinicaltrials.gov/study/NCT02404350), 2020, NCT02404350, Available from: <https://clinicaltrials.gov/study/NCT02404350>, Submitted: March 16 2015.
- [76] Novartis Pharmaceuticals, A Phase III Randomized, Double-Blind, Placebo-Controlled Multicenter Study of Subcutaneous Secukinumab in Prefilled Syringes To Demonstrate the Efficacy At 24 Weeks and To Assess the Long Term Efficacy, Safety and Tolerability Up To 5 Years in Patients with Active Psoriatic Arthritis, [clinicaltrials.gov](https://clinicaltrials.gov/study/NCT01752634), 2020, NCT01752634, Available from: <https://clinicaltrials.gov/study/NCT01752634> Submitted: October 23 2012.

- [77] IB. McInnes, PJ. Mease, AJ. Kivitz, P. Nash, P. Rahman, J. Rech, et al., Long-term efficacy and safety of secukinumab in patients with psoriatic arthritis: 5-year (end-of-study) results from the phase 3 FUTURE 2 study, *Lancet Rheumatol.* 2 (4) (2020) e227–e235, Available from: <https://www.sciencedirect.com/science/article/pii/S2665991320300369>.
- [78] PJ. Mease, R. Landewé, P. Rahman, H. Tahir, A. Singhal, E. Boettcher, et al., Secukinumab provides sustained improvement in signs and symptoms and low radiographic progression in patients with psoriatic arthritis: 2-year (end-of-study) results from the FUTURE 5 study, *RMD Open* 7 (2) (2021).
- [79] Novartis Pharmaceuticals, A Randomized, Double-Blind, Placebo-Controlled Study of Secukinumab To Demonstrate the Efficacy At 24 Weeks and To Assess the Safety, Tolerability and Long Term Efficacy Up To 2 Years in Patients with Active Rheumatoid Arthritis Who Have an Inadequate Response To Anti-TNF α Agents (CAIN457F2302) and a Three Year Extension Study To Evaluate the Long Term Efficacy, Safety and Tolerability of Secukinumab in Patients with Active Rheumatoid Arthritis (CAIN457F2302E1), *clinicaltrials.gov*, 2017, NCT01377012, Available from: <https://clinicaltrials.gov/study/NCT01377012> Submitted: June 17 2011.
- [80] Novartis Pharmaceuticals, A Randomized Double-Blind Placebo- and Active-Controlled Study of Secukinumab To Demonstrate the Efficacy At 24 Weeks and To Assess the Safety, Tolerability and Long Term Efficacy Up To 1 Year in Patients with Active Rheumatoid Arthritis Who Have an Inadequate Response To Anti-TNF- α Agents (CAIN457F2309) and a Four Year Extension Study To Evaluate the Long Term Efficacy, Safety and Tolerability of Secukinumab in Patients with Active Rheumatoid Arthritis (CAIN457F2309E1), *clinicaltrials.gov*, 2016, NCT01350804, Available from: <https://clinicaltrials.gov/study/NCT01350804> Novartis Pharmaceuticals.
- [81] FJ. Blanco, R. Möricke, E. Dokoupilova, C. Codding, J. Neal, M. Andersson, et al., Secukinumab in active rheumatoid arthritis: A phase III randomized, double-blind, active comparator- and placebo-controlled study, *Arthritis Rheumatol.* (Hoboken, NJ) 69 (6) (2017) 1144–1153.
- [82] AM. Mallon, DA. Häring, F. Dahlke, P. Aarden, S. Afyouni, D. Delbarre, et al., Advancing data science in drug development through an innovative computational framework for data sharing and statistical analysis, *BMC Med. Res. Methodol.* 21 (1) (2021) 1–11.
- [83] C. Bishop, *Pattern Recognition and Machine Learning* | Christopher Bishop | Springer, first ed., Springer-Verlag New York, 2006, Available from: <https://www.springer.com/gp/book/9780387310732>.
