Personalized Imputation in metric spaces via conformal prediction: Applications in Predicting Diabetes Development with Continuous Glucose Monitoring Information

Marcos Matabuena*
Health Research Institute of Santiago de Compostela, Santiago de Compostela
Department of Biostatistics, Harvard University, Boston, MA 02115, USA
mmatabuena@hsph.harvard.edu

Carla Díaz-Louzao
Department of Mathematics, University of A Coruña, Spain

Rahul Ghosal
Department of Epidemiology and Biostatistics, University of South Carolina, USA

Francisco Gude-Sampedro
ISCIII Support Platforms for Clinical Research,
Health Research Institute of Santiago de Compostela (IDIS), Santiago de Compostela, Spain
Concepción Arenal Primary Care Center, Santiago de Compostela, Spain
University of Santiago de Compostela, Spain

Abstract

The challenge of handling missing data is widespread in modern data analysis, particularly during the preprocessing phase and in various inferential modeling tasks. Although numerous algorithms exist for imputing missing data, the assessment of imputation quality at the patient level often lacks personalized statistical approaches. Moreover, there is a scarcity of imputation methods for metric space based statistical objects. The aim of this paper is to introduce a novel two-step framework that comprises: (i) a imputation methods for statistical objects taking values in metrics spaces, and (ii) a criterion for personalizing imputation using conformal inference techniques. This work is motivated by the need to impute distributional functional representations of continuous glucose monitoring (CGM) data within the context of a longitudinal study on diabetes, where a significant fraction of patients do not have available CGM profiles. The importance of these methods is illustrated by evaluating the effectiveness of CGM data as new digital biomarkers to predict the time to diabetes onset in healthy populations. To address these scientific challenges, we propose: (i) a new regression algorithm for missing responses; (ii) novel conformal prediction algorithms tailored for metric spaces with a focus on density responses within the 2-Wasserstein geometry; (iii) a broadly applicable personalized imputation method criterion, designed to enhance both of the aforementioned strategies, yet valid across any statistical model and data structure. Our findings reveal that incorporating CGM data into diabetes time-to-event analysis, augmented with a novel personalization phase of imputation, significantly enhances predictive accuracy by over ten percent compared to traditional predictive models for time to diabetes.

*

1 Introduction

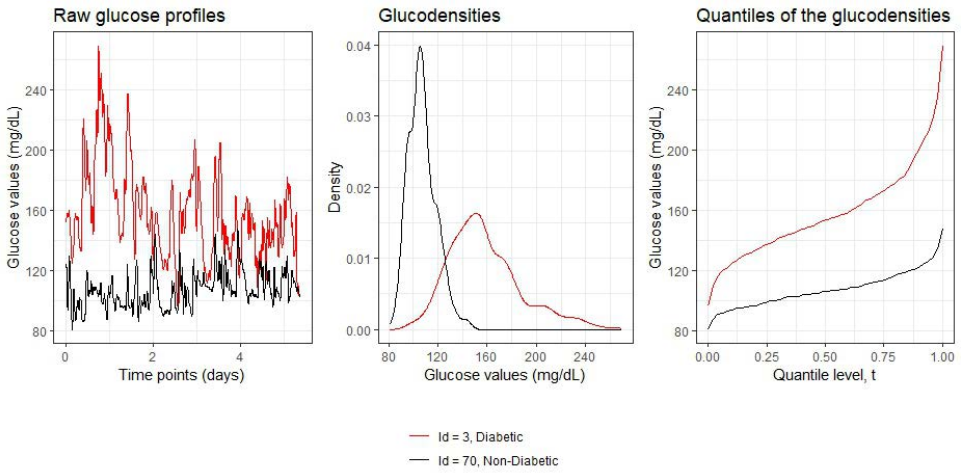
Recent technological advancements are providing new scientific opportunities in biological measurement systems [1]. Consequently, the emerging novel medical tests enable the monitoring of patient conditions in real-time with high-resolution data [2]. This progress has catalyzed the evolution of clinical systems towards precision and digital health paradigms [3, 4]. The next step involves the development of a large number of statistical models to exploit the inherent complexity of these new data structures and support decision-making in the paradigm of personalized medicine [5].

One important example of recent technological advancements is seen with continuous glucose monitoring (CGM) devices [6]. Nowadays, these devices are designed to be minimally invasive and enable the detailed tracking of glucose levels at regular intervals over extended periods, including weeks and months [7]. This provides a comprehensive view of the temporal dynamics of an individual's glucose metabolism [8]. Originally developed to significantly improve the management of potentially dangerous situations, such as low blood sugar episodes in people with type 1 diabetes (hypoglycemia), CGM devices have also proven to be particular useful for managing and monitoring blood sugar levels in individuals with type 2 diabetes.

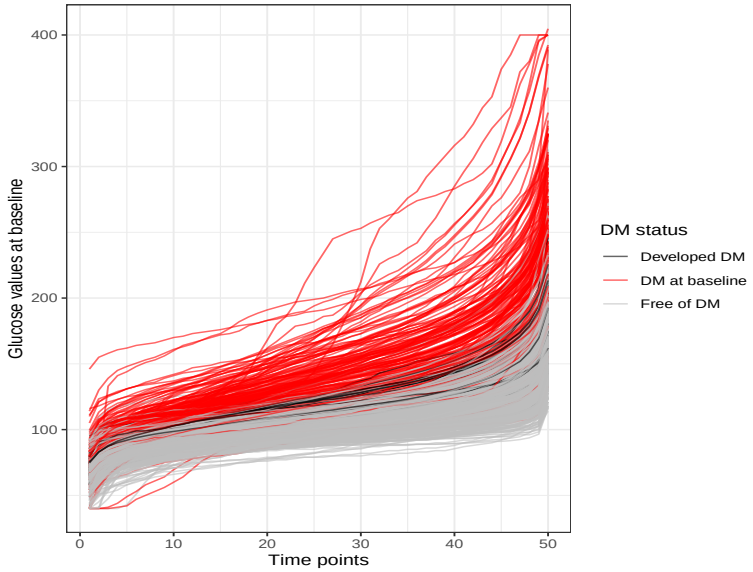
With the increased affordability and enhanced accuracy of non-invasive glucose measurement technologies, their adoption is expanding among healthy populations. In the realm of personalized nutrition, CGM devices are pivotal, facilitating the identification of optimal dietary choices through monitoring real-time glucose fluctuations. Moreover, CGM devices have found application in epidemiological research on medical cohorts composed on health individuals, proving their utility in pinpointing individuals at elevated risk of developing diabetes mellitus [9]. Recent studies have highlighted the advantage of leveraging long-term glucose trends, as reflected by glycosylated hemoglobin levels, over traditional diabetes biomarkers within a general population sample [10, 11]. However, the potential of CGM to predict diabetes incidence and the timing of disease onset in non-diabetic populations is yet to be fully explored. This gap underscores the necessity for efficient imputation strategies within two-step study designs, where only a subset undergoes detailed medical assessments, including CGM monitoring, to address this issue.

