Biostatistics, 2025, **26**(1), kxaf013
https://doi.org/10.1093/biostatistics/kxaf013
Article

Predicting distributions of physical activity profiles in the National Health and Nutrition Examination Survey database using a partially linear Fréchet single index model

Marcos Matabuena[1,*], Aritra Ghosal[2], Wendy Meiring[3], Alexander Petersen[4]

[1]Department of Biostatistics, Harvard University, 677 Huntington Avenue, Building 2, MA 02115, United States
[2]Department of Biostatistics, St. Jude Children's Research Hospital, 262 Danny Thomas Place, TN 38105, United States
[3]Department of Statistics and Applied Probability, University of California, South Hall, Santa Barbara, CA 93106, United States
[4]Department of Statistics, Brigham Young University, 223 TMCB, UT 84602, United States

[*]Corresponding author: Department of Biostatistics, Harvard T.H. Chan School of Public Health, School of Public Health 2, 677 Huntington Ave, Boston, MA 02115, United States. Email: mmatabuena@hsph.harvard.edu

ABSTRACT

Object-oriented data analysis is a fascinating and evolving field in modern statistical science, with the potential to make significant contributions to biomedical applications. This statistical framework facilitates the development of new methods to analyze complex data objects that capture more information than traditional clinical biomarkers. This paper applies the object-oriented framework to analyze physical activity levels, measured by accelerometers, as response objects in a regression model. Unlike traditional summary metrics, we utilize a recently proposed representation of physical activity data as a distributional object, providing a more nuanced and complete profile of individual energy expenditure across all ranges of monitoring intensity. A novel hybrid Fréchet regression model is proposed and applied to US population accelerometer data from National Health and Nutrition Examination Survey (NHANES) 2011 to 2014. The semi-parametric nature of the model allows for the inclusion of nonlinear effects for critical variables, such as age, which are biologically known to have subtle impacts on physical activity. Simultaneously, the inclusion of linear effects preserves interpretability for other variables, particularly categorical covariates such as ethnicity and sex. The results obtained are valuable from a public health perspective and could lead to new strategies for optimizing physical activity interventions in specific American subpopulations.

KEYWORDS: accelerometer devices; complex survey designs; distributional representations; functional data analysis.

1. INTRODUCTION

Medical science is living in a golden age with the expansion of the clinical paradigms of digital and precision medicine (Li et al. 2017; Kosorok and Laber 2019; Topol 2019; Onnela 2021; Javaid et al. 2022). In this new context, it is increasingly common to record participant data that is most faithfully represented using complex statistical objects such as probability distributions

Received: December 2, 2023. **Revised:** April 15, 2025. **Accepted:** April 16, 2025

(Matabuena et al. 2021; Ghosal et al. 2022, 2023b; Matabuena and Petersen 2023) that contain enriched information compared to traditional clinical biomarkers in predictive terms. Distributional representations can be seen as natural digital functional biomarkers to analyze wearable data information. In a series of papers, the performance of the distributional representation was compared with that of existing summary metrics, providing strong evidence of their advantages in diabetes and physical activity domains (Matabuena et al. 2021, 2022; Cui et al. 2023; Ghosal et al. 2023b; Jašková et al. 2023; Matabuena and Petersen 2023). Distributional representations are a direct functional extension of traditional compositional metrics (Battelino et al. 2019; Biagi et al. 2019; Park et al. 2025) and facilitate the creation of synthetic profiles over a continuum of intensities measured by wearable devices that provide an individualized profile of the participant's activity. Importantly, these representations overcome the critical limitations of compositional metrics that require one to define specific cut-off points to categorize participant information that can introduce subjectivity and be highly dependent on the population under analysis.

This work is motivated by the task to uncover factors that are associated with the physical activity patterns of the American population, where these patterns are represented as distributional data objects. As energetic expenditure behaves nonlinearly with age (Schrack et al. 2012), and other anthropometrical measures (Mehta et al. 2017), more advanced and flexible regression models are required to overcome the limitations of the linear model. Here, to provide a good balance between the advantages and disadvantages of linear and nonlinear models, the proposed model extends the partially linear model for scalar responses (Liang et al. 2010) to the case of a distributional response object, yielding a partially linear Fréchet single index model. Analogous to the scalar response case, this can be viewed as an extension of the recently proposed global Fréchet regression and Fréchet single index models (Petersen and Müller 2019; Bhattacharjee and Müller 2023; Ghosal et al. 2023a). Furthermore, the survey weights from the complex survey design of the NHANES are incorporated into the model estimation to obtain reliable population-based results according to the composition of the US population (Lumley 2010).

From a public health point of view, the proposed model is attractive because it elucidates the impact certain variables exert on the American population's average physical activity level profiles along the full range of accelerometer intensities. Moreover, these new findings can help to refine and plan specific health interventions that reduce the gap in physical inactivity in different US sub-populations. For example, one of the follow-up analyses conducted herein extracts clinical phenotypes of individuals to characterize the participants who are more or less active than predicted by the regression model.

The structure of the paper is as follows. Section 2 introduces the NHANES data that will be analyzed, together with a background of the physical activity distributional representations. Section 3 introduces the model and an efficient, spline-based estimator. Section 4 reports the various analyses performed. Finally, Section 5 discusses the results from a public health perspective, this paper's role in the broader statistical literature on regression models in metric spaces, and its opportunities in the medical field to analyze other complex statistical objects. Additional results, included in the Supplementary Materials, explore novel semi-supervised physical activity phenotypes and demonstrate how classical physical activity metrics can be interpreted as specific instances of our distributional models.

1.1 Contributions

The methodological contributions of this paper will first be summarized, followed by the findings from the analysis of the NHANES database using the proposed regression model with responses being the distributional representation of physical activity profiles.

- To the authors' knowledge, the proposed model is the first partially linear Fréchet single index (PL-FSI) regression model for responses that are probability distributions, viewed as elements of the L^2-Wasserstein space. Moreover, for this particular situation, the use of splines are introduced for the first time in the Fréchet regression modeling framework.

- An efficient optimization strategy is proposed to address the complex survey sampling mechanism of the NHANES data that retains the estimator's form of a weighted least squares problem. The key idea of our approach is to estimate the model's nonlinear component by means of regression splines after projecting the variables in the nonlinear term to a single covariate.
- The primary findings of the analyses conducted on the NHANES are:
  1. The proposed single-index model is shown to outperform both the global Fréchet model and a competing, more general, partially linear Fréchet regression model in terms of the adjusted Fréchet R^2 measure (Petersen and Müller 2019).
  2. Interpretations are provided for the effects of ethnicity and other interesting variables on distributional physical activity profiles. These novel analyses can provide new insights into how physical activity varies among the various US sub-populations.
  3. New physical activity phenotypes are constructed corresponding to individuals who do more or less exercise than is predicted by the model using the distributional representation. These analyses are new and help to examine how well individuals adhere to the recommended physical activity guidelines.

The code for reproducing the results presented using the methods proposed in this paper are publicly available on GitHub at https://github.com/aghosal89/FSI_NHANES_Application.

1.2 Literature review

Statistical regression analysis of response data in metrics spaces is an increasingly prominent research direction in the statistical community (Petersen and Müller 2016; Yang et al. 2020; Fan and Müller 2024; Petersen et al. 2021; Dubey and Müller 2022; Jeon et al. 2022; Petersen et al. 2022; Tucker 2022; Zhou and Müller 2022; Bhattacharjee and Müller 2025; Chen and Müller 2023; Chen et al. 2023; Ghosal et al. 2023a; Lin et al. 2023b). The first papers on hypothesis testing (Lyons 2013; Dubey and Müller 2019; Petersen et al. 2021), variable selection (Tucker et al. 2023; Coulter et al. 2024), causal inference (Katta et al. 2024), multilevel models (Matabuena and Crainiceanu 2024), uncertainity quantification (Lugosi and Matabuena 2024; Matabuena et al. 2024), semi-parametric regression models (Bhattacharjee and Müller 2023; Ghosal et al. 2023a), and non-parametric regression (Hanneke 2022), have recently appeared. For classical regression models with univariate response data, single index models, including partially linear ones, have been a topic with particular popularity in the last 20 yr in the statistical and econometrics literature (Carroll et al. 1997; Horowitz 2012). There are several works in this direction, including recent extensions of the model to functional data (Wang et al. 2016; Wong et al. 2019; Xiao et al. 2021; Zhu et al. 2022). However, the authors are not aware of any existing extension of partially linear single index models to response data in metric spaces, even for the special case of distributional response data, or that incorporates the complex survey design into the analysis.

2. MOTIVATING EXAMPLE: DATA ON WEARABLE ACCELEROMETER DEVICES FROM NHANES 2011 TO 2014

The NHANES aims to provide a broad range of descriptive health and nutrition statistics for the non-institutionalized civilian US population (Johnson et al. 2014). Data collection consists of an interview and an examination. The interview gathers personal demographic, health, and nutrition information; the examination includes physical measurements such as blood pressure, a dental examination, and the collection of blood and urine specimens for laboratory testing. Additionally, participants were asked to wear a physical activity monitor, starting on the day of their exam, and to keep wearing this device all day and night for seven full days (midnight to midnight) and remove it on the morning of the 8th day. The device used was the ActiGraph GT3X+ (ActiGraph of Pensacola, FL). Data from the NHANES cohorts 2011 to 2014 were used for the analyses in this paper (Johnson et al. 2014).

Physical activity signals were pre-processed by staff from the National Center for Health Statistics (NCHS) to determine signal patterns that were unlikely to result from human movement. Acceleration measurements were then summarized at the minute level using Monitor-Independent Summary (MIMS) units, an open-source, device-independent universal summary metric (John et al. 2019).

Throughout the figures and text, we will refer to MIMS as Activity Counts for consistency and clarity. In order to further increase the reliability of the analysis, we use the following filter criteria strategy extracted from Smirnova et al. (2020) in order to remove participants with poor quality in their accelerometry data. Those participants who (i) had fewer than 3 d of data with at least 10 h of estimated wear time or (ii) had non-wear periods, identified as intervals with at least 60 consecutive minutes of zero activity counts and at most 2 min with counts between 0 and 5 were deemed by NHANES to have poor quality and hence were removed. These protocol instructions were adopted from high-level accelerometer research (see, for example, Troiano et al. 2008).

2.1 Quantile function representation of physical activity profiles

A novel representation of the resulting data is herein adopted that extends previous compositional metrics to a functional setting (Matabuena and Petersen 2023), aimed at overcoming their dependency on certain physical activity intensity thresholds. This approach also overcomes some previously known limitations of more traditional approaches. Let $i \in \{1, 2, \ldots, n\}$ be the index for participants, where n is the total number of participants in the study. For the ith participant, let M_i indicate the number of days (including partial days) for which accelerometer records are available and n_i be the number of observations recorded in the form of pairs $(m_{ij}, A_{ij}), j \in \{1, \ldots, n_i\}$. Here, the m_{ij} are a sequence of time points in the interval $[0, M_i]$ in which the accelerometer records activity information and A_{ij} is the measurement of the accelerometer at time m_{ij}. No data are available during non-wear periods.

