Statistica Sinica Preprint No: SS-2023-0049

Title	Longitudinal Modeling of Rank-based Global Outcome
Manuscript ID	SS-2023-0049
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0049
Complete List of Authors	Maomao Ding, Jing Ning, Xuming He, Anne-Marie Wills and Ruosha Li
Corresponding Authors	Ruosha Li
E-mails	ruosha.li@uth.tmc.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0049

Longitudinal Modeling of Rank-based Global Outcome

Maomao Ding[1], Jing Ning[2*], Xuming He[3], Anne-Marie Wills[4], Ruosha Li[5*]

[1] *Rice University,* [2] *University of Texas MD Anderson Cancer Center,* [3] *Washington University in St. Louis,* [4] *Massachusetts General Hospital and Harvard Medical School, and* [5] *University of Texas Health Science Center at Houston*

Abstract: Many chronic diseases exhibit multifaceted symptoms that cannot be comprehensively characterized by one outcome. To address this, researchers often adopt a global outcome to combine information from multiple individual outcomes. The global rank-sum facilitates robust integration of multiple outcomes and has been applied in many clinical studies. We consider longitudinal settings and devise a global percentile outcome for depicting patients' time-varying global disease burden. We develop useful regression strategies for the longitudinal global percentile outcome based on a flexible regression framework of the monotonic index model. Posing minimal restrictions, we propose a maximum rank correlation type estimator and show that it entails desirable asymptotic properties. The methods are also extended to accommodate the common missing at random dropout scenarios. We propose a computationally stable and efficient procedure for parameter estimation, as well as a perturbation scheme for consistent variance estimation. Numerical studies show that our method performs well under realistic settings. We apply the proposed method to data from a Parkinson's disease clinical trial to examine risk factors associated with elevated global disease burden and accelerated disease progression. *Key words and phrases:* Global percentile outcome; Longitudinal data; Maximum rank correlation; Monotonic index model; Parkinson's disease.

1. Introduction

Definition of the primary outcome plays a central role in biomedical studies. For many diseases with complex symptoms and multi-dimensional deteriorations, however, there is no single outcome that is comprehensive enough to capture all important aspects of the disease. In these situations, researchers often construct a global outcome that combines information from multiple individual outcomes. For example, the NIH Exploratory Trials in Parkinson's Disease (NET-PD) Long-term Study-1 (LS-1) was a multicenter, double-blind, placebo-controlled, randomized trial aimed at testing whether the daily administration of creatine would slow the global progression of Parkinson's disease (PD) (Kieburtz et al., 2015). PD involves a broad spectrum of symptoms, including movement-related (motor) symptoms such as tremor and stiffness, as well as non-motor symptoms such as impaired sense of smell, sleep disorders, and fatigue. To comprehensively assess the global disease progression, the steering committee of the LS-1

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0049

chose five outcome measures for the primary analysis, accounting for both motor and non-motor symptoms (Elm and Investigators, 2012). A global rank-sum (O'Brien, 1984) was formulated based on the change from baseline to year five in the five outcome measures. Similar rank-based global outcomes have also been adopted in studies of heart failure (Felker and Maisel, 2010; Sun et al., 2012) and amyotrophic lateral sclerosis (Berry et al., 2013).

For each study subject, the global rank-sum is computed by first calculating the subject's rank among the study participants in terms of each individual outcome and then summing up the outcome-specific ranks (O'Brien, 1984). The theoretical properties of the global rank-sum test were discussed in Huang et al. (2005). They also considered the general nonparametric Behrens-Fisher hypothesis problem. The global rank-sum test was extended to censored outcomes by Ramchandani et al. (2016), and extension to clustered multiple endpoints were considered in Zhang et al. (2019).

However, existing methods for the global rank-sum are mostly targeting two-sample or K-sample hypothesis testing problems and could not account for additional covariates, such as confounders or risk factors. What is more, these methods do not accommodate longitudinal data, a data structure that commonly arises in biomedical studies. In the NET-PD LS-1 trial and many similar studies, the outcome measures were collected during pre-planned follow-up visits, providing important opportunities for exploring the time trend of global disease progression and the associated mechanisms. It is desirable to link the covariates to the global disease burden and progression through a sensible regression framework.

In this paper, we consider regression modeling of a global, rank-based outcome for longitudinal data. By compositing the percentile ranks of each outcome, we devise a global percentile outcome (GPO) that amalgamate information from the individual longitudinal outcomes. The GPO provides a robust and scale-free way to quantify patients' global disease levels relative to the study population. It ranges between 0 and 1 and facilitates straightforward interpretation. We adopt a longitudinal monotonic index model that requires minimal distributional assumptions and provides great flexibility in accommodating a broad range of relationships between the GPO and covariates. The maximum rank correlation (MRC) estimator under the monotonic index model has been well studied for cross-sectional data (Han 1987; Cavanagh and Sherman 1998; Sherman 1993; among others). We designed rank-based objective functions, which can be readily implemented in standard statistical software. We show that the proposed methods enjoy desirable theoretical properties despite the non-smooth objective functions and the inter-subject dependence due to the estimated GPO.

The rest of the paper is organized as follows. In Section 2, we describe the proposed regression models and estimation strategies. We start with the case where a single outcome measure is considered and then proceed to the GPO, which becomes a univariate outcome after the combination. Next, we extend the methods to accommodate dropout, a common complication in longitudinal data. Asymptotic properties are rigorously established for the proposed estimators. We provide sensible variance estimation procedures and discuss strategies for numerical implementations. In Section 3, we evaluate our estimators through extensive numerical experiments. We apply the proposed method to the NET-PD LS-1 data to examine risk factors for increased disease burden and accelerated disease progression in Section 4. Some concluding remarks are provided in Section 5.

2. Methods

2.1 Monotonic index regression for single longitudinal outcome

We start from the situation with a single longitudinal outcome. For subject i, let $\mathbf{Y}_i = (Y_{i0}, \ldots, Y_{iM})'$ denote a continuous longitudinal outcome measured at times $t_0, \ldots, t_M$, with $t_0 = 0$ corresponding to the baseline and $\boldsymbol{a}'$ denoting the transpose of $\boldsymbol{a}$. The covariate vector $\mathbf{X}_i = (X_{i1}, ..., X_{ip})'$ is of length p. The observed data consist of i.i.d. samples $\{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^n$, where

n is the sample size. To characterize the relationship between $\mathbf{Y}_i$ and $\mathbf{X}_i$, we consider the following monotonic index model:

$$Y_{im} = \zeta\{\mu(\mathbf{X}_i, t_m, \boldsymbol{\beta}), \varepsilon_{im}\},\ m = 0, \ldots, M, \tag{2.1}$$

where $\zeta(a, b)$ is an unspecified bivariate function that is increasing in both arguments. $\mu(\mathbf{X}_i, t_m, \boldsymbol{\beta})$ is a prespecified function with an unknown parameter vector $\boldsymbol{\beta}$, for which the true value is given by $\boldsymbol{\beta}_0$. It can be parameterized in a similar manner as the mean part of a GEE model, except that no intercept term is needed. For example, one may let $\mu(\mathbf{X}_i, t_m, \boldsymbol{\beta}) = \sum_{l=1}^{p} X_{il}(\beta_{0l} + \beta_{ml}t_m) + \beta_t t_m$ or $\mu(\mathbf{X}_i, t_m, \boldsymbol{\beta}) = \sum_{l=1}^{p} X_{il}\{\beta_{0l} + \beta_{ml}I(m > 0)\} + \beta_{tm}$. The error term ε_{im} follows an unspecified distribution F_ε. The components of $(\varepsilon_{i0}, \ldots, \varepsilon_{iM})$ could be correlated within the same subject but are independent across different subjects. To ensure identifiability, it is necessary to impose a scaling constraint on $\boldsymbol{\beta}$, as the model has minimal assumptions on the link function. Two common ways to impose this constraint in literature are by fixing one of the elements to be 1 (Sherman, 1993) or restricting that $||\boldsymbol{\beta}|| = 1$ (Han, 1987). Without loss of generality, we follow the first approach in the sequel. We have verified through simulations (unreported) that the proposed methods also perform satisfactorily

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0049

when we adopt the unit norm constraint. In practice, one could adopt one of the two constraints according to the clinical interest of the application.

Let $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{\theta}) = (1, \boldsymbol{\theta}')'$ and $\boldsymbol{\beta}_0 = \boldsymbol{\beta}(\boldsymbol{\theta}_0) = (1, \boldsymbol{\theta}_0')'$. To estimate the unknown coefficients, we extend the maximum rank correlation (MRC) estimator for cross-sectional data and define the objective function as

$$\mathcal{L}_n(\boldsymbol{\theta}) = \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} \sum_{u=0}^{M} \sum_{v=0}^{M} I(Y_{iu} > Y_{jv}) I\big[\mu\{\mathbf{X}_i, t_u, \boldsymbol{\beta}(\boldsymbol{\theta})\} > \mu\{\mathbf{X}_j, t_v, \boldsymbol{\beta}(\boldsymbol{\theta})\}\big]. \tag{2.2}$$

$\mathcal{L}_n(\boldsymbol{\theta})$ exploits the degree of concordance between the outcome value and the model-based $\mu\{\mathbf{X}, t, \boldsymbol{\beta}(\boldsymbol{\theta})\}$ across the follow-up times, and we let $\widehat{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}} \mathcal{L}_n(\boldsymbol{\theta})$. The $\mathcal{L}_n(\boldsymbol{\theta})$ is non-differentiable with respect to $\boldsymbol{\theta}$ but can be solved using a nonlinear optimization algorithm such as the Nelder-Mead algorithm. More details about computation are deferred to Section 2.5.