This paper delves into a pertinent scientific inquiry, aiming to develop a robust clinical score for diabetes prediction that incorporates the distinct glucose profile captured by CGM technology. We utilize data from the Spanish longitudinal diabetes study, AEGIS [11, 12], adopting a two-step experimental design with a comprehensive ten-year follow-up [13]. This methodological approach distinguishes our work from other studies that do not incorporate CGM data [14, 15]. Historically, the prohibitive cost of CGM devices limited baseline data collection to a secondary subsample of 580 individuals from a larger cohort of 1,516 randomly selected from the general population. The ongoing advancements and cost reductions in CGM technology anticipate its widespread use in healthy demographics, potentially integrating these devices into routine public health diabetes screenings [16, 17]. This evolution underscores the significance of our novel predictive models based on CGM data.

The concept of glucodensity provides more information than traditional CGM summaries [18]. Figure 1a (top panel) shows the glucodensity profiles of a randomly chosen diabetic and non-diabetic subject, while the bottom panel displays these profiles across all subjects, categorized into three groups: those with pre-existing diabetes, those who developed diabetes during the study, and those free of diabetes at study's end. We hypothesize that glucodensity offers particular utility in non-diabetic populations, a stark contrast to traditional CGM composicional metrics that are specifically tailored for disease-populations. Unlike these traditional CGM metrics, which are defined by specific glucose thresholds applicable to diabetics, our method is designed to discern subtle differences in glucose homeostasis over the complete range of CGM values recorded by CGM monitor.



(a) Glucodensity profiles from raw CGM data for a diabetic and non diabetic individual.



(b) Glucodensities profiles of all subjects with CGM, separated according to the status of diabetes. Red: individuals with diabetes at baseline. Black: individuals without diabetes at baseline who developed diabetes throughout the study. Grey: individuals free of diabetes at the end of the study.

We are interested in improving the reliability and power of predictive models for the time to diabetes [19]. For this purpose, we introduce an imputation step for CGM information that minimizes the impact of using a two-step design [20] in terms of statistical efficiency. Generally, in the field of functional data settings, there is a significant gap in the literature concerning the imputation of statistical objects, even for Hilbert space-valued random variables [21]. Inadequate imputation can severely affect predictive models [20] due to the large dimensionality of statistical functional objects. To address this gap in the literature, our study introduces a novel metric space imputation framework based on a weighted least squares global Fréchet model [22], incorporating a conformal prediction [23] step for robust uncertainty quantification. The incorporation of uncertainty quantification steps provides the opportunity to assess the imputation quality and offer personalized imputation rules in line with precision medicine principles [24] This method is applicable not just to glucodensity data [18] but also to other complex statistical objects in separable metric spaces Ω [22].

In diabetes research, risk scores like FINDRISC [25] and GDRS [26] have been developed using logistic or Cox regression models with scalar lifestyle and demographic variables. However,

the integration of CGM data for long-term diabetes onset prediction in healthy populations remains underexplored due to the scarcity of extensive long-term CGM cohorts. Our personalized framework utilizes novel distributional glucodensity representations [18], enhancing the prediction of diabetes onset and providing a more comprehensive understanding of glucose dynamics and progression than traditional screening methods for diabetes mellitus disease.

1.1 Contributions

We briefly summarize the main methodological contributions of this paper as well as the key findings from the analysis of the AEGIS study for modelling time to diabetes.

- To the best of our knowledge, we propose the first global Fréchet regression model for metric spaces with missing responses. Our new estimators, based on inverse-probability weighted estimators, are utilized to impute missing glucodensity. Additionally, we provide an algorithm for estimating the conditional variance of the quantile-based glucodensity representation, assessing the uncertainty resulting from the imputation.
- We extend conformal inference algorithms to provide prediction regions for distributional representations and define a personalized imputation criterion based on the uncertainty related to the imputation. To the best of our knowledge, this is the first work validating the quality of personalized imputations for functional and distributional responses.
- Utilizing the personalized imputation tools in the AEGIS study, we provide the following scientific insights:
    1. We impute the distributional representations of continuous glucose monitoring (CGM) data using the proposed global Fréchet regression model.
    2. With our conformal inference algorithm, we identify patients in whom glucodensity can be imputed for follow-up and time-to-diabetes analysis in this longitudinal study.
    3. In the time-to-event analysis of diabetes, we demonstrate that our glucodensity approach, aided by personalized imputation, outperforms traditional CGM metrics. The proposed models show superior prediction accuracy compared to existing literature, highlighting the effective incorporation of CGM data as a reliable source of information for the progression of diabetes mellitus.
- The codes, along with the methods proposed and utilized in this study, are available for reproducibility purposes. They are publicly accessible at `https://github.com/CarlaDiaz/Conformal_Imputation`.

2 Background and related work

2.1 Statistical Models in Metric Spaces

One of the most prominent applications of statistical modeling in metric spaces is in biomedical problems, particularly in personalized and digital medicine [27]. These applications often involve complex statistical objects, such as curves and graphs, to record physiological functions and measure brain connectivity patterns at a high resolution. A notable example is the concept of "glucodensity" [18], a distributional representation of glucose profiles. This concept has significantly advanced diabetes research methodologies [11] and has proved useful in analyzing accelerometer data [28, 29, 30]. Methodologically, statistical regression analysis in metric spaces is an emerging field [31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41]. Recent publications have explored hypothesis testing [42, 43, 33, 44], variable selection [45], multilevel models [46], dimension-reduction [47], semi-parametric [35, 48, 49], and non-parametric regression models [50, 51, 52, 53].

2.2 Missing Data Imputation and Statistical Methods for Two-Sample Design Studies

The treatment of missing data, a longstanding issue in statistics, significantly impacts medical study reliability, as emphasized by leading medical journals [54]. However, research addressing missing data in metric space models remains scarce. In spaces embeddable in separable Hilbert spaces (negative-type spaces) [55, 56], we have proposed new statistical hyphotesis for randomized clinical trials and paired design [57, 11]. For a standard functional data, current methods primarily utilize functional principal component analysis [58, 59, 60, 61] and multiple imputation [62, 63]. However, these approaches often inadequately address the uncertainty induced by imputation. For distributional data [64, 28], it's crucial to account for the constraints of the underlying space of the distributional object.