In this paper, each individual's accelerometer measurements, $\{A_{ij}\}_{j=1}^{n_i}$, are studied without regard to their ordering in time. They are thus characterized by the empirical quantile function, $Y_i(t) = \hat{Q}_i(t)$, for $t \in [0, 1]$, which will be used as the response in the regression model. Here, $\hat{Q}_i(t) = \inf\{a \in \mathbb{R} : \hat{F}_i(a) \geq t\}$ is the generalized inverse of $\hat{F}_i(a) = \frac{1}{n_i} \sum_{j=1}^{n_i} 1\{A_{ij} \leq a\}, a \in \mathbb{R}$, the empirical cumulative distribution function for the physical activity values for the ith individual. In order to illustrate clearly the difficulty of analyzing raw physical activity data from participants who are monitored during different periods and in different experimental conditions, Fig. 1 shows the plot of observed A_{ij} against m_{ij} for an arbitrarily chosen participant in the study. In Fig. 2, the left panel shows the empirical quantile functional representation of the physical activity measurements of the participant whose raw measurements are shown in Fig. 1, while the right panel shows the empirical quantile functions of all participants after transforming the raw time series physical activity data into distributions of the physical activity. Quantile function representations of the physical activity trajectories overcome the problems of the traditional summary metrics of physical activity when the raw time series have different lengths. In addition, the new representation uses all accelerometer intensities (over a continuum) to construct the new physical activity functional profile, generalizing traditional metrics of physical activity that summarize the information in a compositional vector.

The precise relationship between the distributional representation and traditional compositional metrics can be understood as follows. Given a subject's cumulative distribution function (CDF) of physical activity, $\hat{F}_i(a)$, a compositional metric is formed by creating bins $[a_{j-1}, a_j]$ for $a_0 < a_1 < \ldots < a_M$. This process results in an M-dimensional compositional vector $\mathbf{U}_i$, where each component is defined as

$$U_{ij} = \hat{F}_i(a_j) - \hat{F}_i(a_{j-1}).$$

Thus, any desired compositional information can always be derived from the distribution, and the distributional representation can be thought of as a continuous extension of the compositional

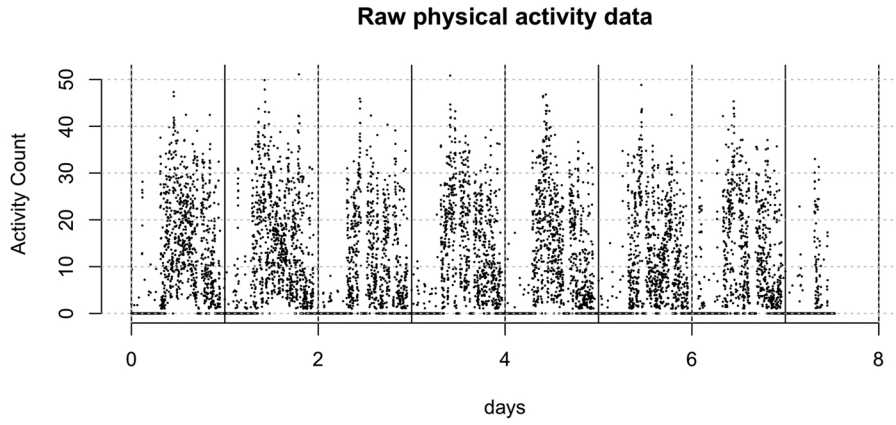

Fig. 1. The plot of physical activity time series A_{ij} of one representative participant (one chosen i) in the NHANES 2011 to 2014 study monitored during 8 d are plotted over the observed time intervals m_{ij}, when the physical activity measurements are counted as described in Section 2.1.

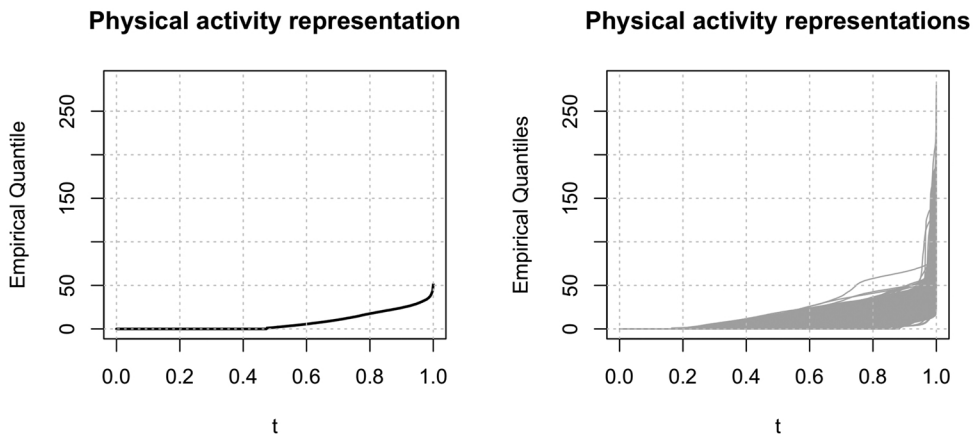

Fig. 2. (Left) The empirical quantile representation $\hat{Q}_i$, of the activity profile of the participant (chosen i) described in Fig. 1 above, also described in the Section 2.1. (Right) The estimated empirical quantile functions of physical activity profiles for all of the 4616 participants in the study.

approach. Most importantly, when physical activity distributions are the responses in a regression model, distributional predictions produced by the regression fit can always provide *any* desired compositional information regardless of the choice of bins or thresholds $a_j, j = 0, 1, \ldots, M$. Hence, from a practical perspective, there are two main advantages to choosing the distributional representation. First, it avoids collapsing the information of physical activity into intervals, preventing potential loss of information. Second, it eliminates the need to define thresholds that, in the context of physical activity, depend on population and physical characteristics such as age, sex, and BMI. This removes arbitrariness and subjectivity from the analysis.

Another key distinction between this work and others that have used distributional representations is that, while these previous papers have focused only on positive observations of physical activity counts (Ghosal et al. 2025; Lin et al. 2023a), the current work also incorporates periods of recorded inactivity. Inactivity behavior is an important component of human physical activity profiles, and the proposed methodological approach allows for the integration of both inactivity and activity behaviors comprehensively.

Table 1. Summaries of the covariates Age, BMI (Body Mass Index), HEI (Healthy Eating Index), TAC (Total Activity Count), and Ethnicity, stratified by Sex.[a]

Covariates	Men	Women
Numeric variables		
Age	47.45 (16.45)	48.082 (16.50)
Body mass index	28.72 (5.73)	29.18 (7.41)
Healthy eating index	53.013 (14.13)	56.63 (14.75)
Total activity count	9.8 (3.1)	9.7 (2.6)
Ethnicities		
Mexican American	8.55 %	6.42 %
Other Hispanic	5.42 %	5.28 %
Non-Hispanic White	70.65 %	72.3 %
Non-Hispanic Black	8.82 %	10.29 %
Non-Hispanic Asian	3.77 %	3.37 %
Other races including multi-racial	2.79 %	2.35 %

[a]In the first column, the levels of the categorical variable Ethnicity are separated from the numerical covariates Age, BMI, HEI, and TAC. In the third and fourth columns, the first four rows present the means and, in brackets, the standard deviations of the continuous variables (Age, BMI, HEI, and TAC) for men and women respectively. Rows 5to 10 in the same columns represent the percentage breakdown of the sub-populations of men and women into their respective ethnicities. The description of the covariates are found in Section 2.2.

2.2 Details of covariates

In addition to the accelerometer data, the covariates used in the model include sociodemographic, dietary, and clinical variables such as age, Body Mass Index (BMI), and Healthy Eating Index (HEI), along with the categorical variables Ethnicity and Sex and their interaction. HEI is a continuous score reflecting the overall diet quality of each participant. The ethnicity variable reported the racial origin of the participants divided into the following six categories: Mexican American, Other Hispanic, Non-Hispanic White, Non-Hispanic Black, Non-Hispanic Asian, and Other races, including Multi-racial. The age range of the participants in the analysis was 20 to 80 yr. The BMI (kg/m^2) was restricted to the range 18.5 to 40 to study individuals ranging from healthy to highly overweight/obese. Under these restrictions, $n = 4616$ individuals were chosen for the analysis. Although not included as a predictor in the models analyzed in Section 4, the Total Activity Count (TAC) is a widely used summary metric for accelerometer data that represents the average activity count over the complete series of physical activity measurements recorded by the device. In the data cohort used for analysis, each element of the sequence $\{A_{ij}\}_{j=1}^{n_i}$ is measured in MIMS units, unlike NHANES 2003 to 2006, which uses Actigraph counts. For consistency, since the TAC variable in both datasets represents the average count units of the monitor, this metric is termed Total Activity Count (TAC) in both settings. More specifically, in the notation of Section 2.1, the TAC of the ith participant is $TAC_i = \frac{1}{n_i}\sum_{j=1}^{n_i} A_{ij}$. In this sense, TAC represents a scalar summary of the distributional representation used in this paper. Supplementary Materials demonstrate how the proposed model, which relates covariates to the full physical activity distribution, can subsequently be used to interpret relationships between covariates and summaries of interest such as TAC. Univariate summaries of each of these covariates, stratified by sex, are displayed in Table 1.

This paper aims to create a parsimonious and straightforward regression model to interpret the several central aspects of energetic expenditure captured by the Age and BMI variables that are expected to behave in a nonlinear way with the response. At the same time, it is of interest to assess the effect of diet on physical exercise. It has been observed that sex and ethnicity differences in the US population tend to interact concerning physical activity. For instance, women tend to be physically less active than men within some ethnic groups (Black, While, Asian, Other races, including Multi-racial) (Troiano et al. 2008), while among the Mexican American and Other Hispanic ethnicities, men and women show very similar physical activity levels (Ortiz-Hernandez and Ramos-Ibanez 2010) and adjusted for demographic factors (Medina et al. 2013). Hence, an

interaction between sex and ethnicity was included to obtain reliable population-based conclusions about the relationships between these covariates and physical activity. The inherent sampling design of the NHANES provides essential advantages to obtaining reliable population measurements that cannot be guaranteed with observational cohorts such as the UK-Biobank due to selection bias. In order to properly exploit this advantage, however, the survey design must be taken into account in the estimation procedure, as described in Section 3.1.

3. THE PARTIALLY LINEAR FRÉCHET SINGLE INDEX REGRESSION MODEL

Let Y_i be the empirical quantile function of daily activity levels corresponding to the ith participant. In what follows, the regression relationship is built by directly modeling the pointwise mean function of $Y_i(t)$ on the covariates, $t \in [0, 1]$. Using the quantile function to characterize the physical activity distribution, as opposed to some other representation, can be justified as follows. First, a density or cumulative distribution function (cdf) representation that ignores inactivity time is inappropriate because the distributions represented by the Y_i are a mixture of a mass at zero and a continuous distribution for positive values. There are other practical reasons to prefer quantile functions. For instance, quantile functions are less constrained than cdfs or density functions. For example, one may add two quantile functions or multiply one by any positive constant to produce another, but not so for cdfs or densities. This is crucial for predictive modeling, since applying a post hoc adjustment to a model prediction to obtain a valid distributional representation can affect both interpretation and introduce distortions that are more prominent for cdfs or densities than for quantile functions (Petersen et al. 2021). Second, modeling the mean quantile function is directly related to *optimal transport* through the *Wasserstein metric*, offering a natural framework for understanding biological phenomena (Villani 2009; Zhang and Müller 2011; Panaretos and Zemel 2019; Peyré et al. 2019).

Briefly, if μ and ν are two suitable measures on $\mathcal{R}$ with finite second moment, and if Q_μ and Q_ν are their corresponding quantile functions, then $d_{W_2}(\mu, \nu)$, the Wasserstein distance between μ and ν, is known to be equivalent to the L^2 distance between Q_μ and Q_ν, that is,

$$d_{W_2}(\mu, \nu) = \left[\int_0^1 (Q_\mu(t) - Q_\nu(t))^2 \mathrm{d}t\right]^{1/2}. \tag{1}$$

As a consequence, under this metric, the Fréchet mean (Fréchet 1948) measure of a random measure is characterized by the pointwise mean of the corresponding random quantile process. Hence, by proposing a regression model for the random quantile function Y_i, one is implicitly constructing a model for the conditional (Wasserstein-)Fréchet mean of the underlying random physical activity distribution measure (Petersen et al. 2021).