For any $\boldsymbol{\theta} \in \Theta$ and $f(\boldsymbol{\theta})$, a function of $\boldsymbol{\theta}$, write ∇_m for the following m-th partial derivative operator with respect to $\boldsymbol{\theta}$, and let $|\nabla_m| f(\boldsymbol{\theta}) = \sum_{i_1,\ldots,i_m} \mid \frac{\partial^m}{\partial\theta_{i_1}\ldots\partial\theta_{i_m}} f(\boldsymbol{\theta}) \mid$. For $\mathbf{z} = (\mathbf{x}', \mathbf{y}')'$ and $\boldsymbol{\theta} \in \Theta$, define

$$\begin{aligned}\tau(\mathbf{z}, \boldsymbol{\theta}) = &\sum_{u=0}^{M} \sum_{v=0}^{M} \mathbb{E}\Big(I(y_u > Y_{jv}) I\big[\mu\{\mathbf{x}, t_u, \boldsymbol{\beta}(\boldsymbol{\theta})\} > \mu\{\mathbf{X}_j, t_v, \boldsymbol{\beta}(\boldsymbol{\theta})\}\big]\Big) \\ &+ \sum_{u=0}^{M} \sum_{v=0}^{M} \mathbb{E}\Big(I(y_u < Y_{jv}) I\big[\mu\{\mathbf{x}, t_u, \boldsymbol{\beta}(\boldsymbol{\theta})\} < \mu\{\mathbf{X}_j, t_v, \boldsymbol{\beta}(\boldsymbol{\theta})\}\big]\Big).\end{aligned}$$

Similar to Sherman (1993), our MRC estimator for longitudinal outcomes is consistent and asymptotically normal, which we summarize in the following theorem. Detailed regularity conditions and proof can be found in the Supplementary Material 1.1.

Theorem 1. *Under model* (2.1) *and conditions (C1)-(C4) in the Supplementary Material, we have* $\widehat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}_0$*, and* $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ *converges in distribution to* $\mathcal{N}(0, V^{-1}\Delta V^{-1})$*, where* $V = \mathbb{E}\nabla_2\tau(\cdot, \boldsymbol{\theta}_0)/2$ *and* $\Delta = \mathbb{E}\nabla_1\tau(\cdot, \boldsymbol{\theta}_0)\nabla_1\tau(\cdot, \boldsymbol{\theta}_0)'$*.*

Alternatively, we can approximate the second indicator function in (2.2) by kernel smoothing. For the MRC with cross-sectional data, Lin and Peng (2013) and Zhang et al. (2018) have used smoothing approaches to overcome the computational challenges posed by the non-smooth objective function. Here, we consider the kernel-smoothed MRC estimator, which maximizes

$$\mathcal{K}_n(\boldsymbol{\theta}; \phi) = \frac{1}{n(n-1)} \sum_{1\le i\ne j\le n} \sum_{u=0}^{M} \sum_{v=0}^{M} I(Y_{iu} > Y_{jv})\phi\left[\frac{\mu\{\mathbf{X}_i, t_u, \boldsymbol{\beta}(\boldsymbol{\theta})\} - \mu\{\mathbf{X}_j, t_v, \boldsymbol{\beta}(\boldsymbol{\theta})\}}{c_n}\right]. \tag{2.3}$$

$\phi(\cdot)$ is a continuous kernel function that is monotonically increasing and satisfies $\lim_{x\to-\infty}\phi(x) = 0$ and $\lim_{x\to\infty}\phi(x) = 1$. We could use a Gaussian kernel by setting $\phi(x)$ as the cumulative distribution function (CDF) of a standard normal distribution, logistic kernel $\phi(x) = (1 + e^{-2x})^{-1}$, or other functions satisfying the above-mentioned properties. The choice of ϕ showed minimal impact in our preliminary simulations, thus we adopted the logistic kernel below. The bandwidth parameter c_n vanishes as $n \to \infty$, and we discuss the choices of c_n in Section 2.5.

2.2 Monotonic index models for the global percentile outcomes

We now adapt our method to handle the global percentile outcome (GPO) that combines information from K individual outcomes, where K is a finite integer. The observed outcomes are Y_{ikm}, where $i = 1, 2, \ldots, n$ is the subject index, $k = 1, \ldots, K$ is the outcome index, and $m = 0, 1, \ldots, M$ is the visit index. Without loss of generality, we assume that higher values are worse for all outcomes. Denote the marginal CDF of the k-th outcome at visit m as $F_{km}(\cdot)$. We formulate a longitudinal GPO for subject i as:

$$\mathcal{P}_{im} = \frac{1}{K}\left\{F_{1m}(Y_{i1m}) + F_{2m}(Y_{i2m}) + \cdots + F_{Km}(Y_{iKm})\right\}, m = 0, 1, ..., M. \tag{2.4}$$

The GPO effectively quantifies the global disease level of the i-th subject at each visit through compositing the subject's percentile rank in each individual outcome. As a percentile-based outcome, it naturally avoids the issue of skewed distributions and outliers. It also features straightforward interpretation as a composite percentile with a range of $[0, 1]$. Specifically,

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0049

higher $\mathcal{P}_{im}$ reflects larger global disease burden, and worsening $\mathcal{P}_{im}$ over m suggests that subject i progresses faster than the general trend. We shall focus on the $\mathcal{P}_{im}$ in (2.2) below, though it is possible to formulate a weighted version of GPO by assigning outcome-specific weights, if certain outcomes are deemed more important than others. When $K > 1$, the GPO is not uniformly distributed and also non-Gaussian. Considering its formulation, the common assumption of linear covariate effect may also be dubious, and it is desirable to avoid stringent assumptions in the modeling procedure.

When dealing with a single outcome, the model in (2.1) is rather general and encompasses many commonly used models. In the presence of multiple longitudinal outcomes, the GPO represents an integrated metric that effectively unifies information from the multiple outcomes. As seen in (2.4), the GPO becomes a single continuous outcome post-combination, making it a natural decision for the application of the generalized monotonic index regression model studied in Section 2.1. We apply this monotonic index model directly to the GPO, without making any specific assumptions on the individual outcomes. Notably, we refrain from imposing specific assumptions on the individual longitudinal outcomes, which, in fact, serves as an advantageous aspect of our proposed method. This global modeling strategy is similar to the win ratio regression (Mao and Wang, 2021), where

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0049

model assumptions are placed directly on a global rank-based composite, bypassing model specification for the individual components. In the following, we consider modeling the GPO using the monotonic index regression,

$$\mathcal{P}_{im} = \zeta\left\{\mu(\mathbf{X}_i, t_m, \boldsymbol{\beta}), \varepsilon_{im}\right\}. \tag{2.5}$$

With the unspecified link $\zeta(a,b)$ and error distribution, it relies on minimal assumptions and accommodates a wide range of the covariate-outcome relationship, including but not limited to a linear transformation model.

Directly modeling the GPO is suitable for settings similar to the motivating example, where the main interest is on the global disease burden, rather than detailed disease aspects captured by individual outcomes. Indeed, when some of the individual outcomes exhibit skewness and non-normality, ensuring accurate model specification for each individual outcome becomes challenging. In this case, it is preferable to directly employ a global regression model through the GPO, rather than constructing K separate regression models for the individual outcomes. Under this global model, the relationship between the covariates and each Y_{ikm} is left unspecified and not required to follow Equation (2.5).

The model readily identifies covariates that affect the global disease burden and/or progression and facilitate straightforward interpretation. The GPO itself is directly interpretable as the global disease level in the study population, with a value of 0 corresponding to the best and 1 corresponding to the worst disease burden. Regarding covariate effects, consider an illustrative example when $\mu(\mathbf{X}_i, t_m, \boldsymbol{\beta})$ is specified as $\sum_{l=1}^{p} X_{il}(\beta_{0l} + \beta_{ml}t_m) + \beta_t t_m$. A positive (or negative) coefficient β_{0l} suggests that a higher value in this covariate is associated with a heavier (or lower) global disease burden. Moreover, a positive (or negative) coefficient of the covariate by time interaction β_{ml} suggests that a higher value in this covariate is indicative of accelerated (or slower) disease progression relative to the general trend of progression in the study population. This facilitates the identification of specific patient subgroups that tend to progress faster over time in global disease level. For monotonic index models, it is standard practice to fix one of the components to 1 for identifiability (Sherman, 1993). One may set this component as a well-acknowledged risk factor for the disease, such as total UPDRS or age in our motivating example. Following this, the magnitude of other coefficients can be easily interpreted as the effect relative to the reference risk factor. Thus, both the sign and magnitude of the fitted coefficients are easily interpretable in real applications.