Recent studies in standard settings have focused on addressing missing data in large cohorts and high-dimensional data, emphasizing the importance of uncertainty quantification [65], dimensionality reduction [66], and imputation step [67]. Machine learning algorithms, such as XGBoost [68], along with optimal transport-based algorithms [69], have shown promise in imputation tasks, proving to be more efficient in certain non-linear settings. In the field of digital medicine, new methods have been developed for wearable data, such as data from accelerometers, using specialized models that focus on aggregate summaries like physical activity counts [70]. Recent advances in two-sample designs have included proposals for optimal subsampling methods and efficient influence function-based estimation techniques [71, 72]. To date, specific studies on imputing functional data, especially density functions in non-vectorial spaces, remain unexplored. The development of specific methods for functional data, such as medical images and distributional representations for wearable information, is increasingly relevant in precision medicine for the proliferation of summarizing the medical conditions of patients with complex statistical objects [5].

3 AEGIS Study Overview and CGM Glucodensity Approach

3.1 Study Background

The A Estrada Glycation and Inflammation Study (AEGIS) [73], spanning over a decade, longitudinally tracks 1516 subjects to explore health dynamics, focusing specifically on diabetes. A distinctive aspect of AEGIS is the adoption of continuous glucose monitoring (CGM) technology, offering in-depth glucose profiles for a significant subset of participants of health individuals at two time points (years 0 and 5), which is a notable deviation from many clinical studies with shorter durations and fewer participants.

3.2 Study Goals

AEGIS aims to: a) Identify biomarkers within CGM data for stratifying diabetes risk and complications. b) Develop dynamic patient phenotypes based on glucose evolution. c) Characterize metabolic changes to enhance personalized clinical interventions in diabetes research.

3.3 Data Collection and Participants

CGM data were recorded every 5 minutes, encompassing about one-third of the study's participants at various time points. Baseline data include dietary habits, laboratory values, and questionnaires assessing metabolic capacity and lifestyle factors. Out of the 1516 participants, 622 were selected for CGM procedures, with 580 successfully completing the protocol and providing analyzable data.

3.4 Data Analysis Goals

The focus of this paper is to establish a novel diabetes risk model using high-resolution CGM data from a 9-year longitudinal follow-up. This model, formulated using baseline data, aims to highlight the superiority of CGM in comparison to traditional diabetes biomarkers such as A1C and FPG.

3.5 CGM Data Collection Protocol

Participants were equipped with an *Enlite™* sensor and *iPro™* CGM device (Medtronic Inc., Northridge, CA, USA). Glucose concentrations were recorded at 5-minute intervals for 7 days. The analysis omits the first day and any day with over 2 hours of data-acquisition failure.

3.6 Ethical Considerations

The study, sanctioned by the Regional Ethics Committee (Comité Ético de Investigación Clínica de Galicia, code: 2012/025), adhered to the Helsinki Declaration guidelines. Informed consent was obtained in writing from all participants.

3.7 Glucodensity Approach and Distributional Representations

Building upon our prior work [18], this paper introduces the "glucodensity approach" to analyze CGM data, a notable advancement beyond conventional time-in-range metrics in diabetes research. These traditional metrics often categorize glucose levels into fixed intervals, which may not capture individual variations, particularly in non-diabetic cohorts.

3.7.1 Rationale Behind the Glucodensity Approach

The glucodensity method offers a refined understanding of glucose profiles by considering the entire distribution of glucose values, rather than just the time within specific ranges. This approach is crucial for unraveling the complexities of glucose dynamics, essential for comprehending the progression to diabetes and its complications.

3.7.2 Modelling Framework

For each participant in the study, denoted as the i-th participant, we analyze their Continuous Glucose Monitoring (CGM) data, represented as G_{ij} where $j = 1, \ldots, n_i$. These data points are crucial for capturing the distributional characteristics of glucose levels, which are essential for understanding individual metabolic patterns.

We focus on the empirical quantile function of each participant's glucose measurements. This function is defined for each participant as $Y_i(\rho) = \hat{Q}_i(\rho)$, where ρ ranges over the interval $[0, 1]$. Here, $\hat{Q}_i(\rho)$ denotes the generalized inverse of the empirical cumulative distribution function (CDF) associated with the participant's glucose levels. The empirical CDF, $\hat{F}_i(a)$, is given by the proportion of glucose measurements that do not exceed a certain level a, mathematically expressed as $\hat{F}_i(a) = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{1}\{G_{ij} \leq a\}$, where G_{ij} are the glucose values recorded for the i-th participant.

3.7.3 Implications for Diabetes Prediction

Employing this distributional perspective on glucose monitoring provides a comprehensive view of an individual's glucose regulation and its deviations from normative patterns. This method enables more precise and personalized risk assessments for diabetes, potentially leading to earlier interventions and improved management strategies, contrasting with traditional diabetes risk models that often rely on singular biomarkers like A1C or FPG.

3.7.4 Advantages Over Traditional Metrics

The glucodensity approach [18] offers several advantages over traditional CGM metrics: i) It captures a comprehensive view of glucose fluctuations over time, including high and low extremes often overlooked in time-in-range analyses. ii) It facilitates identification of subtle glucose patterns that might signify early metabolic changes leading to diabetes. iii) It is adaptable to various populations, including non-diabetic ones, thereby enhancing its utility in preventive medicine.

3.8 Covariates and Analysis

Our analysis included a subset of 580 participants with complete CGM data. Covariates included demographic characteristics (age, sex, body mass index), laboratory measurements (lipid profile, liver enzymes). Statistical models were adjusted for these covariates to isolate the effect of glucose dynamics on diabetes risk prediction. A complete list of variables used in the creation of predictive score are provided in Table 1.

Table 1. Summary of the predictor variables used in the regression analysis, for the whole sample and separated by the Sex and the fact of having or not CGM at baseline. We represent the means and the standard deviations (in brackets) of the continuous variables (Age, Body mass index (BMI), Glycosilated hemoglobin (HbA1c), Fasting plasma glucose (FPG), Albumin, Insulin) for Men and Women with and without CGM at baseline. For the categorical variables Smoking and Diabetes mellitus , we include the absolute frequency and percentage (in brackets).