The partially linear Fréchet single index (PL-FSI) model is formally defined as follows. Let $\boldsymbol{X_i} \in \mathcal{R}^p$ denote the p-dimensional covariate vector in the single index part of the model, while $\boldsymbol{Z_i} \in \mathcal{R}^q$ is the covariate vector considered for the linear part. The PL-FSI model is

$$E(Y_i(t)|\boldsymbol{X}_i, \boldsymbol{Z}_i) = \alpha(t) + \boldsymbol{\beta}(t)^T \boldsymbol{Z}_i + g(\boldsymbol{\theta}_0^T \boldsymbol{X}_i, t), \quad t \in [0, 1], \tag{2}$$

where the vector $\boldsymbol{\theta_0} \in \mathcal{R}^p$, intercept function α, coefficient function $\boldsymbol{\beta}$ and link function g are the unknown parameters.

3.1 Model estimation

For estimating the parameter $\boldsymbol{\theta}_0$ and the identifiability of the PL-FSI model (Lin and Kulasekera 2007), define the parameter space

$$\Theta_p = \{\boldsymbol{\theta} \in \mathcal{R}^p : \|\boldsymbol{\theta}\|_E = 1, \text{first non-zero element being strictly positive}\}$$

where $\|\cdot\|_E$ is the Euclidean norm. To facilitate estimation of the smooth bivariate function g, the expansion

$$g(u,t) \approx \sum_{k=1}^{K+s} \gamma_k(t)\phi_k(u) \tag{3}$$

will be used, where $\{\phi_k\}_{k=1}^{K+s}$ is a B-spline basis of order s on K interior knots, and $\gamma_k(t)$ are the coefficients of the basis as a function of t.

In practice, the tuning parameter K can be chosen by evaluating candidate values using a goodness-of-fit measure, such as the coefficient of determination R^2 or related measures, while $s = 4$ is usually chosen, corresponding to cubic splines. The approximation to the PL-FSI model (2) can thus be written as

$$E(Y_i(t)|\boldsymbol{X}_i, \boldsymbol{Z}_i) \approx \alpha(t) + \boldsymbol{\beta}(t)^T\boldsymbol{Z}_i + \boldsymbol{\gamma}(t)^T\boldsymbol{U}_i(\boldsymbol{\theta}_0), \quad t \in [0,1], \tag{4}$$

where $\boldsymbol{\gamma}(t) = (\gamma_1(t), \ldots, \gamma_{K+s}(t))^T$ and, for any $\boldsymbol{\theta} \in \Theta_p$, $\boldsymbol{U}_i(\boldsymbol{\theta}) = (\phi_1(\boldsymbol{\theta}^T\boldsymbol{X}_i), \ldots, \phi_{K+s}(\boldsymbol{\theta}^T\boldsymbol{X}_i))^T$.

As (4) exhibits a linear form for each fixed value of $\boldsymbol{\theta}$, a semi-parametric least-squares approach can be utilized for estimation. Due to the complex survey design of the NHANES database, a weighted least squares criterion is needed in order to perform inference correctly and obtain reliable results (Lumley 2010). Assume that a sample $\mathcal{D} = \{(Y_i, \boldsymbol{X}_i, \boldsymbol{Z}_i) : i \in S\}$ is available, where Y_i is a response variable, and $\boldsymbol{X}_i, \boldsymbol{Z}_i$ are vectors of covariates taking values in a finite-dimensional space. The index set S represents a sample of n units from a finite population.

To account for sampling, each individual $i \in S$ is associated with a positive weight w_i derived from an experimental design such as multistage random sampling. In the particular case of NHANES, the survey weights w_i are specified by the Centers for Disease Control and Prevention (CDC) [1] and are primarily used to mitigate selection bias, as explained in the CDC guidelines, [2] to obtain reliable conclusions about the US population. In the analyses conducted in this paper, these weights were taken to be the inverse of the probability $\pi_i > 0$ of being selected into the sample, ie $w_i = \frac{1}{\pi_i}$ (Kish 1965; Lumley 2004). These weights are used to construct an estimator of Horvitz-Thompson type (Horvitz and Thompson 1952; Rabe-Hesketh and Skrondal 2006) by constructing a weighted least squares criterion.

The full procedure can be broken down into two steps. In the first step, for any $\boldsymbol{\theta} \in \Theta_p$ and any $t \in [0,1]$, one computes weighted least squares estimates

$$\left(\hat{\alpha}_{\boldsymbol{\theta}}(t), \hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}(t), \hat{\boldsymbol{\gamma}}_{\boldsymbol{\theta}}(t)\right) = \underset{a\in\mathcal{R}, \boldsymbol{b}\in\mathcal{R}^q, \boldsymbol{c}\in\mathcal{R}^{K+s}}{\operatorname{argmin}} \sum_{i=1}^{n} w_i \left[Y_i(t) - a - \boldsymbol{b}^T\boldsymbol{z}_i - \boldsymbol{c}^T\boldsymbol{U}_i(\boldsymbol{\theta})\right]^2. \tag{5}$$

These estimates lead to initial fitted quantile functions

$$Y_i^*(\boldsymbol{\theta}, t) = \hat{\alpha}_{\boldsymbol{\theta}}(t) + \hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}^T(t)\boldsymbol{Z}_i + \hat{\boldsymbol{\gamma}}_{\boldsymbol{\theta}}^T(t)\boldsymbol{U}_i(\boldsymbol{\theta}), \quad t \in [0,1]. \tag{6}$$

However, as a function of t, $Y_i^*(\boldsymbol{\theta}, t)$ may not be monotonically increasing and hence is not a proper quantile function. The typical solution for this is to project $Y_i^*(\boldsymbol{\theta}, t)$, in the $L^2[0,1]$ sense, onto the nearest monotonic function (Petersen and Müller 2019; Petersen et al. 2021), yielding a valid quantile function $\hat{Y}_i(\boldsymbol{\theta}, t)$. For more details, the reader is referred to Algorithm 2 in the Supplementary Material of Petersen et al. (2021). This algorithm produces a non-decreasing

[1] https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Demographics&CycleBeginYear=2011
[2] https://wwwn.cdc.gov/nchs/nhanes/tutorials/weighting.aspx

quantile function because the space of quantile functions is closed and convex, ensuring that the projection is unique. The algorithm used is a quadratic program solver that effectively adjusts the estimated quantile function to enforce monotonicity, thereby guaranteeing that the resulting function is non-decreasing.

Once these initial quantities are formed for any $\boldsymbol{\theta}$ and t, an estimate $\hat{\boldsymbol{\theta}}_0$ can be computed. As justified in Ghosal et al. (2023a), one can use a generalized version of the residual sums of squares to obtain the estimate. In the current context, the survey-weighted criterion

$$W_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} w_i \int_0^1 \left\{ Y_i(t) - \hat{Y}_i(\boldsymbol{\theta}, t) \right\}^2 dt \tag{7}$$

is proposed, and constitutes a weighted average of the squared L^2 norms of the quantile residuals (or, equivalently, of the squared Wasserstein distances between observed and fitted physical activity distributions). Then the estimated parameter is

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta_p}{\operatorname{argmin}}\, W_n(\boldsymbol{\theta}). \tag{8}$$

From this estimate of the index parameter, given any covariate pair $(\boldsymbol{z}, \boldsymbol{x})$, the conditional Wasserstein-Fréchet mean quantile function can be estimated as follows. First, the basis functions are evaluated at the relevant input by computing $\hat{\boldsymbol{u}} = (\phi_1(\hat{\boldsymbol{\theta}}^T \boldsymbol{x}), \ldots, \phi_{K+s}(\hat{\boldsymbol{\theta}}^T \boldsymbol{x}))^T$. Then, (5) is computed at the specified $\hat{\boldsymbol{\theta}}$ and, as in (6), the preliminary estimate

$$Y^*(t; \boldsymbol{z}, \boldsymbol{x}) = \hat{\alpha}_{\hat{\boldsymbol{\theta}}}(t) + \hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\theta}}}^T(t)\boldsymbol{z} + \hat{\boldsymbol{\gamma}}_{\hat{\boldsymbol{\theta}}}^T(t)\hat{\boldsymbol{u}}. \tag{9}$$

is constructed. Finally, the estimated quantile function $\hat{Y}(t; \boldsymbol{z}, \boldsymbol{x})$ is obtained by projecting (if necessary), in the L^2 sense, $Y^*(t; \boldsymbol{z}, \boldsymbol{x})$ onto the space of quantile functions, meaning the nearest monotonically increasing function. In particular, for any set of observed covariates $(\boldsymbol{Z_i}, \boldsymbol{X_i})$, fitted values $\hat{Y}_i(t) = \hat{Y}(t; \boldsymbol{Z_i}, \boldsymbol{X_i})$ are obtained.

Importantly, the proposed estimation method differs from other additive functional regression models in the literature, such as those presented by McLean et al. (2014), which represent quantile functions using a basis. A basis representation for the quantile response objects is not required for the several reasons: first, it is not necessary for implementation of the proposed estimation algorithm, and its use would not simplify computation; second, it does not provide additional insights into the biological problem or aid in interpretation; and, third, the zero pattern in quantile functions does not lend itself well to such representations in general, except for special bases like B-splines. Furthermore, another alternative used in the nonparametric regression literature is to employ kernel smoothing to estimate the nonparametric part. However, local smoothers are more difficult to incorporate with linear effect estimation compared to global smoothers like splines.

3.2 Model inference: global confidence bands via survey boostraping

In order to quantify the uncertainty associated with the estimates $\hat{\beta}(t)$ in the linear component of the model, we employed a survey bootstrap methodology designed to derive global confidence bands for each linear predictor. Global confidence bands for functional predictors provide a more robust statistical interpretation by overcoming the limitations inherent in pointwise confidence bands, which are affected by multiple comparisons.

Our bootstrap resampling procedure explicitly respects the hierarchical structure inherent to the NHANES survey data. Instead of resampling individual activity count pairs (m_{ij}, A_{ij}), the bootstrap procedure involves resampling entire participant-level data from the database. This strategy preserves the correlation structure between activity counts observed from the same participant

and accurately reflects the variability at the participant level. Since the number of physical activity measurements per participant is not linked to the survey sampling design, a two-step bootstrap approach involving separate resampling of measurements within individuals would be inappropriate. Hence, the correct approach is to select individuals, with replacement, along with all associated activity counts, effectively maintaining the dependence structure among multiple measurements within each participant.

Specifically, we adopted the survey bootstrap methodology developed for multistage designs described in Rust and Rao (1996). This methodology employs a multiplier bootstrap approach. For each participant $i \in \{1, \ldots, n\}$ and each bootstrap replicate $b \in \{1, \ldots, B_s\}$, a positive multiplier weight $W_{i,b}$ is generated. These random multipliers depend on tuning parameters such as the number of observations selected from each stratum (based on geographical Primary Sampling Units or PSUs) and the bootstrap sampling rate. The technical details and closed-form expressions for the multiplier calculations in specific cases are provided by Rust and Rao (1996). The bootstrap survey weights are then defined as

$$w_i^{(b)} = W_{i,b} w_i,$$

where w_i are the original NHANES survey weights. The resulting bootstrap weights $w_i^{(b)}$ mimic the distributional behavior of the original survey weights w_i, thereby providing valid uncertainty quantification under complex survey designs. For different survey scenarios, appropriate bootstrap weights can be computed using the `surveybootstrap` package in `R`.