One important distinction between the GPO and a standard outcome is that the true GPO, $\mathcal{P}_{im}$, is not directly observable but can be estimated by $\widehat{\mathcal{P}}_{im} = K^{-1}\sum_{k=1}^{K}\widehat{F}_{km}(Y_{ikm})$, where $\widehat{F}_{km}(\cdot)$ is the empirical CDF of Y_{ikm}. We can then estimate the regression coefficients by maximizing

$$\mathcal{L}_n^{GPO}(\boldsymbol{\theta}) = \frac{1}{n(n-1)}\sum_{1\le i\neq j\le n}\sum_{m=0}^{M}\sum_{v=0}^{M} I(\widehat{\mathcal{P}}_{im} > \widehat{\mathcal{P}}_{jv}) I\big[\mu\{\mathbf{X}_i, t_m, \boldsymbol{\beta}(\boldsymbol{\theta})\} > \mu\{\mathbf{X}_j, t_v, \boldsymbol{\beta}(\boldsymbol{\theta})\}\big], \tag{2.6}$$

and we denote the maximizer as $\widehat{\boldsymbol{\theta}}_{\mathcal{P}}$. A kernel-smoothed version can also be defined accordingly. Similar to the scenario with single outcome, $\widehat{\boldsymbol{\theta}}_{\mathcal{P}}$ is shown to be consistent and asymptotically normal.

Theorem 2. *Under model* (2.5) *and conditions (C5)-(C11) in the Supplementary Material, we have* $\widehat{\boldsymbol{\theta}}_{\mathcal{P}} \xrightarrow{P} \boldsymbol{\theta}_{0,\mathcal{P}}$ *and* $\sqrt{n}(\widehat{\boldsymbol{\theta}}_{\mathcal{P}} - \boldsymbol{\theta}_{0,\mathcal{P}})$ *converges in distribution to* $\mathcal{N}(0, V_{\mathcal{P}}^{-1}\Delta_{\mathcal{P}}V_{\mathcal{P}}^{-1})$.

The detailed regularity conditions, formal proof, and the definitions of $V_{\mathcal{P}}$ and $\Delta_{\mathcal{P}}$ can be found in the Supplementary Material 1.2. Derivations of these theoretical properties need to account for the additional variability introduced by replacing $\mathcal{P}_{im}$ with $\widehat{\mathcal{P}}_{im}$. While the consistency of $\widehat{\boldsymbol{\theta}}_{\mathcal{P}}$ follows from the uniform convergence of $\widehat{\mathcal{P}}_{im}$ to $\mathcal{P}_{im}$, proof of asymptotic normality is challenged by the non-differentiable term $I(\widehat{\mathcal{P}}_{iu} > \widehat{\mathcal{P}}_{jv})$ in the objective function, which makes it difficult to sort out the additional variability due to the estimated $\widehat{\mathcal{P}}_{iu}$. In the proof, we approximate $I(\widehat{\mathcal{P}}_{iu} > \widehat{\mathcal{P}}_{jv})$ by kernel smoothing using $K_{a_n}(\widehat{\mathcal{P}}_{iu} - \widehat{\mathcal{P}}_{jv})$, where $K(\cdot)$ is a continuously differentiable monotonically increasing kernel function satisfying that $K(x) \geq 0$, $K(x) + K(-x) = 1$, $\lim_{x\to\infty} K(\cdot) = 1$, and $K_{a_n}(x) = K(x/a_n)$. We further approximate $K_{a_n}(\widehat{\mathcal{P}}_{iu} - \widehat{\mathcal{P}}_{jv})$ by its first order approximation, $K'_{a_n}(\mathcal{P}_{iu} - \mathcal{P}_{jv})(\widehat{\mathcal{P}}_{iu} - \widehat{\mathcal{P}}_{jv} - \mathcal{P}_{iu} + \mathcal{P}_{jv}) + K_{a_n}(\mathcal{P}_{iu} - \mathcal{P}_{jv})$. It can be shown that the difference between the original objective function and the objective function obtained by the linearization of the indicator $I(\widehat{\mathcal{P}}_{iu} > \widehat{\mathcal{P}}_{jv})$ is negligible, in that the resulting estimators have difference of order $o_P(n^{-1/2})$. For the linearized objective function, the conditions required in Theorem 4 of Sherman (1993) are satisfied, which further implies the asymptotic normality of $\widehat{\boldsymbol{\theta}}_{\mathcal{P}}$. The form of the asymptotic covariance $\Delta_{\mathcal{P}}$ in the Supplementary Material indicates that when replacing $\mathcal{P}_{im}$ with $\widehat{\mathcal{P}}_{im}$, only the pairs with $\mathcal{P}_{iu} - \mathcal{P}_{jv}$ close to zero contribute to the additional variation. This is well expected, as for those pairs with $|\mathcal{P}_{iu} - \mathcal{P}_{jv}|$ distant from zero, $I(\widehat{\mathcal{P}}_{iu} > \widehat{\mathcal{P}}_{jv})$ coincides with $I(\mathcal{P}_{iu} > \mathcal{P}_{jv})$ with a high probability.

The proposed method can also accommodate situations where some, but not all, outcomes are of discrete nature. Assuming, without loss of generality, that the k_{th} outcome is discrete, $F_{km}(x) = \left\{\frac{P(Y_{km} \leq x) + P(Y_{km} < x)}{2}\right\}$ can be used in place of the standard CDF in the definition of the true $\mathcal{P}_{im}$.

Similarly, in the estimation procedure, we use $\widehat{F}_{km}(x) = \frac{1}{2n}\sum_{j=1}^{n}\{I(Y_{km} \leq x) + I(Y_{km} < x) + 1\}$ in place of the empirical CDF in the derivation of $\widehat{\mathcal{P}}_{im}$. The remaining estimation and inference steps remain unchanged.

2.3 Accounting for missing data due to patient dropout

We next discuss how to handle commonly encountered dropout mechanisms, namely missing completely at random (MCAR) and missing at random (MAR). Let $\eta_{im} \in \{0,1\}$ indicate whether subject i completed the m-th visit. Handling dropout under MCAR is straightforward, where we compute the objective function using only the completed visits as

$$\widetilde{\mathcal{L}}_n(\boldsymbol{\beta}) = \frac{1}{n(n-1)} \sum_{1\leq i\neq j\leq n} \sum_{u=0}^{M}\sum_{v=0}^{M} \eta_{im}\eta_{jv} I(Y_{iu} > Y_{jv}) I\{\mu(\mathbf{X}_i, t_u, \boldsymbol{\beta}) > \mu(\mathbf{X}_j, t_v, \boldsymbol{\beta})\},$$

for the scenario with single outcome. This objective can also be used for the GPO, where we compute $\widehat{\mathcal{P}}_{im}$ by estimating $F_{km}(\cdot)$ using only the subjects who completed visit m. For the more general dropout mechanism of MAR, where the missing mechanism can depend on covariates as well as observed outcomes, we adopt the methods of inverse probability weighting (IPW) (Molenberghs et al., 2014; Yi and He, 2009) to account for dropout. Denote $\mathbf{Y}_{im} = (Y_{i1m}, \ldots, Y_{iKm})$. We estimate $\lambda_{im} = \mathbb{P}(\eta_{im} = 1 \mid \eta_{i,m-1} =$

$1, \mathbf{X}_i, \mathbf{Y}_{i0}, \ldots, \mathbf{Y}_{i,m-1})$ and $w_{im} = \prod_{v=1}^{m} \lambda_{iv}$ with $\lambda_{i0} \equiv 1$ under a suitable parametric or semi-parametric model, such as a logistic regression model. Denote $\boldsymbol{\alpha}$ as the vector of parameters used in modeling missingness, and $\widehat{\boldsymbol{\alpha}}$ its estimated counterpart. We propose a weighted objective function:

$$\begin{aligned}\widetilde{\mathcal{L}}_n(\boldsymbol{\theta}, \widehat{\boldsymbol{\alpha}}) &= \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} \sum_{u=0}^{M} \sum_{v=0}^{M} \frac{\eta_{im}\eta_{jv}}{w_{im}(\widehat{\boldsymbol{\alpha}}) w_{jv}(\widehat{\boldsymbol{\alpha}})} \times \\ &\qquad I(Y_{iu} > Y_{jv}) I\left[\mu\{\mathbf{X}_i, t_u, \boldsymbol{\beta}(\boldsymbol{\theta})\} > \mu\{\mathbf{X}_j, t_v, \boldsymbol{\beta}(\boldsymbol{\theta})\}\right]\end{aligned}$$

for the scenario of single longitudinal outcome, where $w_{im}(\widehat{\boldsymbol{\alpha}})$ is the model-based estimation of w_{im}. The inverse weights can be included in a similar manner in the kernel smoothed objective function. Regression of the GPO requires an additional step, where we adopt the weighted marginal CDF

$$\widehat{F}_{km}(y; \widehat{\boldsymbol{\alpha}}) = n^{-1} \sum_{j=1}^{n} \frac{\eta_{jm}}{w_{jm}(\widehat{\boldsymbol{\alpha}})} I(Y_{jkm} \le y)$$

in the calculation of $\widehat{\mathcal{P}}_{im}$. The maximizers of the weighted objective function for single outcome and GPO are denoted as $\widehat{\boldsymbol{\theta}}^{W}$ and $\widehat{\boldsymbol{\theta}}_{\mathcal{P}}^{W}$, respectively.