	Total sample		CGM		Not CGM	
	Men (n = 678)	Women (n = 838)	Men (n = 220)	Women (n = 360)	Men (n = 458)	Women (n = 478)
Age	51.97 (17.58)	53.09 (17.52)	47.85 (14.79)	48.21 (14.48)	53.96 (18.47)	56.76 (18.69)
BMI	28.56 (4.64)	27.99 (5.40)	28.92 (4.74)	27.71 (5.33)	28.39 (4.59)	28.19 (5.46)
HbA1c	5.65 (0.83)	5.57 (0.65)	5.64 (0.88)	5.52 (0.69)	5.66 (0.81)	5.62 (0.62)
FPG	97.54 (24.50)	92.05 (21.03)	97.06 (23.37)	90.96 (20.85)	97.79 (25.04)	92.87 (21.15)
Albumin	4.48 (0.25)	4.35 (0.22)	4.51 (0.23)	4.36 (0.21)	4.46 (0.25)	4.35 (0.22)
Insulin	14.02 (13.50)	11.69 (7.31)	15.54 (18.41)	11.69 (7.41)	13.29 (10.29)	11.69 (7.24)
Smoking						
Ex-smoker	267 (39.38%)	128 (15.27%)	77 (35.00%)	77 (21.39%)	190 (41.48%)	51 (10.67%)
Smoker	172 (25.37%)	124 (14.80%)	56 (25.45%)	62 (17.22%)	116 (25.33%)	62 (12.97%)
Diabetes mellitus	101 (14.90%)	82 (9.79%)	33 (15.00%)	31 (8.61%)	68 (14.85%)	51 (10.67%)

4 Mathematical Models

4.1 Imputation Step from Distributional Representations

4.1.1 Linear Regression Model for Metric Space Responses: Global Fréchet Model

We consider the multivariate random variable $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, where $X \in \mathcal{X} = \mathbb{R}^p$ and $Y \in \mathcal{Y}$, with $\mathcal{Y} = (\Omega, d)$ being a separable metric space that adheres to the regularity conditions introduced in [22]. These conditions ensure the existence and uniqueness of the Fréchet conditional mean denoted as $m(x) = \mathbb{E}(Y|X = x)$ for all $x \in \mathcal{X}$.

In this article, our focus lies on scenarios where the conditional Fréchet mean is expressed through a linear regression model between the predictor and response variables. Such a regression model is known as a global Fréchet regression and is defined as:

$$m(x) = \arg\min_{y \in \mathcal{Y}} \mathbb{E}\left(\left[1 + (x - \mu)^\intercal \Sigma^{-1}(X - \mu)\right] d^2(Y, y)\right), \tag{1}$$

where $\Sigma = \mathrm{Cov}(X, X)$, and $\mu = \mathbb{E}(X)$. Given an i.i.d. (independent, identical, and distributed) random sample $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$, we can construct an estimator $\hat{m}(\cdot)$ from the

Global Fréchet model as follows:

$$\hat{m}(x) = \arg\min_{y \in \mathcal{Y}} \sum_{i=1}^{n} \omega_{in}(x) d^2(y, Y_i), \tag{2}$$

where $\omega_{in}(x) = \frac{1}{n}\left[1 + (x - \overline{X})^\intercal \hat{\Sigma}^{-1}(X_i - \overline{X})\right]$, with $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, and $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})^\intercal$.

Global Fréchet Regression from Weighted Least Squares

We extend the concept of global Fréchet regression, to incorporate sampling mechanisms induced by missing data patterns, by following scalar response regression models with missing responses. In particular, consider the weighted least squares (WLS) linear regression for scalar response $Y_i \in \mathbb{R}$. Let $\mathcal{D}_n = \{(X_i, Y_i, w_i)\}_{i=1}^n$ be the observed random sample, where w_i denotes the weight of participant i. For $Y_i \in \mathbb{R}$, under a WLS model, the objective function is given by:

$$\begin{aligned}\hat{\beta} &= \arg\min_{\beta} \sum_{i=1}^{n} w_i (Y_i - \langle X_i, \beta \rangle)^2 \\ &= \arg\min_{\beta} \|\sqrt{W}(Y - X\beta)\|^2,\end{aligned}$$

where $\beta = (\beta_1, \ldots, \beta_p)^\intercal$ is a vector of model parameters, $W = \mathrm{diag}(w_1, \ldots, w_n)$ is a weight matrix, $Y = (Y_1, \ldots, Y_n)$, $X = (X_1, \cdots, X_n)$, and $\|\cdot\|$ is the Euclidean norm. The solution is $\hat{\beta} = (X^\intercal W X)^{-1} X^\intercal W Y$.
A future prediction at any $x \in \mathbb{R}^p$ is given by:

$$\begin{aligned}x\hat{\beta} &= x(X^\intercal W X)^{-1} X^\intercal W Y \\ &= k(x) Y \\ &= \sum_{i=1}^{n} k_{i,w}(x) Y_i,\end{aligned}$$

where $k_w(x) = (X^\intercal W X)^{-1} X^\intercal W$ with $k_w(x) = (k_{1,w}(x), \ldots, k_{n,w}(x))$ and $k_{i,w}(x) = \frac{s_{i,w}(x)}{\sum_{i=1}^{n} s_{i,w}(x)}$.
The estimator for conditional mean can be reformulated as:

$$\hat{m}(x) = \arg\min_{y \in \mathbb{R}} \sum_{i=1}^{n} k_{i,w}(x)(Y_i - y)^2. \tag{3}$$

Proposition 1. *The WLS estimator from the global Fréchet model is:*

$$\hat{m}(x) = \arg\min_{y \in \mathcal{Y}} \frac{1}{n} \sum_{i=1}^{n} \omega_{in}(x) d^2(y, Y_i), \tag{4}$$

where

$$\omega_{in}(x) = \frac{w_i \left[1 + (x - \overline{X})^\intercal \hat{\Sigma}^{-1}(X_i - \overline{X})\right]}{\sum_{j=1}^{n} w_j \left[1 + (x - \overline{X})^\intercal \hat{\Sigma}^{-1}(X_j - \overline{X})\right]}$$

and $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, *and* $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})^\intercal$.

This is directly obtained by extending the WLS estimation criterion (3) for global Fréchet model.

Linear Regression Model for Missing Metric Space Responses

Estimation with Missing Observations

In cases where some distributional observations Y_i are missing ($\delta_i = 0$ for some $i \in \{1, \ldots, n\}$), we introduce a special weighting estimator of the form:

$$\hat{m}(x) = \arg\min_{y \in \mathcal{Y}} \sum_{i=1}^{n} \omega_{in}(x) d^2(y, Y_i), \tag{5}$$

$$\omega_{in}(x) = \frac{w_i \left[1 + (x - \overline{X})^\intercal \hat{\Sigma}^{-1} (X_i - \overline{X})\right]}{\sum_{j=1}^{n} w_j \left[1 + (x - \overline{X})^\intercal \hat{\Sigma}^{-1} (X_j - \overline{X})\right]}, w_i = \frac{\frac{\delta_i}{\hat{P}(\delta=1|X=X_i)}}{\sum_{i=1} \frac{\delta_i}{\hat{P}(\delta=1|X=X_i)}}. \tag{6}$$

In essence, here we perform inverse-probability weighting (IPW) and consider only the non-missing observations ($\delta_i = 1$) in the construction of regression model.