Let $\hat{\beta}_r^{(b)}(t)$ be the estimate of the rth regression parameter $\beta_r(t)$, $r \in \{1, \ldots, q\}$, for the bth simulation and the tth quantile, using the same estimation process as described in Section 3.1, replacing the survey weights w_i with $w_i^{(b)}$. Define the pointwise boostrap means and variances as

$$\bar{\hat{\beta}}_r(t) = \frac{1}{B_s} \sum_{b=1}^{B_s} \hat{\beta}_r^{(b)}(t), \quad s^2_{r,\hat{\beta}}(t) = \frac{1}{B_s - 1} \sum_{b=1}^{B_s} \left(\hat{\beta}_r^{(b)}(t) - \bar{\hat{\beta}}_r(t) \right)^2,$$

then calculate

$$u_{r,b} = \sup_{t \in (0,1)} \frac{\left| \hat{\beta}_r^{(b)}(t) - \hat{\beta}_r(t) \right|}{s_{r,\hat{\beta}}(t)}, \quad b \in \{1, 2, \ldots, B_s\}; \ r \in \{1, 2, \ldots, q\}.$$

The upper 95% quantile of $\{u_{r,1}, u_{r,2}, \ldots, u_{r,B_s}\}$ is denoted by $q^{(r)}_{0.05}$. Hence, the 95% bootstrap confidence interval for the parameter $\beta_r(t)$ is

$$\left(\hat{\beta}_r(t) - q^{(r)}_{0.05} s_{r,\hat{\beta}}(t), \quad \hat{\beta}_r(t) + q^{(r)}_{0.05} s_{r,\hat{\beta}}(t) \right).$$

3.3 Computational details

Details regarding the implementation of the proposed PL-FSI model and its estimation for the NHANES database will now be provided. In the models implemented below in Section 4, the nonlinear covariate $\boldsymbol{X_i}$ for the ith individual consists of their BMI and Age, so the dimension for this component is $p = 2$, ie $\boldsymbol{\theta}_0 \in \Theta_2$. For the spline basis in (3), computations were performed using the `dbs` function in the package `splines2` (Wang and Yan 2021). Knot placement was determined internally by the default option of the `dbs` function and varied with the value of $\boldsymbol{\theta}$. Specifically, for any K, equally spaced values r_k, $k = 1, \ldots, K$, were computed, where $r_0 = 0 < r_1 < \cdots < r_K < r_{K+1} = 1$; the kth interior knot was then taken as the r_kth empirical quantile of the values $\{\boldsymbol{\theta}^T X_i;\ i = 1, \ldots, n\}$. In the experiments conducted, $s = 4$ and $K = 5$ were used, so the

number of spline regression parameters was $K + s = 9$. The choice $K = 5$ was made based on a preliminary investigation of the PL-FSI model fit for potential values $K = 4, \ldots, 10$. The adjusted Fréchet R^2 criterion, defined below in Section 4, showed a jump from $K = 4$ to $K = 5$, followed by stable values for larger K. The covariates in the linear component $\boldsymbol{Z}_i$ consist of HEI (continuous) and indicator variables for Sex and Ethnicity, as well as the interaction between these categorical variables.

The estimates of parameters in (5) can be efficiently computed as a weighted least squares problem for any fixed $\boldsymbol{\theta}$ and $t \in [0, 1]$. However, this can only be done in practice for a finite ordered grid of values $t \in T_m = \{t_1, \ldots, t_m\} \subset [0, 1]$. These initial survey-weighted least squares computations were done using R package `survey` (Lumley 2004, 2010, 2020), which allows for the introduction of splines into the regression model while simultaneously computing and incorporating the survey weights w_i.

For any given $\boldsymbol{\theta}$ and grid point t, computation of (6) is straightforward. To execute the projection step, observe that monotonicity can only be achieved in the discrete sense in dependence on the chosen grid T_m. We refer to Petersen et al. (2021) for a simple description of this projection algorithm, which can be done using any basic quadratic program solver. Consequently, for a given $\boldsymbol{\theta}$, (7) is approximated by numerical integration. Finally, to perform the optimization in (8), the function `optim` in R was used with the L-BFGS-B algorithm by repeatedly performing the above steps to evaluate $W_n(\boldsymbol{\theta})$ for different values of $\boldsymbol{\theta}$ across iterations. To deal with the possibility of local minima, four different starting values (taken to be equally spaced in their angular representation) in Θ_2 were used for this optimization step, yielding four (possibly not unique) candidate estimators at convergence. The final estimator was taken to be the candidate yielding the smallest value of W_n.

4. EXPERIMENTAL RESULTS

This section explores the PL-FSI model when it is fitted to the NHANES database, with physical activity distributions as the response functions. First, the PL-FSI model is compared to two competing models, the global Fréchet regression model and an alternative semi-parametric Fréchet regression model. This latter model replaces the single index term in (2) with two separate additive terms $g_1(X_{i1}) + g_2(X_{i2})$, and is termed a Partially Linear (Additive) Fréchet (PLF) model. The PLF model is a natural candidate for including nonlinear effects for Age and BMI. While it does not impose a single index structure, it also does not account for any interaction between these two covariates, which is a critical biological feature. The PLF model was similarly fitted using two separate cubic spline terms, each with $K = 5$ knots.

After these initial model comparisons, the fitted linear and nonlinear components of the PL-FSI model are interpreted in turn in order to identify the implied associations with the various covariates. Beginning with the predictors in the linear component of the model, differences between diverse subpopulations corresponding to ethnicity and sex are investigated. From there, the combined associations of BMI and age are elucidated through the spline fit of the nonlinear model component. Throughout, due to the use of the quantile representation of physical activity, comparisons are made across the range of physical activity intensities, rather than at the mean intensity or some pre-specified selection of quantiles as is the common practice.

4.1 Comparison of competing models

To provide an even comparison, the global Fréchet (GF) model was slightly modified by introducing the specific survey weights in the estimation criterion. The covariates used were the same in each of the GF, PL-FSI, and PLF models, with the only difference being that the latter two included the BMI and Age in their nonlinear single index and nonlinear additive components, respectively. Hence, the global Fréchet model can be considered as a special case of both the PL-FSI and PLF models, in which all covariates are included in the linear component. To facilitate interpretation, all numerical covariates were centered and scaled prior to fitting the models.

To quantify differences in the quality of model fits, the capacity of the models to explain differences in physical activity distributions across individuals can be evaluated using the survey-weighted Fréchet R^2

$$R^2_{\oplus} = 1 - \frac{\sum_{i=1}^{n} w_i \int_0^1 (Y_i(t) - \hat{Y}_i(t))^2 dt}{\sum_{i=1}^{n} w_i \int_0^1 (Y_i(t) - \overline{Y}(t))^2 dt} \tag{10}$$

where $\overline{Y}(t) = \left(\sum_{i=1}^{n} w_i\right)^{-1} \sum_{i=1}^{n} w_i Y_i(t)$ is the weighted sample Wasserstein-Fréchet mean of the observed physical activity distributions and w_i is the survey weight corresponding to ith observation. Since the weights w_i are positive, $\overline{Y}$ is a convex combination of the individual quantile functions Y_i, ensuring that it is non-decreasing and thus a valid quantile function.

Because the models have differing levels of complexity even with the same set of predictors, $R^2_{\oplus}$ does not provide a fair comparison. Hence, the adjusted Fréchet R^2, denoted as $\bar{R}^2_{\oplus}$, was also computed (Petersen and Müller 2019). With n being the number of observations and q' the number of regression parameters included in the model being considered,

$$\bar{R}^2_{\oplus} = R^2_{\oplus} - \left(1 - R^2_{\oplus}\right) \frac{q'}{n - q' - 1} \tag{11}$$

enables a fair comparison of the quality of model fit that adjusts for the different complexity of each model.

The PLF and PL-FSI models had $\bar{R}^2_{\oplus}$ values of 0.147 and 0.146, respectively, which are roughly 24% higher relative to the 0.118 value obtained by the global Fréchet model. This suggests that, although the predictive capacity all three models is moderate, the additional parameters introduced by the spline representation of the smooth terms in the partially linear models improved the variance explained. The overall moderate statistical associations are not surprising given the multitude of factors that can influence high-dimensional physical activity distributions, many of which are not included in our model or the NHANES database. Nonetheless, we acknowledge this limitation and emphasize the interpretative value of the detected associations in the next sections with functional coefficients despite the moderate R^2 values.

To provide a more reliable predictive evaluation between the three models, cross-validation was conducted using 40 independent data splits (90% training and 10% testing), and the empirical performance across three models was compared in terms of the mean square prediction error on the test sets, quantified as the average squared Wasserstein distance between predicted and observed physical activity quantile functions. Empirically, the GF and PLF models are both outperformed in terms of out-of-sample prediction by the proposed PL-FSI model. Figure 3 visualizes the error comparison of the three models considered, each with different levels of complexity and structural properties. The PL-FSI model strikes a balance between the restrictive global Fréchet regression model and the more flexible PLF model, while also accounting for the age-BMI interaction through the single index—a feature that neither of the other models accommodates.

4.2 Interpretation of the PL-FSI model fit

A salient advantage of the PL-FSI model is its diverse model components that incorporate linear, categorical, and nonlinear effects. These components are harmonized into the final regression model through an additive regression structure. It is important to stress that, while the model is accurately described as linear in the way that it models covariate associations at each $t \in [0, 1]$, the nature of the model is entirely nonparametric in its treatment of the variations across t of both the coefficient functions and, by extension, the fitted quantile functions of physical activity.

A important practical issue arises from the fact that, due to the final projection step in obtaining fitted quantile functions, the interpretability of coefficient function estimates from the linear component could be questionable. In the setting of global Fréchet regression, Lemma 2 in Petersen

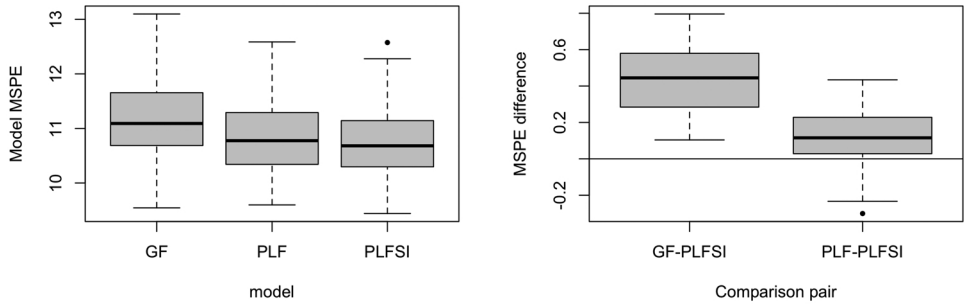

Fig. 3. (Left) Out-of-sample mean square prediction errors across 40 independent data splits for the Global Fréchet (GF), Partially Linear Fréchet (PLF), and Partially Linear Fréchet Single Index (PL-FSI) models. (Right) Differences between out-of-sample mean square prediction errors of GF and PLF models, respectively, relative to the PL-FSI model, with horizontal solid line at 0 for reference.

et al. (2021) demonstrated that, as n grows to infinity, the probability that the projection step is necessary for computing a prediction for any of the n data points converges to zero. Given that the analyzed cohort consists of over 4,000 individuals, the effect of projection is expected to be minor. This is verified empirically in the Supplementary Material.