- [84] WR. Zwick, WF. Velicer, Comparison of five rules for determining the number of components to retain, *Psychol. Bull.* 99 (3) (1986) 432–442, Place: US Publisher: American Psychological Association.
- [85] JC. Hayton, DG. Allen, V. Scarpello, Factor retention decisions in exploratory factor analysis: a tutorial on parallel analysis, *Organ. Res. Methods* 7 (2) (2004) 191–205, Available from: <http://dx.doi.org/10.1177/1094428104263675>, Publisher: SAGE Publications Inc.
- [86] JL. Horn, A Rationale and test for the number of factors in factor analysis, *Psychometrika.* 30 (1965) 179–185.
- [87] LW. Glorfeld, An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain, *Educ. Psychol. Meas.* 55 (3) (1995) 377–393, Place: US Publisher: Sage Publications.
- [88] A. Dinno, Exploring the sensitivity of horn's parallel analysis to the distributional form of random data, *Multivar. Behav. Res.* 44 (3) (2009) 362–388, Publisher: Routledge _eprint: <https://doi.org/10.1080/00273170902938969> Available from: <http://dx.doi.org/10.1080/00273170902938969>.
- [89] A. Dinno, Implementing Horn's parallel analysis for principal component analysis and factor analysis, *Stata J.* 9 (2) (2009) 291–298, Publisher: SAGE Publications. Available from: <http://dx.doi.org/10.1177/1536867X0900900207>.
- [90] SN. Wood, Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models, *J. R. Stat. Soc.: Ser. B (Stat. Methodol.)* 73 (1) (2011) 3–36.
- [91] J. Macdougall, Analysis of dose–response studies—Emax model, in: N. Ting (Ed.), *Dose Finding in Drug Development*, in: *Statistics for Biology and Health*, Springer, New York, NY, 2006, pp. 127–145, Available from: http://dx.doi.org/10.1007/0-387-33706-7_9.
- [92] JC. Pinheiro, D. Bates, *Mixed-Effects Models in S and S-PLUS*, Springer Science & Business Media, 2009, Google-Books-ID: y54QDUTmvDcC.
- [93] ZF. Fisher, Y. Kim, BL. Fredrickson, V. Pipiras, Penalized estimation and forecasting of multiple subject intensive longitudinal data, *Psychometrika* 87 (2) (2022) 1–29, Available from: <http://dx.doi.org/10.1007/s11336-021-09825-7>.
- [94] KP. Murphy, *Probabilistic Machine Learning: An Introduction*, MIT Press, 2022, Available from: probml.ai.
- [95] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, et al., Learning phrase representations using RNN encoder–decoder for statistical machine translation, 2014, arXiv preprint [arXiv:1406.1078](https://arxiv.org/abs/1406.1078).
- [96] S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Comput.* 9 (8) (1997) 1735–1780.
- [97] G. Kulldorff, On the conditions for consistency and asymptotic efficiency of maximum likelihood estimates, *Scand. Actuar. J.* 1957 (3–4) (1957) 129–144, Publisher: Taylor & Francis _eprint: <https://doi.org/10.1080/03461238.1957.10405966> Available from: <http://dx.doi.org/10.1080/03461238.1957.10405966>.
- [98] KP. Murphy, *Machine Learning: A Probabilistic Perspective*, in: *Adaptive computation and machine learning series*, MIT Press, Cambridge, MA, 2012.
- [99] N. Medvidovic, DS. Rosenblum, DF. Redmiles, JE. Robbins, Modeling software architectures in the Unified Modeling Language, *ACM Trans. Softw. Eng. Methodol.* 11 (1) (2002) 2–57, Available from: <https://dl.acm.org/doi/10.1145/504087.504088>.
- [100] G. Nicholson, M. Blangiardo, M. Briers, PJ. Diggle, TE. Fjelde, H. Ge, et al., Interoperability of statistical models in pandemic preparedness: principles and reality, *Stat. Sci.* 37 (2) (2022) 183–206.