4.2 Closed-Projection Algorithm for 2-Wasserstein Metric

In the context of our research, each observation indexed by $i = 1, 2, \ldots, n$ represents a patient under study, with Y_i denoting the distribution or the functional outcome corresponding to the i-th participant. We proceed by constructing the regression model, directly focusing on modeling the point-wise mean of the quantile function $Y_i(t)$, $t \in [0, 1]$ as a function of the covariates. This choice is motivated by the connection of the quantile function to the $2-$Wasserstein distance, and essentially models the Wasserstein barrycenter of the distributional outcome [74] based on the covariates.

The 2-Wasserstein distance, denoted as $d_{W_2}(\mu, \nu)$, serves as a powerful tool for measuring the dissimilarity between probability measures, making it an ideal choice for our analysis. When considering μ and ν as two suitable measures on $\mathcal{R}$ with finite second moments, and Q_μ and Q_ν as their respective quantile functions, the Wasserstein distance $d_{W_2}(\mu, \nu)$ between μ and ν is known to be equivalent to the L^2 distance between Q_μ and Q_ν, as expressed in Equation 7:

$$d_{W_2}(\mu, \nu) = \left[\int_0^1 (Q_\mu(t) - Q_\nu(t))^2 \mathrm{d}t\right]^{1/2}. \tag{7}$$

This elegant equivalence allows us to bridge the gap between probability measures and quantile functions, offering profound insights into the Wasserstein distance's significance. See [64] for various advantages offered by the quantile function based distributional representation as opposed to histogram or densities.

As a consequence, the Fréchet mean [75] of a random measure can be characterized by the point-wise mean of the corresponding random quantile process. Therefore, by introducing a regression model for the random quantile function Y_i, we implicitly construct a model for the conditional Fréchet mean of the underlying glucose distribution measure [33].

Let $X_i \in \mathcal{R}^p$ represent the p-dimensional covariate vector. In this scenario, the global linear Fréchet model takes the form:

$$m(X_i, t) = E(Y_i(t)|X_i) = \alpha(t) + \beta(t)^\intercal X_i, \quad t \in [0, 1]. \tag{8}$$

Here, $\alpha(t)$ represents the intercept function, and $\beta(t)$ denotes the coefficient function.

Assuming we have access to a sample $\mathcal{D}_n = \{(X_i, Y_i, w_i)\}_{i=1}^n$, where Y_i serves as the response quantile function and $X_i \in \mathbb{R}^p$, we employ the weighted least squares criterion to estimate the parameters. The procedure can be outlined in two steps. Firstly, for any $t \in [0, 1]$, we compute the estimates:

$$\left(\hat{\alpha}(t), \hat{\beta}(t)\right) = \underset{a\in\mathcal{R}, b\in\mathcal{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} w_i \left[Y_i(t) - a - b^\intercal X_i\right]^2. \tag{9}$$

These estimates lead to the initial fitted quantile functions:

$$Y_i^*(t) = \hat{\alpha}(t) + \hat{\beta}^\intercal(t) X_i, \quad t \in [0,1]. \tag{10}$$

However, as a function of t, it may occur that $Y_i^*(t)$ is not monotonically increasing. Hence we project this fitted value onto the nearest monotonic function in the $L^2[0,1]$ sense, resulting in valid fitted quantile functions $\hat{Y}_i(t)$ [22]. This process yields fitted values $\hat{Y}_i(t)$ for any set of observed covariates X_i, thus providing valuable insights into the conditional Fréchet mean (based on 2−Wasserstein metric) within the context of our research.

4.3 Conformal inference for distributional representation and missing responses

Conformal inference, a framework for uncertainty quantification in diverse settings, has emerged as a significant tool in statistics, especially in medical applications. Key advantages of conformal prediction include: i) Providing model-independent prediction regions, ii) Offering non-asymptotic guarantees under broad exchangeability assumptions, iii) Delivering fundamentally non-parametric predictive regions.

This paper introduces a novel algorithm for conformal inference in distributional regression models, tailored for responses lying within a 2-Wasserstein space. This framework facilitates the definition of point-wise residuals and involves predictors in a separable Hilbert space, denoted as $\mathcal{H}$. By leveraging the supremum norm, we streamline computation of prediction regions for response quantile functions, focusing on conditional scenarios with covariates to establish Type II tolerance regions. Practically, we connect the regression model, symbolized by $m(\cdot,\cdot)$, with the conditional mean estimator $(m(X_i, t))$. Our goal is to construct a prediction region, $\mathcal{C}^\alpha(X)$, with a confidence level α, ensuring $P(Y \in \mathcal{C}^\alpha(X)) = 1 - \alpha$. This region either minimizes volume or conforms to specified geometric constraints. Utilizing conformal inference on a random sample, $\mathcal{D}_n$, we assure non-asymptotic guarantees: $P(Y \in \widehat{\mathcal{C}}_n^\alpha(X)) \geq 1 - \alpha$, converging to the oracle prediction region as $n \to \infty$.

In this study, we observe $\mathcal{D}_n = \{(X_i, Y_i, \delta_i)\}_{i=1}^n$, where i indexes patient data, and δ_i indicates missing data. The missingness, contingent on covariates Y (i.e., $Y \perp \delta | X$), disrupts sample exchangeability. Achieving marginal non-asymptotic guarantees hinges on the true missing data weights w_i, a rarity in practice. However, with increasing n, the marginal coverage is guaranteed, assuming precise estimation of the regression model and missing data mechanism. The core steps of our proposed conformal prediction algorithm are delineated in Algorithm 1. The calibration sample, $\mathcal{D}_{\text{calibration}}$, is pivotal in conformal inference. It is used to calibrate the algorithm and establish necessary confidence levels or significance thresholds for prediction regions, based on nonconformity scores. This calibration ensures the regions accurately reflect the intended uncertainty level and maintain correct coverage probabilities in the non-asymptotic regime.