In Sections 4.2.1 to 4.2.3, the estimated regression parameters in (9) will be interpreted and discussed. Due to the boundary effects induced by knot placement in the spline representation, as well as the inherently noisy nature of the observations of each physical activity profile in the right tail, these parameters will be displayed and interpreted over the restricted interval $t \in [0, 0.98]$. For the linear components $\hat{\alpha}_{\hat{\theta}}(t)$ and $\hat{\boldsymbol{\beta}}_{\hat{\theta}}(t)$, the interpretations are aided by the use global confidence bands using the survey boostrap methodology described in Section 3.2.

4.2.1. *Linear effect of healthy eating index*

The Healthy Eating Index (HEI) reflects the diet quality of the participants. The outcomes of this analysis provide the statistical association of the diet patterns with physical activity patterns, with the corresponding coefficient function estimate being plotted in Fig. 4. For $t < 0.25$, the estimated coefficient takes the constant value zero, reflecting the inactive portion of each participant's physical activity distribution. Even beyond the inactive region of quantile levels, HEI does not exhibit strong relevance for low to moderately high physical activity intensities ($0.25 < t < 0.90$). However, for individuals with very high physical activity intensities (quantiles $t > 0.90$), the plot suggests that healthier diet quality (higher HEI), is associated, on average, with greater physical activity intensities, after accounting for the other covariates in the model. These findings underscore the intricate relationship between diet quality, lifestyles (Patterson et al. 1994; Leroux et al. 2015) and physical activity (Scarmeas et al. 2009), especially in contexts where a prevalence of individuals actively engage in rigorous exercise regimens, such as high-intensity cardio and/or resistance training and/or physically-demanding manual labor. An interesting point is that establishing directionality, rather than merely interpreting statistical associations, would be ideal. However, to rigorously achieve this, we would require a mediation model (Feng et al. 2021) within metric spaces using the 2-Wasserstein distance.

4.2.2. *Associations with categorical predictors: ethnicity and sex*

The inclusion of categorical covariates as predictors of physical activity patterns is critical, as this allows for the identification of potential disparities in physical activity across various subgroups of the American population, opening the possibility for the development of targeted clinical interventions. The subsequent interpretations of the PL-FSI model fits are designed to address

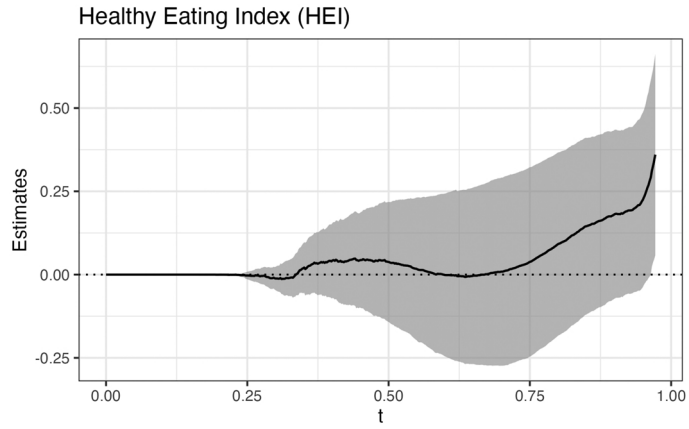

Fig. 4. The estimated functional coefficient for the covariate HEI in the PL-FSI model (4) is shown here as the solid black line and its 95% global confidence band is indicated by the grey shaded region. The dotted black line at 0 is for reference.

the following inquiries regarding significant epidemiological and public health questions related to differences across sexes and ethnicities.

1. Are there variations in physical activity levels between men and women, and how do these variations differ among the represented ethnicities?
2. Do women of different ethnicities exhibit differences in physical activity levels, and if so, how do these differences vary?
3. Do men of different ethnicities demonstrate disparities in physical activity levels, and if so, how do these differences vary?

To address question 1, Fig. 5 illustrates the estimated coefficient functions for male participants subtracted from those for female participants within each ethnic group. The estimated differences plotted as functions of t for each ethnicity, along with their 95% global confidence intervals, reveal that, for quantile levels t around 0.30 to 0.98, men exhibit significantly higher physical activity levels, on average, than women among White, Black, Asian, and Other Races, including Multi-Racial ethnicities. However, in Mexican American and Other Hispanic groups, men and women display statistically similar levels of physical activity, on average.

Addressing question 2 above, Fig. 6 illustrates pairwise differences in estimated model intercepts for females across different ethnicities, along with 95% pointwise confidence intervals. These comparisons condition on fixed Age, HEI and BMI. Over the range of quantile levels t from 0.30 to 0.98, Mexican American and Other Hispanic women, on average, showcase higher physical activity levels compared to certain other ethnicities. Women of White, Black, Asian, and Other Races, including Multi-Racial backgrounds, exhibit statistically similar physical activity levels, on average. Similarly, Mexican American and Other Hispanic women, on average, display comparable levels of physical activity.

Correspondingly, to address the question 3, Fig. 7 presents results akin to Fig. 6, focusing on male participants. These comparisons condition on fixed Age, HEI and BMI. The estimates indicate that, for low to high physical activity levels ($t \in [0.30, 0.98]$), males identified as Mexican American or Other Hispanic ethnicities are more physically active, on average, compared to males of Black, White, Asian, and Other Races (including Multi-Racial backgrounds). While there is scientific evidence highlighting disparities in physical activity levels among different ages, sexes, and ethnicities (Caspersen et al. 2000; Ji et al. 2024) in U.S. populations, to the best of the authors' knowledge, this is the first time that these disparities are addressed across the full range of

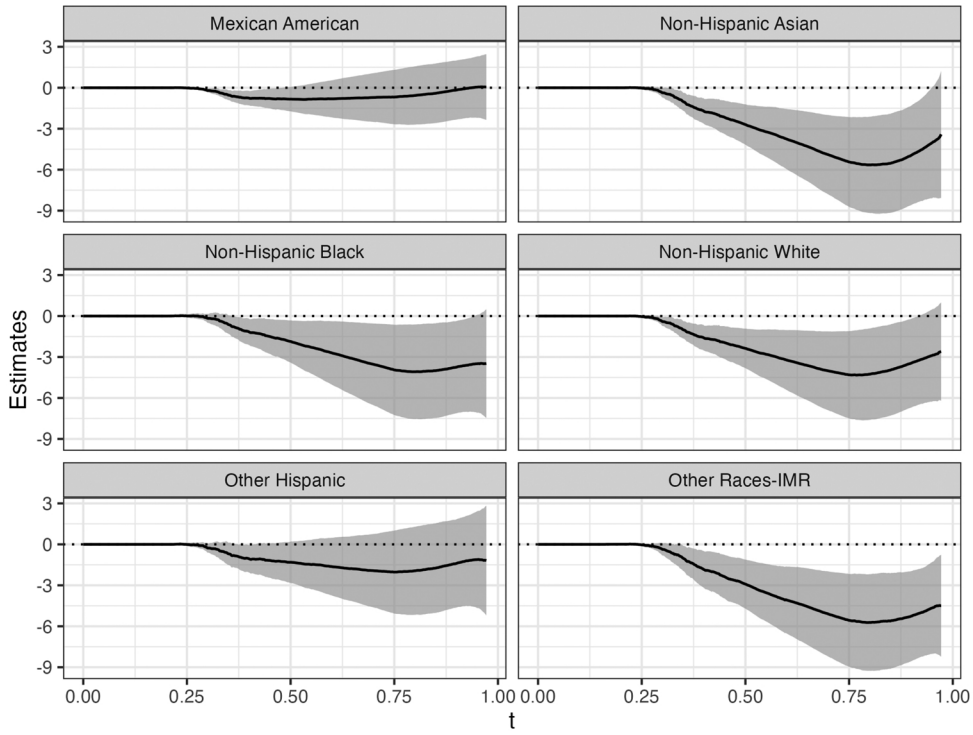

Fig. 5. The model intercepts for the PL-FSI model (4), are considered for male and female participants of different ethnic backgrounds. The intercepts for the males are subtracted from the intercepts of the females for each ethnicity, considering the numeric variables HEI, Age and BMI as fixed. The respective estimated parameter combinations are computed along with their 95% global Confidence Intervals, and plotted as solid black lines and grey shaded regions respectively. The dotted black line at 0 is for reference. The differences are considered for the ethnicities: Mexican American, Other Hispanic, Non-Hispanic White, Non-Hispanic Black, Non-Hispanic Asian, Other races including Multi-Racial.

accelerometer intensities using a distributional representation approach. The model fit also suggests that men of White and Black ethnicities exhibit slightly higher physical activity levels than Asian men for certain quantile values. Men of Other Races, including Multi-Racial individuals, show slightly lower activity levels, on average, compared to Black and White men, but statistically similar levels to Asian men.

From a public health perspective, this underscores the need for tailored interventions to promote physical activity based on accelerometer data, considering variations among sexes and ethnicities. Such targeted strategies are pivotal in enhancing public health outcomes and addressing disparities. Public health policies aimed at promoting health should not be uniform across all groups, differing, for instance, between Asian men and Asian women or across ethnicities like Mexican American and Asian men, when conditioning on the same HEI, Age and BMI values. It is crucial to note that conducting sex-stratified analyses enhances the reliability of the presented findings as they reveal important interactions between these categorical variables when associated with physical activity intensities, when HEI, Age and BMI are the same.

4.2.3. Nonlinear associations with age and BMI

Due to the semiparametric model structure, the interpretation of estimated associations between physical activity levels and the single index composed of Age and BMI requires some care. As already

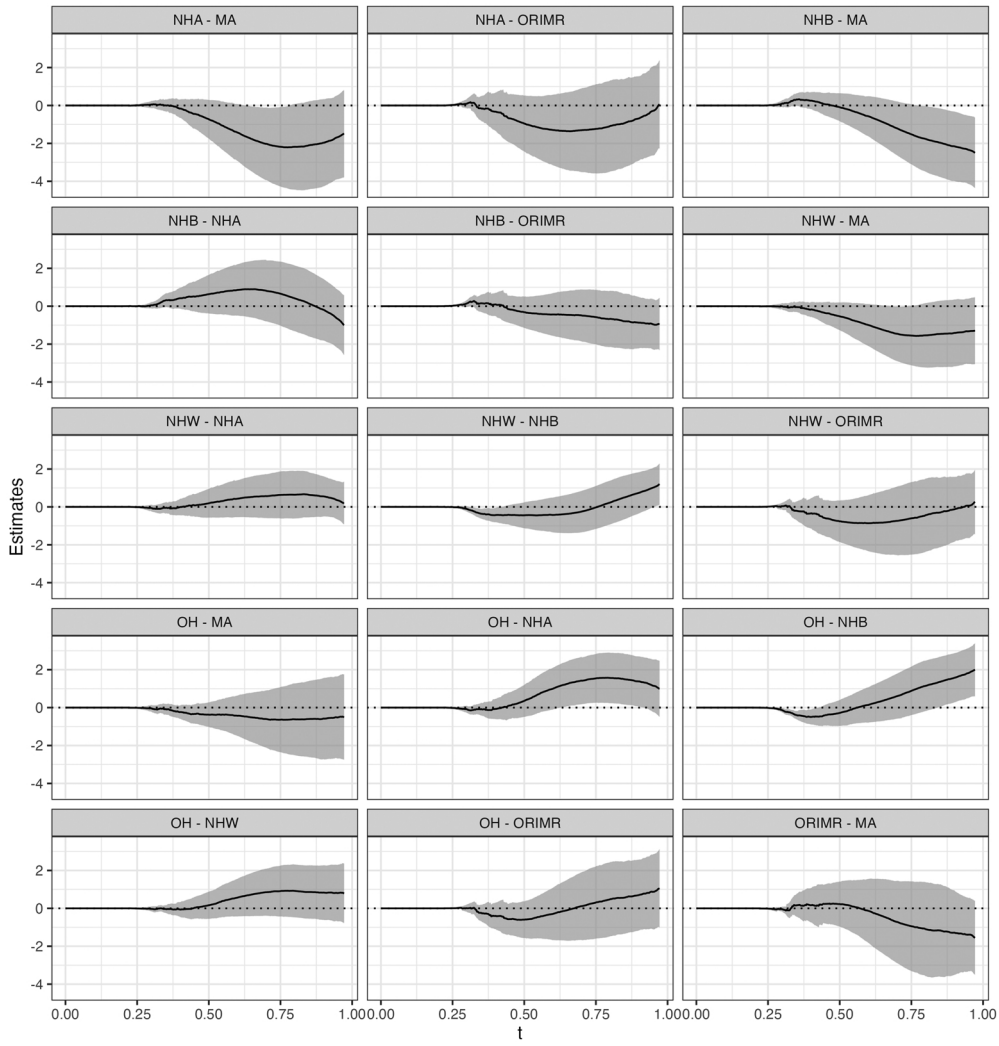

Fig. 6. The estimated model intercepts for the PL-FSI model are compared for females of different ethnic backgrounds. The pairwise differences of such intercepts are displayed, along with their 95% Confidence Intervals, as solid black lines and grey shaded regions respectively. The dotted red line at 0 is for reference. The title for each panel indicates the order of the differences of the intercepts. The abbreviations for the ethnicities are, OH: Other Hispanic, MA: Mexican American, NHW: Non-Hispanic White, NHB: Non-Hispanic Black, NHA: Non-Hispanic Asian, and ORIMR: Other Races Including Multi-Racial. As an example, the title "OH—MA" indicates that the estimated intercepts for females identifying as Mexican American were subtracted from the estimated intercepts for females identifying as Other Hispanic ethnicity.