For $\mathbf{z} = (\mathbf{x}, \mathbf{y}, \boldsymbol{\eta})$, $\boldsymbol{\alpha} \in \mathbf{A}$ and $\boldsymbol{\theta} \in \Theta$, define

$$\tau(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\alpha}) = \sum_{u=0}^{M} \sum_{v=0}^{M} \frac{\eta_u}{w(\mathbf{x}, \mathbf{y}, t_u, \boldsymbol{\alpha})} \mathbb{E} \left(\frac{\eta_{jv}}{w_{jv}(\boldsymbol{\alpha})} I(y_u > Y_{jv}) I[\mu(\mathbf{x}, t_u, \boldsymbol{\theta}) > \mu\{\mathbf{X}_j, t_v, \boldsymbol{\beta}(\boldsymbol{\theta})\}] \right.$$
$$\left. + \frac{\eta_{jv}}{w_{jv}(\boldsymbol{\alpha})} I(y_u < Y_{jv}) I[\mu\{\mathbf{x}, t_u, \boldsymbol{\beta}(\boldsymbol{\theta})\} < \mu\{\mathbf{X}_j, t_v, \boldsymbol{\beta}(\boldsymbol{\theta})\}] \right),$$

and $V(\boldsymbol{\alpha}) = \mathbb{E}\nabla_2 \tau(\cdot, \boldsymbol{\theta}_0, \boldsymbol{\alpha})/2$, where $w(\mathbf{x}, \mathbf{y}, t_u, \boldsymbol{\alpha})$ is the value of $w_{iu}(\alpha)$ when $\mathbf{X}_i = \mathbf{x}$ and $(\mathbf{Y}_{i0}, \ldots, \mathbf{Y}_{iM}) = \mathbf{y}$. Under the scenario with only single outcome, we establish the following theorem.

Theorem 3. *Under model* (2.1), *conditions (C1)-(C3) and (C12)-(C15) in the Supplementary Material, we have* $\widehat{\boldsymbol{\theta}}^W \xrightarrow{P} \boldsymbol{\theta}_0$, *and* $\sqrt{n}(\widehat{\boldsymbol{\theta}}^W - \boldsymbol{\theta}_0)$ *converges in distribution to* $\mathcal{N}\{0, V(\boldsymbol{\alpha}_0)^{-1}\Delta(\boldsymbol{\alpha}_0)V(\boldsymbol{\alpha}_0)^{-1}\}$, *where*

$$\Delta(\boldsymbol{\alpha}_0) = \mathbb{E}\{\nabla_1 \tau(\cdot, \boldsymbol{\theta}_0, \boldsymbol{\alpha}_0) + \mathbb{E}\nabla_{\boldsymbol{\alpha}} \nabla_1 \tau(\cdot, \boldsymbol{\theta}_0, \boldsymbol{\alpha}_0)\kappa(\cdot)\}^{\otimes 2},$$

and $\kappa(\cdot)$ *is the influence function of* $\widehat{\boldsymbol{\alpha}}$.

The uniform convergence of $w_{im}(\widehat{\boldsymbol{\alpha}})$ ensures the uniform convergence of $\widetilde{\mathcal{L}}_n(\boldsymbol{\theta}, \widehat{\boldsymbol{\alpha}})$, and thus guarantees the consistency of $\widehat{\boldsymbol{\theta}}^W$. Next, let $\mathbb{P}_n$ denote the average over the observed data. For asymptotic normality, the key step is to show that

$$\begin{aligned}\widetilde{\mathcal{L}}_n(\boldsymbol{\theta},\boldsymbol{\alpha})-\widetilde{\mathcal{L}}_n(\boldsymbol{\theta}_0,\boldsymbol{\alpha}) &= \frac{1}{2}\left(\boldsymbol{\theta}-\boldsymbol{\theta}_0\right)^T V(\boldsymbol{\alpha})\left(\boldsymbol{\theta}-\boldsymbol{\theta}_0\right)\\ &+ \frac{1}{\sqrt{n}}\left(\boldsymbol{\theta}-\boldsymbol{\theta}_0\right)^T W_n(\boldsymbol{\alpha})+o_P\left(\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2\right)+o_P\left(n^{-1}\right),\end{aligned}$$

uniformly over $O_P(1/\sqrt{n})$ neighbourhoods of $(\boldsymbol{\theta}_0,\boldsymbol{\alpha}_0)$, where

$$\begin{aligned}W_n(\boldsymbol{\alpha}) &= \sqrt{n}\mathbb{P}_n\nabla_1\tau(\cdot,\boldsymbol{\theta}_0,\boldsymbol{\alpha})\\ &= \sqrt{n}\mathbb{P}_n\nabla_1\tau(\cdot,\boldsymbol{\theta}_0,\boldsymbol{\alpha}_0)+\mathbb{E}\nabla_{\boldsymbol{\alpha}}\nabla_1\tau(\cdot,\boldsymbol{\theta}_0,\boldsymbol{\alpha}_0)\sqrt{n}(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)+o_P(1),\end{aligned}$$

It follows that

$$\begin{aligned}\widetilde{\mathcal{L}}_n(\boldsymbol{\theta},\widehat{\boldsymbol{\alpha}})-\widetilde{\mathcal{L}}_n(\boldsymbol{\theta}_0,\widehat{\boldsymbol{\alpha}}) &= \frac{1}{2}\left(\boldsymbol{\theta}-\boldsymbol{\theta}_0\right)^T V(\widehat{\boldsymbol{\alpha}})\left(\boldsymbol{\theta}-\boldsymbol{\theta}_0\right)+\\ &\quad \frac{1}{\sqrt{n}}\left(\boldsymbol{\theta}-\boldsymbol{\theta}_0\right)^T W_n(\widehat{\boldsymbol{\alpha}})+o_P\left(\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2\right)+o_P(n^{-1}),\end{aligned} \tag{2.7}$$

uniformly over $o_P(1)$ neighbourhoods of $\boldsymbol{\theta}_0$. The normality result then follows from (2.7), the consistency of $V(\widehat{\boldsymbol{\alpha}})$ to $V(\boldsymbol{\alpha}_0)$, asymptotic normality of $W_n(\widehat{\boldsymbol{\alpha}})$ and Theorem 4 of Sherman (1993). Finally, we can combine all of the techniques used in the proof of Theorems 3 and 4 to show the asymptotic properties of $\widehat{\boldsymbol{\theta}}_{\mathcal{P}}^{W}$ under the GPO setting with dropout, as stated in Theorem 4 below. Formal proofs of Theorems 3–4 are detailed in the Supplementary Material 1.3–1.4, together with the detailed forms of $\Delta_{\mathcal{P}}(\boldsymbol{\alpha}_0)$ and $V_{\mathcal{P}}(\boldsymbol{\alpha}_0)$.

Theorem 4. *Under model* (2.5)*, conditions (C5)-(C7), (C12)-(C14) and (C16)-(C19) in the Supplementary Material, we have* $\widehat{\boldsymbol{\theta}}_{\mathcal{P}}^{W} \xrightarrow{P} \boldsymbol{\theta}_{0,\mathcal{P}}$*, and* $\sqrt{n}(\widehat{\boldsymbol{\theta}}_{\mathcal{P}}^{W} - \boldsymbol{\theta}_{0,\mathcal{P}})$ *converges in distribution to* $\mathcal{N}\{0, V_{\mathcal{P}}(\boldsymbol{\alpha}_0)^{-1}\Delta_{\mathcal{P}}(\boldsymbol{\alpha}_0)V_{\mathcal{P}}(\boldsymbol{\alpha}_0)^{-1}\}$.

2.4 Variance estimation

Variance estimation for $\widehat{\boldsymbol{\theta}}$ are important for making inferences about the regression coefficients. However, several unknown quantities are involved in the asymptotic variance-covariance matrix of $\widehat{\boldsymbol{\theta}}$, preventing direct variance estimation according to the variance formula. A perturbation-resampling method (Jin et al., 2001; Cai et al., 2005) could be adopted instead. We illustrate the procedure for GPO under MAR below, and similar steps could be carried out for the kernel-smoothed estimators and for the scenario with only single longitudinal outcome.