Theorem 1. *For any function estimator of the regression function $m(\cdot,\cdot)$, $\hat{m}(\cdot,\cdot)$, invariant to permutations, and a random sample $\mathcal{D}_n = \{(X_i, Y_i, w_i)\}_{i=1}^n$ that is exchangeable (assuming knowledge of w_i), the prediction region $\widehat{\mathcal{C}}_n^\alpha(X)$ for a new observation X, defined by Algorithm 1, satisfies:*

$$P(Y \in \widehat{\mathcal{C}}_n^\alpha(X)) \geq 1 - \alpha$$

Proof. Available in the Supplementary Material. □

Algorithm 1: Conformal Prediction Algorithm for Distributional Responses and Missing Data

1: Partition the sample set $\mathcal{D}_n$ into three distinct and independent random samples: $\mathcal{D}_{train1}$, $\mathcal{D}_{train2}$, and $\mathcal{D}_{calibration}$.
2: Estimate the regression function $m(\cdot,\cdot)$ using the global linear Fréchet model (8) as $\hat{m}(\cdot,\cdot)$ using the random sample $\mathcal{D}_{train1}$.
3: For each observation $i \in \mathcal{D}_{\text{train2}}$ and time point $t \in [0,1]$, do steps 4 ,5 and 6.
4: Compute the estimated response $\hat{m}(X_i,t)$.
5: Calculate the residual $r_i(t) = |Y_i(t) - \hat{m}(X_i,t)|$.
6: Derive the modulation function $\hat{s}(X_i,t)$ from the sample $\{(X_i,r_i)\}_{i\in\mathcal{D}_{\text{train2}}}$, where $\hat{s}(X_i,t) = \hat{sd}(X_i,t)$.
7: For each observation $i \in \mathcal{D}_{\text{calibration}}$ perform steps 8 and 9.
8: Define the nonconformity score $R_i = \sup_{t\in[0,1]} \frac{|Y_i(t)-\hat{m}(X_i,t)|}{\hat{s}(X_i,t)}$.
9: Estimate the empirical distribution $\widetilde{G}^*(t)$ as $\frac{1}{\sum_{i\in\mathcal{D}_{\text{calibration}},\delta_i=1} w_i} \sum_{i\in\mathcal{D}_{\text{calibration}},\delta_i=1} w_i \mathbf{1}\{R_i \le t\}$.
10: Compute the empirical quantile $\hat{q}_{1-\alpha}$ at level $1-\alpha$.
11: Construct the prediction region $\widehat{\mathcal{C}}_n^{\alpha}(X,t) = [\hat{m}(X,t) - \hat{q}_{1-\alpha}\hat{s}(X,t), \hat{m}(X,t) + \hat{q}_{1-\alpha}\hat{s}(X,t)]$.

4.4 Personalized Imputation with Conformal Inference

This paper's primary objective is to develop a rigorous mathematical framework for assessing the validity of imputed response values in patient data. For each patient i, we consider the scenario where the random response $Y_i(t) = \hat{Y}_i(t)$ for $t \in [0,1]$ is imputed, signified by $\delta_i = 0$. The appropriateness of the imputation for each patient is evaluated in light of the associated uncertainty, encapsulated by the parameter $\hat{r}_i$. Specifically, for each patient i and a given confidence level $\alpha \in (0,1)$, the uncertainty radius $\hat{r}_i$ is calculated as the maximum deviation across the interval $[0,1]$, defined by $\hat{r}_i = \max_{t\in[0,1]} |\hat{q}_{1-\alpha}\hat{s}(X,t)|$, where $\hat{q}_{1-\alpha}$ represents the quantile associated with the confidence level α, and $\hat{s}(X,t)$ denotes the standard deviation of the imputed values at time t.

For a given threshold $\gamma > 0$, we define the set of imputed observations as follows:

$$\mathcal{S}_\delta = \{i \in [n]; \delta_i = 0 \text{ and } \hat{r}_i \le \gamma\}. \tag{11}$$

Our goal is to ascertain or estimate the optimal threshold parameter $\hat{\gamma}$ that yields high-quality imputations. This determination is based on an interval quality measure of the response Y or a surrogate outcome Z, especially in cases involving binary events (like disease occurrence) or time-to-event responses (such as censored responses).

4.4.1 Practical Implementation and Model Evaluation

In practice, we evaluate a set of m threshold values for γ, denoted as $\gamma^m = \{\gamma_1 < \gamma_2 < \cdots < \gamma_m\}$. For each threshold value γ_s, we assess the performance of a statistical model T, which is constructed using observations from the set:

$$\mathcal{B}_{\gamma_s} = \{i \in [n]; \delta_i = 0 \text{ and } i \in \mathcal{S}_{\gamma_s}\}. \tag{12}$$

The model at threshold γ_s is then given by:

$$T_{\gamma_s} = T(\{(X_i, \widehat{Y}_i) : i \in \mathcal{B}_{\gamma_s}\} \cup \{(X_i, Y_i) : i : \delta_i = 1\}). \tag{13}$$

4.4.2 Model Selection and Contextual Application

The model T encompasses a variety of statistical methods, including but not limited to regression models, logistic regression, and survival models. The choice of T is contingent upon the nature of the data and the specific research question, aiming to capture the relationship between covariates X and response variable Y or surrogate outcome Z.

In scenarios predicting binary events, T may be a logistic regression model, whereas for time-to-event data, a survival analysis model may be more appropriate. The selection and evaluation of T are thus context-dependent, guided by the specific objectives and characteristics of the study.

4.5 Asymptotic Theory for Linear Imputation in Metric Spaces

To clarify and enhance the presentation of the statistical justification for the consistency of mean imputation within a bounded metric space, denoted by (Ω, d), we revise the description and notation for improved readability. The foundational equations and assumptions are outlined as follows:

We define the functions:

$$M(\gamma, x) = \mathbb{E}\left[\omega(X, x)d^2(Y, \gamma)\right], \quad M_n(\gamma, x) = \frac{1}{n}\sum_{i=1}^{n}\omega_{in}(x)d^2(Y_i, \gamma). \tag{14}$$

These functions are critical for assessing the effectiveness of mean imputation, with M representing the expected metric deviation squared between observed values and an imputation parameter γ, and M_n denoting its empirical counterpart based on a sample of size n.

The following assumptions are necessary for a fixed $x \in \mathbb{R}^p$:

(P0) Both the theoretical and empirical minimizers $\mu_p(x)$ and $\hat{\mu}_p(x)$ are confirmed to exist and be unique, with $\hat{\mu}_p(x)$ being almost surely unique. Moreover, for any $\epsilon > 0$, we ensure that $\inf_{d(\gamma,\mu_p(x))>\epsilon} M(\gamma, x) > M(\mu_p(x), x)$, guaranteeing a unique minimum.

(P1) There exists a lower bound $\epsilon > 0$ for the propensity score $\pi(x)$ for any fixed $x \in \mathbb{R}^p$, ensuring the practical applicability of the propensity score.

(P2) The difference between the estimated propensity score $\hat{\pi}(x)$ and the true propensity score $\pi(x)$ diminishes at the rate of $\mathcal{O}_p(n^{-1/2})$, confirming the reliability of the propensity score estimation as the sample size grows.