demonstrated, this increased complexity yields a measurably improved model fit and predictive capacity. In addition, by appropriately examining the estimated model components, an intuitive but nuanced association emerges. To begin, the estimated index parameter was $\hat{\boldsymbol{\theta}} = (0.2661, 0.9639)$ for the standardized variables BMI and Age respectively. The nonparametric function g cannot distinguish between age and BMI since it only accounts for their linear combination as defined by the single index model. However, since both variables have been scaled to have a standard deviation

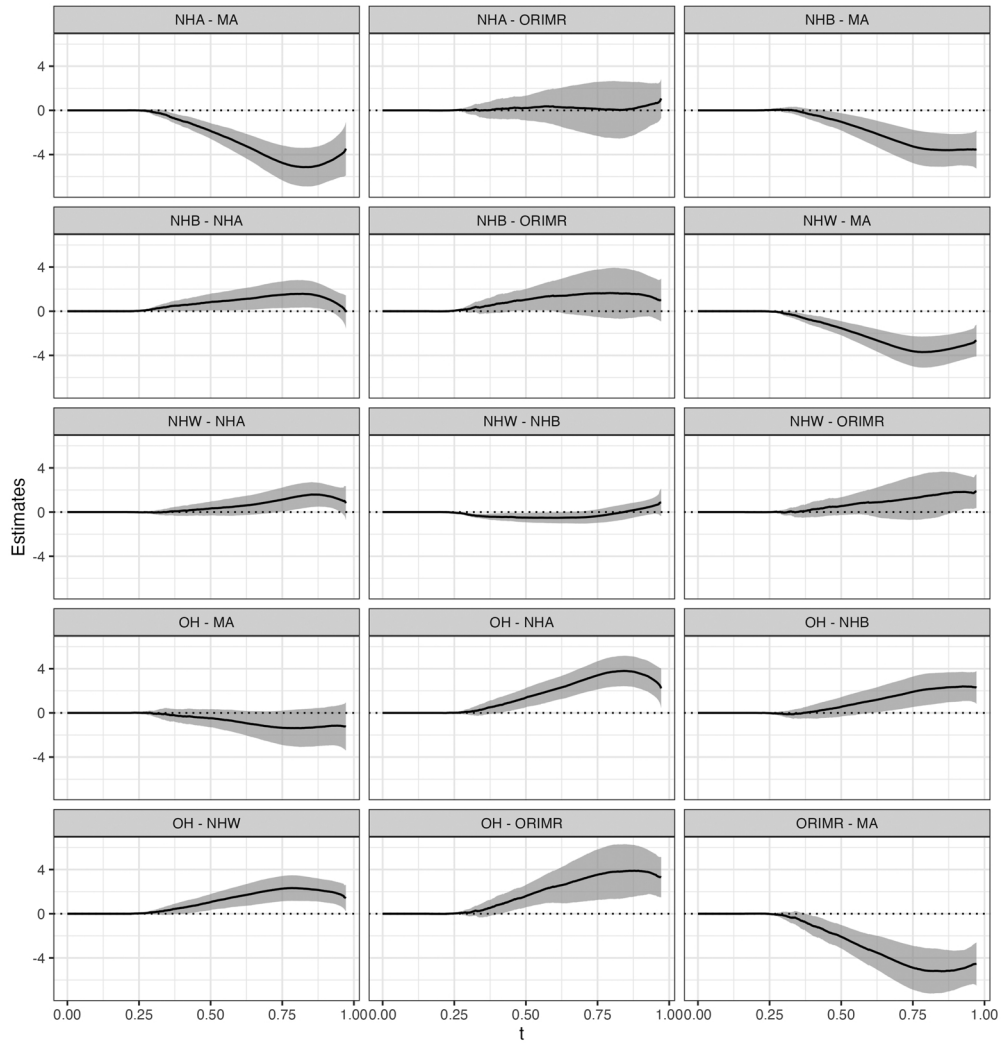

Fig. 7. Estimated model intercepts for the PL-FSI model are compared for males of different ethnic backgrounds, conditional on fixed HEI, Age and BMI. The pairwise differences of such intercepts are computed along with their 95% Confidence Intervals, and plotted here as solid black lines and grey shaded regions, respectively. The dotted black line at 0 is for reference. The abbreviations for ethnicities as well as the order of the differences are the same as in Fig. 6.

of 1, the relative magnitudes of $\hat{\boldsymbol{\theta}}$ do indeed reflect the relative contributions of each variable. In this case, age has a higher coefficient, suggesting a stronger association with the response variable compared to BMI.

As the effect is nonlinear, each of the four panels in Fig. 8 plots the fitted quantile values $Y^*(t, \boldsymbol{z}, \tilde{\boldsymbol{x}})$ for quantile levels $t = 0.50, 0.75, 0.90$, and 0.97; here, the linear covariate $\boldsymbol{z}$ was fixed to represent the reference groups for sex and ethnicity and the median value of HEI, while $\tilde{\boldsymbol{x}}$ represents the standardized value of the Age and BMI combinations present along the horizontal and vertical axes of each panel. The choice of displayed quantile levels, reflects the finding that the single index only exhibits a notable association with physical activity levels near or above the median intensities.

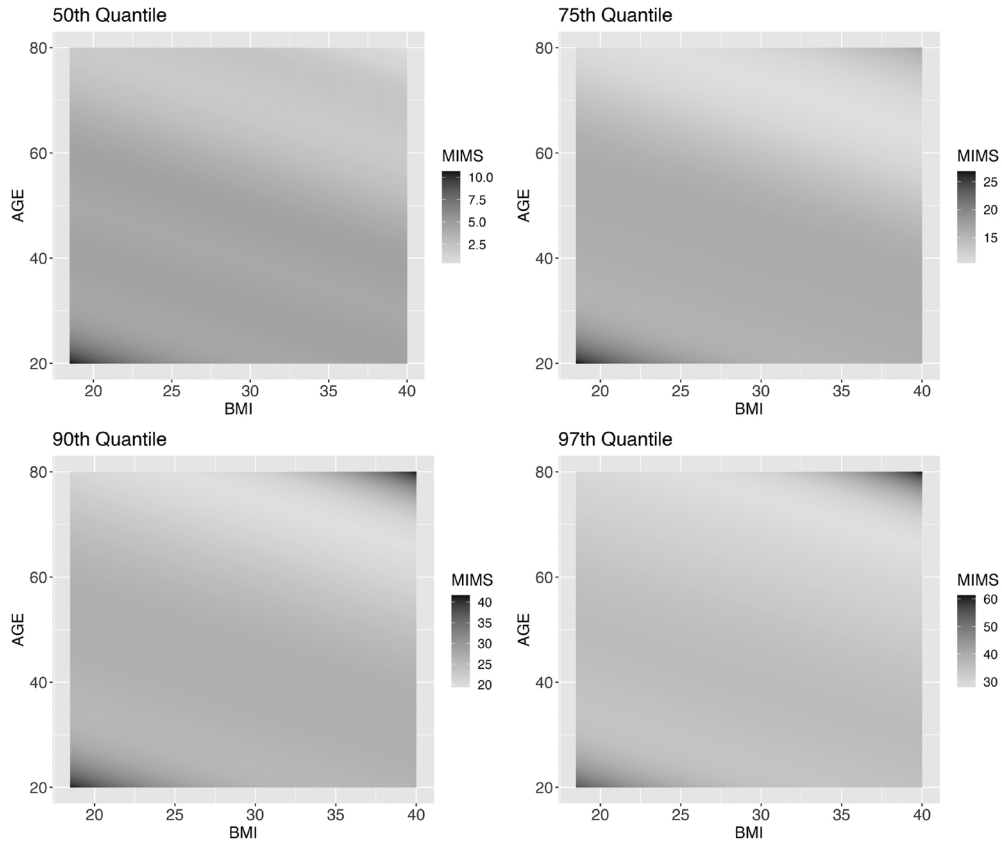

Fig. 8. Heatmap plot of $\hat{Y}(\hat{\boldsymbol{\theta}}, t)$ across different quantiles, $t = 0.50$ (top left), 0.75 (top right), 0.90 (second row left), 0.97 (second row right) respectively. A 2-dimensional grid was considered for the covariates BMI (in range $[18.5,40]$) and Age (in range $[20,80]$) for the single index component of the PL-FSI regression model. The categorical covariates in the linear component were fixed at their baseline levels (ie sex male, ethnicity Mexican American) while the numerical covariate HEI was fixed at the median level. This plot describes the estimated nonlinear interaction in the conditional mean function between age and BMI.

For the lowest BMI group (< 20), people in the age range 55 to 70 are estimated by the model to perform the largest physical activity levels in terms of the median and third quartile ($t = 0.50, 0.75$). However, for the same BMI range, people in the age range 25 to 35 perform the highest physical activity in the extreme right tail corresponding to quantiles $t = 0.90, 0.97$. In each of the panels (or, quantiles) the age range for highest physical activity linearly decreases with increase in BMI. For the highest BMI in our study (≈ 40), the highest average physical activity for quantiles $t = 0.50, 0.75$ is estimated to occur within age range 40 to 55. However, in the quantiles $t = 0.90, 0.97$, the highest physical activity are shown by the BMI range 25 to 30 in the age range 20 to 25. Hence, these panels indicate that the nonlinear association of average physical activity intensity with Age and BMI is more pronounced for larger values of t. This estimated association is more robust for middle age individuals with intermediate BMI values than in the rest of the range. As the age increases and the BMI decreases, individuals are more likely to have lower activity intensities, especially in the extremes of Age and/or BMI.