Our perturbation scheme is designed to account for the variability in all components of the estimation procedure. Specifically, let $\{V_i\}_{i=1}^{n}$ be n i.i.d. positive random variables with unit mean and variance, such as $V_i \sim \text{Exp}(1)$. We start from the inverse probability weights for missing data. When logistic regression is used for modeling the missing mechanism, for example, we could use V_i as subject-specific weight in R function **glm**, and the resulting parameter estimates $\widehat{\boldsymbol{\alpha}}^*$ render perturbed inverse probability weights $w_{im}(\widehat{\boldsymbol{\alpha}}^*)$. To account for the variability in the estimated GPO, we calculate $\widehat{\mathcal{P}}^*_{im} = K^{-1}\sum_{k=1}^{K}\widehat{F}^*_{km}(Y_{ikm})$, where $\widehat{F}^*_{km}(y) = n^{-1}\sum_{j=1}^{n} V_j\eta_{jm}I(Y_{jkm} \leq y)/w_{jm}(\widehat{\boldsymbol{\alpha}}^*)$ using the same set of $\{V_i\}_{i=1}^n$. Finally, we construct

$$\mathcal{L}_n^{GPO*}(\boldsymbol{\theta}) = \frac{1}{n(n-1)}\sum_{1\leq i\neq j\leq n}\sum_{m=0}^{M}\sum_{v=0}^{M}\frac{V_iV_j\eta_{im}\eta_{jv}}{w_{im}(\widehat{\boldsymbol{\alpha}}^*)w_{jv}(\widehat{\boldsymbol{\alpha}}^*)}\Big[I(\widehat{\mathcal{P}}^*_{im} > \widehat{\mathcal{P}}^*_{jv}) \times I\left\{\mu(\mathbf{X}_i, t_m, \boldsymbol{\theta}) \leq \mu(\mathbf{X}_j, t_v, \boldsymbol{\theta})\right\}\Big],$$

and denote its maximizer as $\widehat{\boldsymbol{\theta}}^*$. Following the same lines as the proof of Theorems 1-4, it can be shown that $\sqrt{n}(\widehat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}_0) = n^{-1/2}\sum_{i=1}^{n} V_i\iota_i + o_p(1)$, where ι_i is the influence function of $\widehat{\boldsymbol{\theta}}$. Thus, the conditional distribution of $\widehat{\boldsymbol{\theta}}^*$ given the observed data has the same asymptotic distribution as the unconditional distribution of $\widehat{\boldsymbol{\theta}}$. In practice, we repeat the above process for B times, where B is a predetermined large number such as 200 or 400. A sample covariance matrix of $\widehat{\boldsymbol{\theta}}^*$ can be used as an estimator for the asymptotic covariance matrix of $\widehat{\boldsymbol{\theta}}$.

2.5 Numerical procedures

To maximize the objective functions, it is beneficial to supply good initial values in the optimization. We generate multiple (e.g., five) initial values for each dataset to protect against local minima. We design a procedure to effectively obtain multiple sets of good initial values. Following Clémençon et al. (2008), we adopt a convexified objective function,

$$\mathcal{C}_n(\boldsymbol{\theta};\psi) = \frac{1}{n(n-1)} \sum_{1\le i\ne j\le n} \sum_{u=0}^{M} \sum_{v=0}^{M} \psi\left[-\mathrm{sgn}(Y_{iu}-Y_{jv})\left\{\mu(\mathbf{X}_i,t_u,\boldsymbol{\theta}) - \mu(\mathbf{X}_j,t_v,\boldsymbol{\theta})\right\}\right],$$

where $\psi(\cdot)$ is a monotonically increasing convex function satisfying $\psi(0) = 1$ and $\psi(x) \ge I(x \ge 0)$. The minimizer of $\mathcal{C}_n(\boldsymbol{\beta};\psi)$ can be used as the initial values after standardization (i.e., dividing each of its component by the first component). For each initial value, we minimize $\mathcal{C}_n(\boldsymbol{\theta};\psi)$ over a subsample (with replacement) of m subjects from the original data. In our simulation, we found that this mechanism consistently produced good initial values with the choice of $\psi(x) = e^x$ and $m = \lfloor 3n^{1/2} \rfloor$, where $\lfloor x \rfloor$ is the maximum integer that does not exceed x. This specific choice costs minimal computation time at the order of $O(n)$ and was used throughout our numerical studies. For the GPO setting, we only need to replace $\mathrm{sgn}(Y_{iu} - Y_{jv})$ in $\mathcal{C}_n(\boldsymbol{\theta};\psi)$ with $\mathrm{sgn}(\widehat{\mathcal{P}}_{iu} - \widehat{\mathcal{P}}_{jv})$. When the data are MAR, inverse probability weighting is incorporated in computing $\mathcal{C}_n(\boldsymbol{\theta};\psi)$. We fed these initial values to the Nelder-Mead algorithm to maximize $\mathcal{K}_n(\boldsymbol{\theta};\phi)$ and retained the solution with the highest objective function. The bandwidth parameter c_n was set to $0.125n^{-1/3}\hat{\sigma}$ in our simulations, where $\hat{\sigma}$ was the sample standard error of $\mu\{\mathbf{X}_i, 0, \boldsymbol{\beta}(\widehat{\boldsymbol{\theta}}_C)\}$ with $\widehat{\boldsymbol{\theta}}_C$ being the minimizer of the convexified objective function computed on the full dataset.

3. Simulation study

3.1 Simulations setups

We conducted extensive simulations to examine the numerical performance of the proposed estimator. For each setup, we implemented the proposed method on 2000 simulated datasets, with sample sizes of $n = 200$ and 400. We conducted $B = 400$ perturbations for each dataset and obtained 95% percentile intervals. For the purpose of comparison, we implemented the GEE (Højsgaard et al., 2006) with an identity link and working independence correlation structure under setups S1-S3 below. We fix the first argument of the estimators in both models to 1, and we compute the standard error of rescaled estimators from the GEE model using the delta method.

3.1.1 Setups with single outcome

We considered $p = 3$ covariates, where $X_{i1} \sim \mathcal{N}(0, 1)$, $X_{i2} \sim Bernoulli(0.5)$ and $X_{i3} \sim \text{Unif}(0, 1)$. There were three visits indexed by $m = 0$ (baseline), 1, and 2. We considered the following three setups regarding the relationship between outcome and covariates:

Setup S1. $Y_{im} = \mu(\mathbf{X}_i, m, \boldsymbol{\beta}) + \varepsilon_{im}$, $\varepsilon_{im} \sim \mathcal{N}(0, 0.2)$,

Setup S2. $Y_{im} = H_1 \left\{\mu(\mathbf{X}_i, m, \boldsymbol{\beta}) + \varepsilon_{im}\right\}$, $H_1(x) = x/(1 + |x|)$, and $\varepsilon_{im} \sim \mathcal{N}(0, 0.16)$,

Setup S3. $Y_{im} = \exp\left[\left\{\mu(\mathbf{X}_i, m, \boldsymbol{\beta}) + 6\right\}/2.5\right] \times \varepsilon_{im}$, $\varepsilon_{im} \sim Unif(0.2, 0.6)$,

where $\mu(\mathbf{X}_i, m, \boldsymbol{\beta}) = X_{i1}\{\beta_{10} + \beta_{1m}I(m > 0)\} + X_{i2}\{\beta_{20} + \beta_{2m}I(m > 0)\} + X_{i3}\{\beta_{30} + \beta_{3m}I(m > 0)\} + \beta_t m$, $m = 0, 1, 2$. For all the setups, the error terms $(\varepsilon_{i0}, \varepsilon_{i1}, \varepsilon_{i2})$ followed the Frank copula (Nelsen, 2007) with parameter $\theta = 5.72$. We set the true value for $\boldsymbol{\beta} = (\beta_{10}, \beta_{20}, \beta_{30}, \beta_{11}, \beta_{21}, \beta_{31}, \beta_{12}, \beta_{22}, \beta_{32}, \beta_t)'$ as $(1, 0, 0.5, -0.2, -0.5, 0.6, -0.2, -0.5, 1.1, 0.5)'$. The outcome model in S1 featured a linear relationship between the risk score $\mu(\mathbf{X}_i, m, \boldsymbol{\beta})$ and outcome, for which the GEE model was correctly specified. True models under S2 and S3 were more complicated than a linear model, and the outcome was nonlinear in the risk score $\mu(\mathbf{X}_i, m, \boldsymbol{\beta})$. The error terms ε_{im} were additive in Model S2 and multiplicative in Model S3. The latter two settings were chosen to demonstrate the versatility of the proposed regression procedures under various covariate-outcome relationships.

We next incorporated missing data in these setups, by setting $\eta_{i0} \equiv 1$ for the baseline visit and generating the missing indicators η_{im} from the following logistic regression model for $m = 1, 2$:

$$\text{logit}\left\{\mathbb{P}(\eta_{im=1}|\eta_{i,m-1} = 1, Y_{i,m-1}, \mathbf{X}_i)\right\} = \alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \alpha_3 X_{i3} + \alpha_4 Y_{i,m-1}.$$

We set $\boldsymbol{\alpha} = (1.4, -0.5, 0, 0.5, -0.5)'$ for setup S1, $\boldsymbol{\alpha} = (1, 0, 0, 0.5, -0.5)'$ for setup S2, and $\boldsymbol{\alpha} = (1.5, 0.5, 0, 0.5, -0.1)'$ for setup S3. The missing rates were around 23% for visit 1 and 38% for visit 2 under all three setups.