Assumption (P0) is crucial for establishing the consistency of the M-estimator $\hat{\mu}_p(x)$, implying that the convergence of M_n to M in the empirical process ensures the convergence of their minimizers. The existence of these minimizers is straightforward if Ω is a compact set. Assumption (P1) introduces a necessary condition related to the propensity score, and (P2) addresses the accuracy of the propensity score estimate.

Theorem 2. *Under assumptions (P0) to (P2) and with the condition that Ω is bounded, for any fixed $x \in \mathbb{R}^p$, the following convergence holds:*

$$d(\hat{m}(x), m(x)) = o_p(1), \tag{15}$$

where $\hat{m}(x)$ and $m(x)$ represent the imputed and actual mean values, respectively. This equation demonstrates that the imputed means converge in probability to the actual means as the sample size increases, validating the consistency of the mean imputation method.

5 Simulation Study

In this Section, we investigate the performance of our proposed distributional imputation method via simulations. For this purpose, we consider the following data generating scenarios.

5.1 Generation of the simulated data

In this Subsection, we describe the data generation mechanisms for the simulation scenarios. We consider three following scenarios for the missing data mechanism.

Missing data mechanism

- **Non-dependent**: The probability that the response is missing does not depend on the covariates. In particular, we set $P(\delta_i = 0|X_i) = p_i = 0.5$.
- **Linear**: The probability that the response is observed (not missing) depends on a linear combination of the covariates as $logit(P(\delta_i = 1|X_i)) = X_i^T\beta$, where $logit(x) = log(\frac{x}{(1-x)})$. Further, we considered scenarios with $p_X \in \{1, 2, 5\}$, denoting the number of scalar predictors. The models for missing responses ($\delta_i = 0$ denotes missing) in each case are as follows:
  - **1 covariate**:
  $$logit(P(\delta_i = 1|X_i)) = -0.75 + 1.55X_i$$
  - **2 covariates**:
  $$logit(P(\delta_i = 1|X_i)) = -0.75 + 1.89X_{1i} - 0.37X_{2i}$$
  - **5 covariates**:
  $$logit(P(\delta_i = 1|X_i)) = -0.75 + 0.82X_{1i} - 0.37X_{2i} + 0.09X_{3i} + 0.53X_{4i} + 0.75X_{5i}$$
- **Non-linear**: The probability that the response is not missing is a non-linear function of the covariates. The models for missing responses ($\delta_i = 0$ denotes missing) for varying number of covariates are:
  - **1 covariate**:
  $$logit(P(\delta_i = 1|X_i) = -0.75 + 2.55X_i^2$$
  - **2 covariates**:
  $$logit(P(\delta_i = 1|X_i) = -0.75 + 1.89X_{1i}^3 - 0.37X_{2i}^2 + 0.75X_{1i}$$
  - **5 covariates**:
  $$logit(P(\delta_i = 1|X_i) = -0.75 + 1.12X_{1i}^3 - 0.37X_{2i}^2 + 1.09X_{3i} + 0.1\sin(2\pi X_{4i}) + 0.75\cos(2\pi X_{5i})$$

In all three cases above, the mean probability that the response is missing is 0.5. Also, the main data generation mechanism for the distributional outcome and scalar covariates is the same for all scenarios and described below.

Data generation mechanism

The data generation mechanism considered for the distributional outcome $Y_i(t)$ and the scalar predictors X_i is same across all the scenarios. The scalar covariates are independently generated as $X_{ij} \sim \mathrm{U}[0, 1]$, $j \in \{1, \ldots, p_X\}$ and $i \in \{1, \ldots, n\}$. Let $T = [0, 1]$ be the quantile grid, which in this case is composed by 50 equidistant points in $[0, 1]$. The distributional response $Y_i(t)$ are generated as

$$Y_i(t) = \sum_{j=1}^{p_X} X_{ij}\beta_j(t) + \frac{\sigma_{lp}}{\sqrt{(SNR)}}\varepsilon_i. \tag{16}$$

The signal-to-noise-ratio (SNR) was set to $SNR = 30$. Here σ_{lp} is the empirical standard deviation of the linear predictor $\sum_{j=1}^{p_X} X_{ij}\beta_j(t)$, and $\varepsilon_i \sim \mathrm{N}(0,1)$ independently. We set $\beta_j(t) = t$ across all the covariates. Four different sample size $n \in \{500, 1000, 2000, 5000\}$ are considered for this simulation study across all possible combination of scenarios (missing data mechanism and p_X).

In Figure 2 we display the trajectories of the distributional outcomes for one simulated dataset with sample size 500 for 1 (left panel), 2 (middle panel), and 5 covariates (right panel), which can all noticed to be non-decreasing.

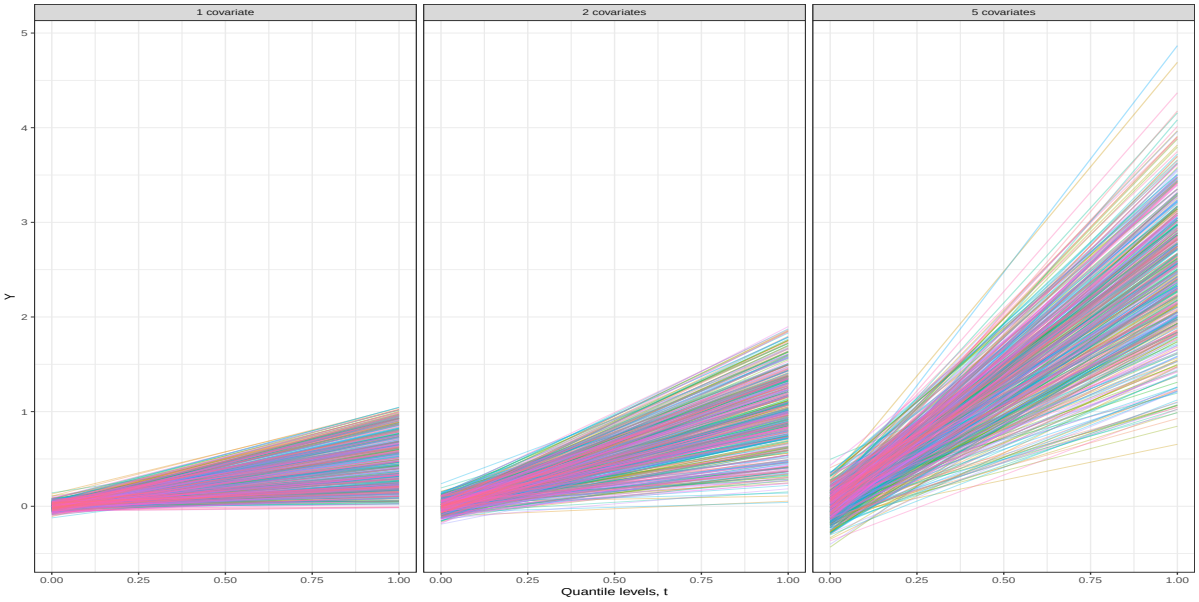


Figure 2. Trajectories of the distributional outcomes $Y_i(t)$ for one simulated dataset with sample size 500 for 1 (left panel), 2 (middle panel), and 5 covariates (right panel).