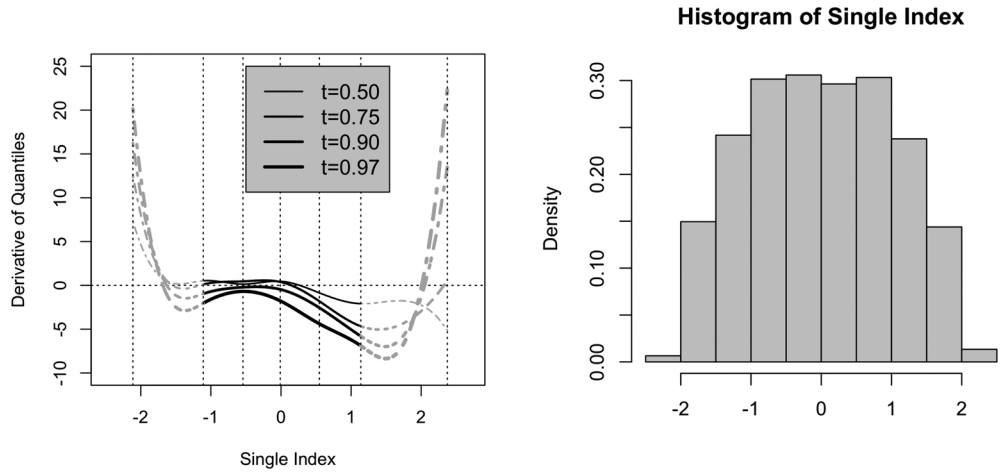

Fig. 9. (Left) Plot of $(\partial\hat{g}/\partial u)(u,t)$, the derivative of the spline-based estimate of the nonlinear component defined in (12), for u across the empirical range of observed index values $\hat{\boldsymbol{\theta}}^T\boldsymbol{X}_i$, $i = 1,\ldots,n$, and for quantile levels $t = 0.50, 0.75, 0.90$, and 0.97 given in increasing thickness of lines respectively. Knot locations are shown as vertical dotted lines, and the derivative curves are depicted as solid black (respectively, dotted grey lines) within (resp., outside of) the interior knot range. (Right) Histogram of observed index values $\hat{\boldsymbol{\theta}}^T\boldsymbol{X}_i$, $i = 1,\ldots,n$. (Right) Histogram of the distribution of the single index projection evaluation for 4,616 included participants in NHANES 2011 to 2014.

As an additional visualization of the nonlinear association with these covariates in the model, consider a direct analysis of the derivative of the nonlinear fit, namely

$$\frac{\partial\hat{g}}{\partial u}(u,t) = \sum_{k=1}^{K+s}\hat{\gamma}_{\hat{\boldsymbol{\theta}},k}(t)\left[\frac{d}{du}\phi_k(u)\right]. \tag{12}$$

The reason to consider the derivative is that, if $\hat{g}$ were a linear function in u, then this derivative could be interpreted in the same way as the linear coefficient estimates in the previous sections. In other words, it is the direct counterpart of the linear coefficient functions for this nonlinear term. The left panel of Fig. 9 displays the behavior of (9) across the relevant range of values u, with respect to the empirical values $\hat{\boldsymbol{\theta}}^T\boldsymbol{X}_i$ (whose distribution is depicted via a histogram in the right panel of the figure) that were used to generate the estimates, for $t = 0.5, 0.75, 0.9, 0.97$. As the observed shape of the curves within the interior knots is most reliable, this region is indicated by the solid portion of each curve in the figure. The various derivative curves suggest that, given the covariates in the linear term, for negative values of the single index containing BMI and Age, there is little to no association with the physical activity quantile response; for positive single index values, the estimated association becomes negative. Since both elements of $\hat{\boldsymbol{\theta}}$ are positive, these findings imply that the model reflects a negative association between physical activity quantiles and BMI/Age when at least one of these is large. Furthermore, the strength of this negative association increases as the quantile level t becomes larger, as evidenced by the increasingly negative derivative estimates in the left panel as t increases. For instance, the derivative for $t = 0.5$ (the median physical activity quantile) is only slightly negative for values of u near zero, whereas it steadily decreases as one examines the curves for $t = 0.75, 0.9, 0.97$. In short, for positive values of $u = \hat{\boldsymbol{\theta}}^T\boldsymbol{x}$, average physical activity quantiles above the median tend to decrease, and the rate of decrease becomes more pronounced for both larger values of u and quantile level t.

5. DISCUSSION

The main contribution of this paper is to propose a new PL-FSI regression model to analyze responses of a distributional functional nature. The new methods have been implemented to analyze the physical activity data from the NHANES database 2011 to 2014, for participants aged 20 to 80 and in BMI range 18.5 to 40. The NHANES survey weights were incorporated into the PL-FSI algorithm using the sampling mechanism of the Horvitz-Thompson type estimator (Kish 1965) to construct a weighted least squares criterion to estimate the model parameters.

The findings from application of this new model are summarized as follows.

1. The discrepancies in physical activity levels between men and women of different ethnicities were examined in the American population. Associations between physical activity and the continuous variables HEI, BMI, and Age were quantified, visualized, and interpreted across the range of human physical activity intensities, as opposed to just the mean or median activity level, due to the new quantile distributional representation of physical activity. For example, it was shown that diet is important only in the high-intensity levels of physical activity range; a better diet, according to the HEI score, is related to more exercise. We also show that the Mexican American and Other Hispanic groups are the most active individuals in the American population for both men and women. An interaction between Age and BMI was discovered and exploited in determining their association with energetic expenditure, valid more specifically in the moderate to higher intensities of physical activity levels.
2. The modeling advantages of the new PL-FSI algorithm over the classical global Fréchet regression model were shown in terms of adjusted Fréchet R-squared and mean square prediction error. In addition, interpretation of the nonlinear term in the PL-FSI model was demonstrated using the gradient of a conditional mean function.

From a practical perspective, these new results illustrate the variation in physical activity across the range of accelerometer intensities, unlike previous models that focus on scalar summaries and averages (Leroux et al. 2019). The results derived from the PL-FSI model, primarily based on demographic variables, show that we can define expected physical activity levels in different U.S. populations. From a public health perspective, this approach can be generalized—for instance, by proposing tolerance regions for physical activity and creating new recommendations about expected physical activity levels, following Matabuena et al. (2024). However, the model in this previous work does not provide interpretable statistical associations as our case, as they are no longer explainable.

From a methodological point of view, we propose the first PL-FSI regression model in the context of object data analysis to bridge the gap between the global Fréchet regression (Petersen and Müller 2019) and the Frechet single index model (Ghosal et al. 2023a), while preserving the interpretability of the predictors and parameter estimates. To the best of our knowledge, this is also the first regression model to incorporate survey data in the context of object data analysis.

The most popular approach to analyzing accelerometer data is through finite dimensional compositional metrics. Here, the functional extension of these metrics (Matabuena and Petersen 2023), was instead used to capture more information about physical activity from an individual by adopting the mathematical framework of the L^2-Wasserstein space. Due to the positive probability at zero physical activity level for each individual (corresponding to periods of inactivity), the quantile function, which is intimately connected with the Wasserstein metric, provides a natural functional representation of such mixed distributions. In addition, the range of values measured by the accelerometer varies widely among individuals and groups, which can present difficulties when trying to apply the standard distributional data analysis methods in this setting (Matabuena and Petersen 2023). For example, functional compositional transformations can be an alternative strategy to creating a regression model about physical activity in a linear space (Van den Boogaart et al. 2014; Hron et al. 2016; Petersen and Müller 2016). However, the distributional physical activity representation arises from a mixed-stochastic process (see Figs 1 and 2 for more details)

that prevents the naive use of the linear functional data methods that typically utilize a basis of smooth functions to represent the functional response and/or the functional parameters in the model, due to the discontinuity of the quantile function in the transition from inactivity to activity in the physical exercise. While specialized basis functions that allow for jumps in the functional parameters or their derivatives could be used, the proposed model and its estimation procedure demonstrate that it is not necessary to do so. As future work, we propose generalizing the distributional variable selection model proposed in Coulter et al. (2024) for survey data, and extending our PL-FSI semi-parametric approach to select the most relevant predictors in settings with a larger pool of variables. Additionally, providing prediction regions could be highly valuable for modeling scientific problems. Leveraging a previous framework for uncertainty quantification in metric spaces (Matabuena et al. 2024), adapted for survey data and applied to the PL-FSI model, can be especially relevant. For instance, defining tolerance regions from a distributional perspective on physical activity could address emerging scientific challenges that are currently focused on scalar variables as step counts in the literature.

Missing data is another significant challenge in wearable data analysis, especially when studying younger populations or during shorter monitoring periods. Although accelerometers generally provide more consistent data quality, smartphone-based tracking of physical activity often introduces greater variability and data gaps, which complicate the reliability of analyses. Addressing these issues necessitates the development of innovative methods and stringent criteria specifically tailored to smartphone-derived data to ensure data integrity. Implementing these approaches is essential for producing accurate and robust results when applying distributional representations of physical activity as both predictor and response variables.

The analysis of complex statistical objects in biomedical science provides an excellent opportunity to create new clinical biomarkers that enrich those available for medical decision-making beyond those commonly used to monitor the health and evolution of diseases. For example, distributional representations are a significant advancement in digital medicine (Javaid et al. 2022) as a digital biomarker (Matabuena et al. 2021; Zhang et al. 2022). However, the generality of techniques introduced also enables the application of the methods developed here to other complex statistical objects such as connectivity graphs, shapes, and directional objects. These methods have potential to introduce new clinical findings in a broad list of clinical situations, for example in neuroimaging and in phylogenetic tree analysis (Yuan et al. 2012; Nye et al. 2017; Relión et al. 2019; Dubey and Müller 2022; Zhou and Müller 2022). Furthermore, with the increasing availability of data from large cohort studies, such as from longitudinal surveys with carefully designed subpopulation sampling weights, the methods provided here will gain more popularity among practitioners. The use of complex statistical objects will undoubtedly enhance daily statistical practice in biomedical applications.

ACKNOWLEDGMENTS

We sincerely thank the two anonymous reviewers and the editors for their insightful and constructive comments, which significantly improved the quality and clarity of our manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Biostatistics Journal* online.

FUNDING

This research was funded in part by National Science Foundation grant DMS-2310943.

CONFLICT OF INTEREST

None declared.

REFERENCES

Battelino T et al 2019. Clinical targets for continuous glucose monitoring data interpretation: recommendations from the international consensus on time in range. Diabetes Care. 42:1593–1603.

Bhattacharjee S, Müller H-G. 2023. Single index frechet regression. Ann Stat. 51:1770–1798.

Bhattacharjee S, Müller H-G. 2025. Geodesic mixed effects models for repeatedly observed/longitudinal random objects. J Am Stat Assoc. 1–23. https://doi.org/10.1080/01621459.2025.2474267

Biagi L et al 2019. Individual categorisation of glucose profiles using compositional data analysis. Stat Methods Med Res. 28:3550–3567.

Carroll RJ, Fan J, Gijbels I, Wand MP. 1997. Generalized partially linear single-index models. J Am Stat Assoc. 92:477–489.

Caspersen CJ, Pereira MA, Curran KM. 2000. Changes in physical activity patterns in the United States, by sex and cross-sectional age. Med Sci Sports Exerc. 32:1601–1609.

Chen H, Müller H-G. 2023. Sliced Wasserstein regression [preprint], arXiv, arXiv:2306.10601.

Chen Y, Lin Z, Müller H-G. 2023. Wasserstein regression. J Am Stat Assoc. 118:869–882.

Coulter A, Aurora RN, Punjabi NM, Gaynanova I. 2024. Fast variable selection for distributional regression with application to continuous glucose monitoring data [preprint], arXiv, arXiv:2403.00922.

Cui EH, Goldfine A, Quinlan M, James DA, Sverdlov O. 2023. Investigating the value of glucodensity analysis of continuous glucose monitoring data in type 1 diabetes: an exploratory analysis. Front Clin Diabetes Healthc. 4:1244613.