3.1.2 Setups for the global percentile outcomes

We generated three covariates, where $X_{i1} \sim \mathcal{N}(0, 0.5)$, $X_{i2} \sim \text{Bernoulli}(0.5)$ and $X_{i3} \sim \mathcal{N}(0, 0.5)$. There were 3 visits, and there were 4 individual outcomes denoted as Y_{i1m}, Y_{i2m}, Y_{i3m}, Y_{i4m}, $m = 0, 1, 2$. The marginal distributions of $Y_1, \ldots, Y_4$ were $\mathcal{N}(0, 1)$, $\text{Exp}(1)$, $t(df = 3)$, and $\text{Unif}(-1, 1)$, respectively. The following two setups were considered for GPO:

Setup S4. $\mathcal{P}_{im} = H_2\left\{\mu(\mathbf{X}_i, m, \boldsymbol{\beta}) + \varepsilon_{im}\right\}$, $\varepsilon_{im} \sim \text{Exp}(1/\sqrt{5})$,

Setup S5. $\mathcal{P}_{im} = H_3\left[\exp\{\mu(\mathbf{X}_i, m, \boldsymbol{\beta})\}\varepsilon_{im}\right]$, $\varepsilon_{im} \sim \text{Unif}(0.2, 0.6)$, where $H_2(\cdot)$ and $H_3(\cdot)$ are monotonically increasing functions, $\mu(\mathbf{X}_i, m, \boldsymbol{\beta}) = X_{i1}\{\beta_{10} + \beta_{1m}I(m > 0)\} + X_{i2}\{\beta_{20} + \beta_{2m}I(m > 0)\} + X_{i3}\{\beta_{30} + \beta_{3m}I(m > 0)\} + \beta_t m$, $m = 0, 1, 2$, and error terms $(\varepsilon_{i0}, \varepsilon_{i1}, \varepsilon_{i2})$ followed the Frank copula model with parameter 5.72. True $\boldsymbol{\beta}_0 = (1, -1, 1.5, 0.3, 0, -0.25, 0.5, 0, -0.5, 0)'$. We incorporated missing data by generating η_{im}, $m = 1, 2$, from

$$\text{logit}\left\{\mathbb{P}(\eta_{im=1} \mid \eta_{i,m-1} = 1, \mathbf{Y}_{i,m-1}, \mathbf{X}_i)\right\} = \alpha_0 + \alpha_{X1}X_{i,1} + \alpha_{X2}X_{i,2} + \alpha_{X3}X_{i,3} + \alpha_{Y1}Y_{i,m-1,1} + \alpha_{Y2}Y_{i,m-1,2} + \alpha_{Y3}Y_{i,m-1,3} + \alpha_{Y4}Y_{i,m-1,4},$$

where $(\alpha_0, \alpha_{X1}, \alpha_{X2}, \alpha_{X3}, \alpha_{Y1}, \alpha_{Y2}, \alpha_{Y3}, \alpha_{Y4}) = (1.4, -0.5, 0, 0.5, 0, 0, 0, 0.1)$. This yielded around 20% missingness at visit 1 and about 36% missingness at visit 2 under both setups.

3.2 Simulation results

For the setups with single longitudinal outcome, Figure 1 summarizes the parameter estimation results under Setup S1–S3, with and without missing data. The corresponding summary tables are in Tables S1-S2 in the Supplementary Materials. The proposed estimator showed negligible biases, and the empirical bias shrinks when sample size increases. The empirical standard deviation (ESD) decreases with sample size roughly at the $n^{-1/2}$ rate, which agrees with our theoretical results. When missingness is present, our estimator has bias similar to the complete data case. The ESDs and ASEs were slightly larger when compared to their counterparts under complete data. The average standard error (ASE) based on perturbation resampling provides good approximation to the ESD, and the empirical coverage probability (ECP) of 95% confidence intervals is close to the nominal level, with improving performance as sample size increases.

Table 1 compares our method on complete data with the GEE method assuming an identity link for $n = 200$, and the corresponding summary table for $n = 400$ are presented in Tables S3 in the Supplementary Material. As expected, under Setup S1, the GEE was correctly specified and provided unbiased estimation of $\boldsymbol{\beta}$. Our estimator had larger ESDs than the GEE estimator, as we posed weaker model assumptions and only utilized the rank information from the outcomes. Under Setups S2–S3, where the GEE model was mis-specified, the GEE estimator could be severely biased with empirical coverage probability reaching 0 for some coefficients. The ESD of the two estimators are comparable to each other under Setups S2–S3.

Figure 2 summarizes the results with the longitudinal GPO under Setups 4 and 5, and the corresponding summary tables are presented in Tables S4-S5 in the Supplementary Material. Additionally, Supplemental Table S6 presents the simulation results under a modified Setup 4, where one of the

Table 1: Comparison of parameter estimation results by MRC and GEE on complete data for a single longitudinal outcome ($n = 200$).

Setup	Method		β_{20}	β_{30}	β_{11}	β_{21}	β_{31}	β_{12}	β_{22}	β_{32}	β_t
S1	MRC	Bias	0.001	0.003	0.005	-0.004	0.001	0.011	-0.008	0.015	0.005
		ESD	0.077	0.134	0.043	0.071	0.094	0.067	0.088	0.179	0.047
		ASE	0.079	0.137	0.050	0.082	0.113	0.074	0.100	0.202	0.053
		ECP	0.931	0.929	0.939	0.938	0.950	0.947	0.943	0.944	0.949
	GEE	Bias	-0.001	-0.003	0.000	-0.001	0.001	0.001	-0.001	0.001	0.001
		ESD	0.065	0.108	0.024	0.051	0.069	0.025	0.054	0.099	0.034
		ASE	0.063	0.109	0.024	0.051	0.069	0.024	0.055	0.097	0.033
		ECP	0.940	0.950	0.937	0.942	0.949	0.942	0.952	0.941	0.937
S2	MRC	Bias	0.001	0.003	0.005	-0.004	0.002	0.010	-0.006	0.014	0.004
		ESD	0.069	0.119	0.038	0.063	0.084	0.060	0.079	0.159	0.042
		ASE	0.071	0.123	0.045	0.073	0.101	0.066	0.089	0.181	0.047
		ECP	0.934	0.935	0.947	0.943	0.951	0.944	0.943	0.953	0.944
	GEE	Bias	-0.005	-0.011	-0.082	0.059	-0.075	-0.329	0.207	-0.669	0.010
		ESD	0.074	0.125	0.032	0.063	0.091	0.040	0.076	0.133	0.047
		ASE	0.072	0.126	0.031	0.061	0.090	0.039	0.074	0.130	0.045
		ECP	0.943	0.950	0.261	0.810	0.848	0.000	0.214	0.002	0.941
S3	MRC	Bias	0.002	0.005	0.009	-0.007	0.003	0.022	-0.016	0.033	0.009
		ESD	0.129	0.223	0.073	0.118	0.157	0.117	0.149	0.302	0.079
		ASE	0.132	0.231	0.084	0.135	0.185	0.127	0.167	0.337	0.089
		ECP	0.938	0.940	0.944	0.940	0.951	0.946	0.946	0.946	0.944
	GEE	Bias	0.010	0.018	0.181	-0.141	0.179	0.539	-0.340	1.066	0.011
		ESD	0.120	0.199	0.062	0.103	0.165	0.087	0.151	0.300	0.084
		ASE	0.116	0.198	0.060	0.101	0.162	0.085	0.149	0.290	0.082
		ECP	0.942	0.949	0.153	0.713	0.819	0.000	0.361	0.029	0.949

Bias = empirical average of the bias for the parameter estimate, ESD = empirical standard deviation of the estimates over replications, ASE = average of estimated standard errors over replications, and ECP = empirical coverage probability of 95% confidence interval.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0049

individual outcomes was dichotomized to be binary. Under all these GPO scenarios, the patterns were quite similar to those observed for the single longitudinal outcome scenario, and the empirical bias was negligible in all setups considered. Again, the proposed rank estimator performed satisfactorily for both the additive-effect scenario (Setup S4) and the multiplicative-effect scenario (Setup S5). The empirical coverage rates were all close to the nominal level of 95% . Though the GPO were estimated rather than observed, the ASE continued to provide a satisfactory estimate for ESD. Indeed, the impacts of estimated GPO and missing weights on the ESD were quite small empirically, as shown in Supplemental Table S7, though theoretically they lead to additional components in the variance formula. Supplemental Table S8 displays the simulation results with $B = 200$, 400, and 600 under Setup S4. The results demonstrate that the ASE and C95 are comparable for all values of B, suggesting that the results are relatively robust to the choice of B, as long as B is sufficiently large. We also conducted sensitivity studies to investigate the impact of using a mis-specified probit regression model to estimate the missing weights, while the true mechanism follows the aforementioned logistic regression model. The results, as shown in Supplemental Table S9, remain satisfactory, indicating that our proposed methods are relatively insensitive to the choice of weighting scheme model.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0049

Taken together, these simulation results demonstrate the robustness of our methods in the presence of various outcome distributions, link functions, and covariate effect mechanisms.

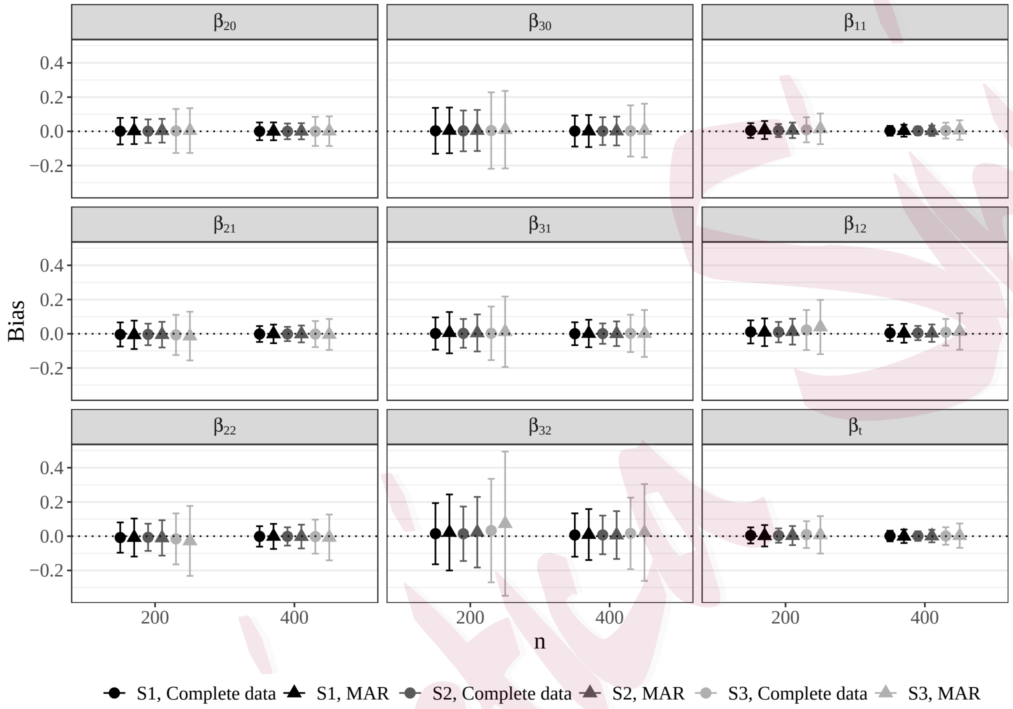

Figure 1: Performance of the proposed method for single longitudinal outcome under Setups S1-S3. The points in the figures denote the average bias over 2000 repetitions and the error bars correspond to the empirical standard deviation.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0049

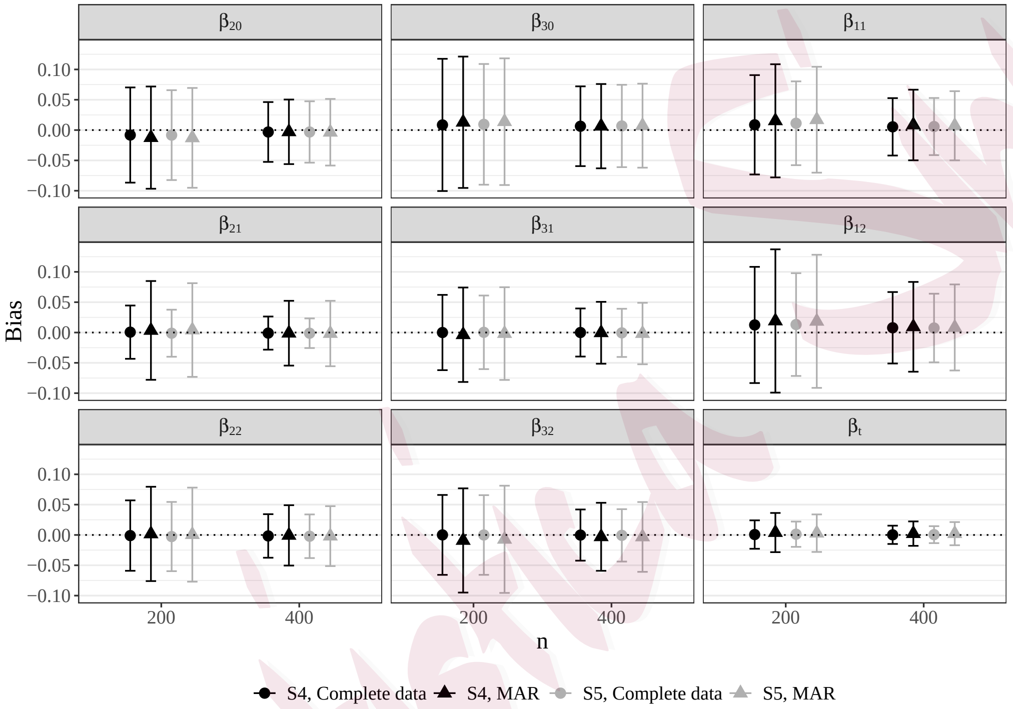

Figure 2: Performance of the proposed method for longitudinal GPO formed by the four individual outcomes under Setups S4-S5. The points in the figures denote the average bias over 2000 repetitions and the error bars correspond to the empirical standard deviation of the bias.

4. Real data analysis

We applied the proposed methods to the data from NET-PD LS-1 study (Kieburtz et al., 2015). The study was a multi-center, double-blind, placebo-controlled, 1:1 randomized efficacy trial, with the objective to determine whether creatine monohydrate was effective in slowing the progression of long-term clinical decline among 1741 patients with early and treated PD. In the analysis here, we are interested in examining risk factors associated with the global disease burden and/or the progression of PD. Considering that PD is a multi-factorial disease, we follow Kieburtz et al. (2015) and capture the global disease burden using five individual outcomes, including the Modified Schwab and England Activities of Daily Living Scale (ADL), 39-item Parkinson's Disease Questionnaire Summary Index (PDQ-39), UPDRS questions related to ambulatory capacity (AC), Symbol Digit Modalities (SDM) test, and the Modified Rankin Scale (mRS) measured at baseline and at yearly follow-up visits. ADL and SDM were reversely coded, such that higher values are worse for all outcomes. These measures collectively capture patient's motor, cognitive and behavioral disability (Elm and Investigators, 2012). Participants were randomized to placebo or creatine and were projected to follow up for a minimum of 5 years. The study was terminated early based on an interim analysis, where the two-sample global rank-sum test yielded a 2-sided p-value of 0.45, indicating that there was limited significant evidence that creatine monohydrate could improve clinical outcomes. In our analytic dataset, the median and maximum follow-up years were 4 and 6 years, respectively. The missing rates at year 1 to 6 were 7.8%, 13.7%, 20.2%, 39.8%, 61.2% and 82.6%. The missing mechanism was deemed to satisfy the MAR, as the main reason of a missed visit was administrative and due to study termination.

We applied the proposed method to the longitudinal GPO formed by the five outcomes. We considered as covariates a comprehensive set of baseline risk factors, time (in years), as well as the interaction between these baseline factors and time if significant. Under this model, the main effect of a covariate corresponds to its association with the level of the global disease burden during the follow-up. In addition, a positive interaction with time would imply a higher rate of progression in terms of the global percentile, such that patients with higher values in this baseline factor tend to have worsening global percentile/ranks. Following Bega et al. (2015) and based on our preliminary analyses, we included the following covariates measured at baseline: age at symptom onset (Onset Age), treatment arm (creatine vs. placebo), Beck Depression Inventory (BDI) score, EuroQOL-5D (EQ-5) score, Levodopa equivalent dose (LED), predominant PD symptom type (postural-instability gait predominant [PIGD] subtype vs other), Race (non-Hispanic white vs others), Scales for Outcomes of Parkinson's Disease Cognition (SCOPA-COG) score, Total Functional Capacity (TFC) score, and total UPDRS score. We also identified significant interactions with follow-up time (t) for Onset Age and Race. All continuous variables except follow-up time were centered at the mean, and we scaled the continuous variables using meaningful units, e.g., per 10 years for Onset Age, see Table 1. The missingness was modeled by logistic regression, where the following variables were included based on clinical insights and model selection: follow up time, duration of diagnosis at time of study entry, Onset Age, income, LED, and AC and SDM in the previous year. For the purpose of comparison, we also fitted a weighted GEE model with independent working correlation structure using the *geeglm* function in R from the **geepack** package. In our analysis, the coefficient for total UPDRS score were fixed to 1 under both our method and the weighted GEE model. The standard errors were computed from 400 perturbations.

Table 1 summarizes the results from the two methods, which are consistent for most of the risk factors. As expected, the study treatment creatine showed no significant effect, which is consistent with the primary conclusion of the trial based on a two-sample test of the global rank-sum.