Dubey P, Müller H-G. 2019. Fréchet analysis of variance for random objects. Biometrika. 106:803–821.

Dubey P, Müller H-G. 2022. Modeling time-varying random objects and dynamic networks. J Am Stat Assoc. 117:2252–2267.

Fan J, Müller H-G. 2024. Conditional Wasserstein barycenters and interpolation/extrapolation of distributions. IEEE Transactions on Information Theory.

Feng Q et al 2021. The role of body mass index in the association between dietary sodium intake and blood pressure: a mediation analysis with nhanes. Nutr Metab Cardiovasc Dis. 31:3335–3344.

Fréchet M. 1948. Les éléments aléatoires de nature quelconque dans un espace distancié. Annales de l'Institut Henri Poincaré. 10:215–310.

Ghosal A, Meiring W, Petersen A. 2023a. Fréchet single index models for object response regression. Electron J Statist. 17:1074– 1112.

Ghosal R et al 2023b. Distributional data analysis via quantile functions and its application to modeling digital biomarkers of gait in Alzheimer's Disease. Biostatistics. 24:539–561.

Ghosal R, Ghosh SK, Schrack JA, Zipunnikov V. 2025. Distributional outcome regression via quantile functions and its application to modelling continuously monitored heart rate and physical activity. J Am Stat Assoc. 1–13. https://doi.org/10.1080/01621459.2025.2460232

Ghosal R et al 2022. Scalar on time-by-distribution regression and its application for modelling associations between daily-living physical activity and cognitive functions in Alzheimer's Disease. Sci Rep. 12:11558–16.

Hanneke S. 2022. Universally consistent online learning with arbitrarily dependent responses. International Conference on Algorithmic Learning Theory. PMLR. p. 488–497.

Horowitz JL. 2012. Semiparametric methods in econometrics, Vol. 131. Springer Science & Business Media.

Horvitz DG, Thompson DJ. 1952. A generalization of sampling without replacement from a finite universe. J Am Stat Assoc. 47:663–685.

Hron K, Menafoglio A, Templ M, Hrůzová K, Filzmoser P. 2016. Simplicial principal component analysis for density functions in Bayes spaces. Comput Stat Data Anal. 94:330–350.

Javaid A et al 2022. Medicine 2032: the future of cardiovascular disease prevention with machine learning and digital health technology. Am J Prev Cardiol. 12:100379.

Jašková P et al 2023. Compositional functional regression and isotemporal substitution analysis: methods and application in time-use epidemiology. Stat Methods Med Res. 32:2064–2080.

Jeon JM, Lee YK, Mammen E, Park BU. 2022. Locally polynomial Hilbertian additive regression. Bernoulli. 28:2034–2066.

Ji H et al 2024. Sex differences in association of physical activity with all-cause and cardiovascular mortality. J Am Coll Cardiol. 83:783–793.

John D, Tang Q, Albinali F, Intille S. 2019. An open-source monitor-independent movement summary for accelerometer data processing. J Meas Phys Behav. 2:268–281.

Johnson CL, Dohrmann SM, Burt VL, and, Mohadjer LK. 2014. National health and nutrition examination survey: sample design, 2011–2014, Number 2014. US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics.

Katta S, Parikh H, Rudin C, Volfovsky A. 2024. Interpretable causal inference for analyzing wearable, sensor, and distributional data. International Conference on Artificial Intelligence and Statistics. PMLR. p. 3340–3348.

Kish L. 1965. Survey sampling. Wiley.
Kosorok MR, Laber EB. 2019. Precision medicine. Annu Rev Stat Appl. 6:263–286.
Leroux A et al 2019. Organizing and analyzing the activity data in NHANES. Stat Biosci. 11:262–287.
Leroux C et al 2015. In adult patients with type 1 diabetes healthy lifestyle associates with a better cardiometabolic profile. Nutr Metab Cardiovasc Dis. 25:444–451.
Li X et al 2017. Digital health: tracking physiomes and activity using wearable biosensors reveals useful health-related information. PLOS Biol. 15:e2001402.
Liang H, Liu X, Li R, Tsai C-L. 2010. Estimation and testing for partially linear single-index models. Ann Stat. 38:3811–3836.
Lin W, Kulasekera KB. 2007. Identifiability of single-index models and additive-index models. Biometrika. 94:496–501.
Lin Z, Kong D, Wang L. 2023a. Causal inference on distribution functions. J R Stat Soc Ser B Stat Methodol. 85:378–398.
Lin Z, Müller H-G, Park BU. 2023b. Additive models for symmetric positive-definite matrices, riemannian manifolds and lie groups. Biometrika. 110:361–379.
Lugosi G, Matabuena M. 2024. Uncertainty quantification in metric spaces [preprint], arXiv, arXiv:2405.05110.
Lumley T. 2004. Analysis of complex survey samples. J Stat Soft. 9:1–19.
Lumley T. 2010. Complex surveys: a guide to analysis using R. John Wiley and Sons.
Lumley T. 2020. survey: analysis of complex survey samples. R package version 4.0.
Lyons R. 2013. Distance covariance in metric spaces. Ann Probab. 41:3284– 3305.
Matabuena M, Crainiceanu M. 2024. Multilevel functional distributional models with application to continuous glucose monitoring in diabetes clinical trials [preprint], arXiv, arXiv:2403.10514.
Matabuena M, Félix P, Hammouri ZAA, Mota J, del Pozo Cruz B. 2022. Physical activity phenotypes and mortality in older adults: a novel distributional data analysis of accelerometry in the NHANES. Aging Clin Exp Res. 34:3107–3114.
Matabuena M et al 2024. Conformal uncertainty quantification using kernel depth measures in separable hilbert spaces [preprint], arXiv, arXiv:2405.13970.
Matabuena M, Petersen A. 2023. Distributional data analysis of accelerometer data from the NHANES database using nonparametric survey regression models. J R Stat Soc Ser C Appl Stat. 72:294–313.
Matabuena M, Petersen A, Vidal JC, Gude F. 2021. Glucodensities: a new representation of glucose profiles using distributional data analysis. Stat Methods Med Res. 30:1445–1464.
McLean MW, Hooker G, Staicu A-M, Scheipl F, Ruppert D. 2014. Functional generalized additive models. J Comput Graph Stat. 23:249–269.
Medina C, Janssen I, Campos I, Barquera S. 2013. Physical inactivity prevalence and trends among mexican adults: results from the national health and nutrition survey (ensanut) 2006 and 2012. BMC Public Health. 13:1063–10.
Mehta JN, Gupta AV, Raval NG, Raval N, Hasnani N. 2017. Physiological cost index of different body mass index and age of an individual. Natl J Physiol Pharm Pharmacol. 7:1–1317.
Nye TM, Tang X, Weyenberg G, Yoshida R. 2017. Principal component analysis and the locus of the Fréchet mean in the space of phylogenetic trees. Biometrika. 104:901–922.
Onnela J-P. 2021. Opportunities and challenges in the collection and analysis of digital phenotyping data. Neuropsychopharmacology. 46:45–54.
Ortiz-Hernandez L, Ramos-Ibanez N. 2010. Sociodemographic factors associated with physical activity in Mexican adults. Public Health Nutr. 13:1131–1138.
Panaretos VM, Zemel Y. 2019. Statistical aspects of Wasserstein distances. Annu Rev Stat Appl. 6:405–431.
Park J, Kok N, Gaynanova I. 2025. Beyond fixed thresholds: optimizing summaries of wearable device data via piecewise linearization of quantile functions [preprint], arXiv, arXiv:2501.11777.
Patterson RE, Haines PS, Popkin BM. 1994. Health lifestyle patterns of us adults. Prev Med. 23:453–460.
Petersen A, Liu X, Divani AA. 2021. Wasserstein *F*-tests and confidence bands for the Fréchet regression of density response curves. Ann Statist. 49:590–611.
Petersen A, Müller H-G. 2016. Functional data analysis for density functions by transformation to a Hilbert space. Ann Statist. 44:183–218.
Petersen A, Müller H-G. 2019. Fréchet regression for random objects with Euclidean predictors. Ann Statist. 47:691–719.
Petersen A, Zhang C, Kokoszka P. 2022. Modeling probability density functions as data objects. Econometrics Stat. 21:159–178.
Peyré G, Cuturi M. 2019. Computational optimal transport: with applications to data science. Found Machine Learn. 11:355–607.
Rabe-Hesketh S, Skrondal A. 2006. Multilevel modelling of complex survey data. J R Stat Soc Ser A (Stat Soc). 169:805–827.

Relión JD, Arroyo Kessler D, Levina E, Taylor SF. 2019. Network classification with applications to brain connectomics. Ann Appl Stat. 13:1648–1677.
Rust KF, Rao J. 1996. Variance estimation for complex surveys using replication techniques. Stat Methods Med Res. 5:283–310.
Scarmeas N et al 2009. Physical activity, diet, and risk of Alzheimer disease. JAMA. 302:627–637.
Schrack JA, Simonsick EM, Chaves PHM, Ferrucci L. 2012. The role of energetic cost in the age-related slowing of gait speed. J Am Geriatr Soc. 60:1811–1816.
Smirnova E et al 2020. The predictive performance of objective measures of physical activity derived from accelerometry data for 5-year all-cause mortality in older adults: National Health and Nutritional Examination Survey 2003–2006. J Gerontol Ser A. 75:1779–1785.
Topol EJ. 2019. A decade of digital medicine innovation. Sci Transl Med. 11. https://www.science.org/doi/abs/10.1126/scitranslmed.aaw7610
Troiano RP et al 2008. Physical activity in the United States measured by accelerometer. Med Sci Sports Exerc. 40:181–188.
Tucker DC. 2022. Modeling non-Euclidean data via Fréchet regression [Ph.D. thesis]. University of Illinois at Chicago.
Tucker DC, Wu Y, Müller H-G. 2023. Variable selection for global Fréchet regression. J Am Stat Assoc. 118:1023–1037.
Van den Boogaart KG, Egozcue JJ, Pawlowsky-Glahn V. 2014. Bayes Hilbert spaces. Aus NZ J of Statistics. 56:171–194.
Villani C. 2009. Optimal transport: old and new, Vol. 338. Springer.
Wang J-L, Chiou J-M, Müller H-G. 2016. Functional data analysis. Annu Rev Stat Appl. 3:257–295.
Wang W, Yan J. 2021. Shape-restricted regression splines with R package splines2. J Data Sci. 19:498–517.
Wong RKW, Li Y, Zhu Z. 2019. Partially linear functional additive models for multivariate functional data. J Am Stat Assoc. 114:406–418.
Xiao W, Wang Y, Liu H. 2021. Generalized partially functional linear model. Sci Rep. 11:23428–14.
Yang H, Baladandayuthapani V, Rao AU, Morris JS. 2020. Quantile function on scalar regression analysis for distributional data. J Am Stat Assoc. 115:90–106.
Yuan Y, Zhu H, Lin W, Marron JS. 2012. Local polynomial regression for symmetric positive definite matrices. J R Stat Soc Series B Stat Methodol. 74:697–719.
Zhang J, Merikangas KR, Li H, Shou H. 2022. Two-sample tests for multivariate repeated measurements of histogram objects with applications to wearable device data. Ann Appl Stat. 16:2396–2416.
Zhang Z, Müller H-G. 2011. Functional density synchronization. Comput Stat Data Anal. 55:2234–2249.
Zhou Y, Müller H-G. 2022. Network regression with graph Laplacians. J Mach Learn Res. 23:14383–14423.
Zhu H, Zhang R, Liu Y, Ding H. 2022. Robust estimation for a general functional single index model via quantile regression. J Korean Stat Soc. 51:1041–1070.