The PIGD subtype, lower SCOPA-COG, TFC, EQ-5, and higher BDI total score, were explanatory of worse global disease burden during the study follow-up. While LED was insignificant in our model, it was associated with worse disease burden under the GEE. Considering that levodopa is the main medication for management of PD symptoms, the significance of LED under GEE may be an artifact of confounding or model misspecification. In terms of risk factors for accelerated disease progression, the time-varying effect of Onset Age and Race were identified in both models. Specifically, patients with older onset age tend to have faster progression. Though PD is most prevalent among the non-Hispanic White, our analysis suggests that patients from other racial and ethnicity groups tend to experience faster progression, in that they would have worsening global percentile relative to the study population as time elapses. Our method led to a much larger coefficient for the race by time interaction when compared to the GEE, suggesting a larger racial disparity in terms of the rate of global disease progression. Our results are insensitive to the range and scale change over time for the outcome variables, and our modeling procedure thus provides robust insights into the progression global PD status.

Table 2: Analysis of the NET-PD LS-1 data.

	MRC			weighted GEE		
Risk factor	Estimate	SE	P-value	Estimate	SE	P-value
Onset Age (per 10 years)	0.200	0.135	0.137	0.252	0.070	0.000
PIGD subtype	1.031	0.230	0.000	0.924	0.165	0.000
Race (non-Hispanic White)	-0.034	0.458	0.941	0.011	0.208	0.956
LED (per 1000 units)	0.318	0.449	0.478	1.024	0.275	0.000
BDI	0.141	0.029	0.000	0.088	0.017	0.000
EQ-5 (per 0.1 unit)	-0.442	0.097	0.000	-0.310	0.057	0.000
SCOPA-COG	-0.114	0.024	0.000	-0.099	0.016	0.000
TFC	-0.637	0.131	0.000	-0.422	0.080	0.000
Creatine	0.106	0.206	0.608	0.039	0.135	0.774
Time (t)	0.434	0.195	0.026	0.272	0.108	0.012
Onset Age $\times$ t	0.209	0.048	0.000	0.165	0.032	0.000
Race $\times$ t	-0.481	0.203	0.018	-0.242	0.111	0.029

The coefficient for baseline total UPDRS score (per 10 units) were fixed to 1.

5. Discussion

In this paper, we proposed rank regression strategies for univariate longitudinal outcomes and longitudinal global percentile outcomes. Our method requires minimal assumptions and allows for a broad range of relationships between covariates and the longitudinal outcome, thereby offering robust estimation of the regression coefficients. We established consistency and asymptotic normality of our proposed estimators. For univariate longitudinal outcome subject to missing at random, we demonstrated that the proposed estimator is asymptotically normal, provided that the estimator of the missing data model has asymptotic normality. Under the longitudinal global percentile outcome setting, the non-differentiable term $I(\widehat{\mathcal{P}}_{iu} > \widehat{\mathcal{P}}_{jv})$ poses an additional challenge to the theoretical justification. We overcome this difficulty by approximating the indicator function with a smooth function and then prove that this approximation does not affect the asymptotic properties of the proposed estimator. Our methods perform satisfactorily under various simulation studies and are readily applicable to many longitudinal studies.

In this work, we have focused on longitudinal data with pre-planned follow-up times. When the follow-up times are irregular, it would be difficult to estimate the GPO directly using the empirical CDF. However, this difficulty could be resolved by first estimating a time-specific marginal CDF using kernel smoothing, borrowing information from patients with similar visit times. Next, our method can be extended to the high-dimensional setting by incorporating suitable penalization terms. Finally, we have focused on the regression coefficients in this paper, but it may be desirable to also estimate the unspecified link function. These directions are beyond the scope of the current work, but they will be investigated in future research.

Supplementary Material

Supplemental Material contains detailed proof for Theorems 1–4 and additional simulation results.

Acknowledgments

Authors are grateful to the editor, associate editor, and the reviewers for their valuable comments. This work was partially supported by grants from the National Institutes of Health (R03NS108136, U24CA230144, R01CA269696) and National Science Foundation (DMS-2345035).

References

Bega, D., S. Kim, Y. Zhang, J. Elm, J. Schneider, R. Hauser, A. Fraser, and T. Simuni (2015). Predictors of functional decline in early parkinson's disease: NET-PD LS1 cohort. *Journal of Parkinson's Disease 5*(4), 773–782.

Berry, J. D., R. Miller, D. H. Moore, M. E. Cudkowicz, L. H. Van Den Berg, D. A. Kerr, Y. Dong, E. W. Ingersoll, and D. Archibald (2013). The combined assessment of function and survival (cafs): A new endpoint for als clinical trials. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration 14*(3), 162–168.

Cai, T., L. Tian, and L. Wei (2005). Semiparametric box–cox power transformation models for censored survival observations. *Biometrika 92*(3), 619–632.

Cavanagh, C. and R. P. Sherman (1998). Rank estimators for monotonic index models. *Journal of Econometrics 84*(2), 351–382.

Clémençon, S., G. Lugosi, N. Vayatis, et al. (2008). Ranking and empirical minimization of u-statistics. *The Annals of Statistics 36*(2), 844–874.

Elm, J. J. and N. N.-P. Investigators (2012). Design innovations and baseline findings in a long-term parkinson's trial: The national institute of neurological disorders and stroke exploratory trials in parkinson's disease long-term study–1. *Movement Disorders 27*(12), 1513–1521.

Felker, G. M. and A. S. Maisel (2010). A global rank end point for clinical trials in acute heart failure. *Circulation: Heart Failure 3*(5), 643–646.

Han, A. K. (1987). Non-parametric analysis of a generalized regression model: The maximum rank correlation estimator. *Journal of Econometrics 35*(2-3), 303–316.

Højsgaard, S., U. Halekoh, and J. Yan (2006). The R package geepack for generalized estimating equations. *Journal of statistical software 15*, 1–11.

Huang, P., B. C. Tilley, R. F. Woolson, and S. Lipsitz (2005). Adjusting O'Brien's test to control type I error for the generalized nonparametric behrens–fisher problem. *Biometrics 61*(2), 532–539.

Jin, Z., Z. Ying, and L. Wei (2001). A simple resampling method by perturbing the minimand. *Biometrika 88*(2), 381–390.

Kieburtz, K., B. C. Tilley, J. J. Elm, D. Babcock, R. Hauser, G. W. Ross, A. H. Augustine, E. U. Augustine, M. J. Aminoff, I. G. Bodis-Wollner, et al. (2015). Effect of creatine monohydrate on clinical progression in patients with parkinson disease: A randomized clinical trial. *JAMA 313*(6), 584–593.

Lin, H. and H. Peng (2013). Smoothed rank correlation of the linear transformation regression model. *Computational Statistics & Data Analysis 57*(1), 615–630.

Mao, L. and T. Wang (2021). A class of proportional win-fractions regression models for composite outcomes. *Biometrics 77*(4), 1265–1275.

Molenberghs, G., G. Fitzmaurice, M. G. Kenward, A. Tsiatis, and G. Verbeke (2014). *Handbook of missing data methodology.* CRC Press.

Nelsen, R. B. (2007). *An introduction to copulas.* Springer Science & Business Media.

O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics 40*(4), 1079–1087.

Ramchandani, R., D. A. Schoenfeld, and D. M. Finkelstein (2016). Global rank tests for multiple, possibly censored, outcomes. *Biometrics 72*(3), 926–935.

Sherman, R. P. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica: Journal of the Econometric Society 61*(1), 123–137.

Sun, H., B. A. Davison, G. Cotter, M. J. Pencina, and G. G. Koch (2012). Evaluating treatment efficacy by multiple end points in phase ii acute heart failure clinical trials: Analyzing data using a global method. *Circulation: Heart Failure 5*(6), 742–749.

Yi, G. Y. and W. He (2009). Median regression models for longitudinal data with dropouts. *Biometrics 65*(2), 618–625.

Zhang, J., Z. Jin, Y. Shao, and Z. Ying (2018). Statistical inference on transformation models: a self-induced smoothing approach. *Journal of Nonparametric Statistics 30*(2), 308–331.

Zhang, W., A. Liu, L. L. Tang, and Q. Li (2019). A cluster-adjusted rank-based test for a clinical trial concerning multiple endpoints with application to dietary intervention assessment. *Biometrics 75*(3), 821–830.

Department of Statistics, Rice University, Houston, TX.

E-mail: dmmatustc@gmail.com

Department of Biostatistics,The University of Texas MD Anderson Cancer Center

E-mail: jning@mdanderson.org

Phone: 713-792-5310

Department of Statistics and Data Science, Washington University in St. Louis

E-mail: HEX@WUSTL.EDU

Department of Neurology, Massachusetts General Hospital and Harvard Medical School

E-mail: AWILLS@mgh.harvard.edu

Department of Biostatistics & Data Science, The University of Texas Health Science Center at

Houston

E-mail: ruosha.li@uth.tmc.edu

Phone: 713-500-9572