5.2 Simulation Results

As explained in Section 4, the WLS estimator from the the global Fréchet model includes inverse probability weighting in the estimations, these weights being proportional to the probability that the response is not missing. We estimate $\hat{P}(\delta_i = 1|X_i)$ using a generalized additive model (GAM) for the binary response δ. The `mgcv` R package [76] is used for the GAM implementation. Finally, for each subject $i \in \{1, \ldots, n\}$, the weights are defined as:

$$w_i = \frac{\mathbf{1}_{\delta_i=1}}{\hat{P}(\delta_i = 1|X_i)} = \begin{cases} 0 & \text{if the distributional outcome is missing,} \\ \frac{1}{\hat{P}(\delta_i=1|X_i)} & \text{if the distributional outcome is known,} \end{cases} \tag{17}$$

Next, we impute the missing responses using the WLS global linear Fréchet model (8). The final imputed values for the missing responses $\hat{Y}_i(t)$ are obtained using the closed–projection algorithm for 2− Wasserstein Metric illustrated in Section 4.2. Finally, we evaluate marginal coverage of the proposed conformal inference algorithm on the dataset with missing responses (where $\delta_i = 0$) for a 95% confidence region, i.e., $\alpha = 0.05$.

Figure 3 displays the distribution of estimated coverage across all simulation scenarios. When the mechanism of missing response does not depend on the covariates (Non-dependent scenario, top panel), the median coverage is close to the nominal coverage of 0.95, regardless of the sample size or the number of covariates. The variability in the estimated coverage, decrease with increasing sample size. In contrast, for scenarios where the missing response mechanism depends on the covariates (linearly or non-linearly), the median coverage is somewhat lower, but

always higher than 0.9 and increasing with sample size. The drop in the estimated coverage in these cases is expected, as we are using estimated weights instead of actual weights, and it is known that conformal inference is not accurate for cases where the exact weights are unknown and must be estimated. Nonetheless, this reduction is small, particularly for higher sample sizes, and the nominal coverage rate of 95% lies within the two standard error limit of the average estimated coverage.

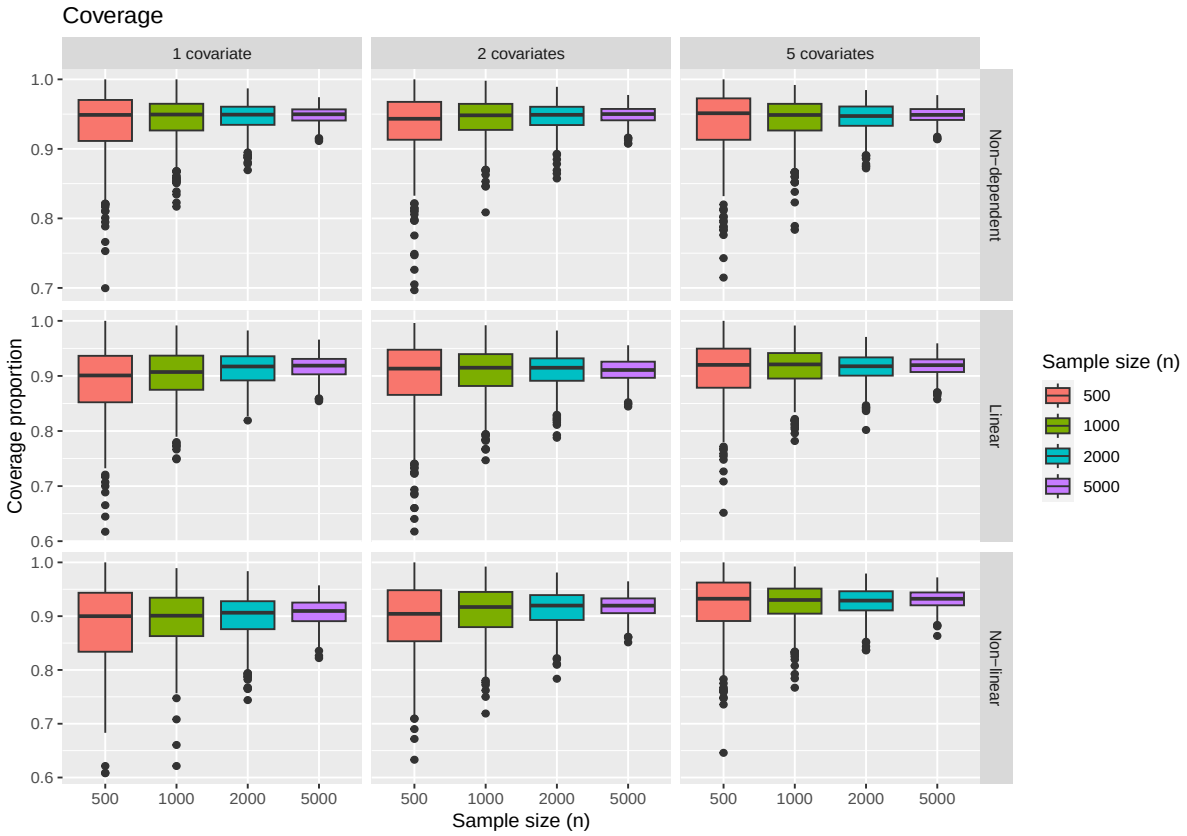


Figure 3. Boxplots of the estimated coverage provided by the conformal inference algorithm for 500 M.C replications across every simulation scenario.

The prediction performance of the WLS global linear Fréchet model (4) was evaluated by means of in-sample R^2 and out of sample Root Mean Squared Error (RMSE). The distribution of R^2 across all the simulation scenarios are displayed in Figure 4. The prediction performance appears to be pretty robust, the median R^2 being higher than 0.8, with a low variability with increasing sample size. The more complex the data generating mechanism, the lower the R^2, which is expected. The Root Mean Squared Error (RMSE) on the test data ($\delta_i = 0$) was computed as,

$$RMSE = \sqrt{\frac{1}{50|I_{test}|} \sum_{i \in I_{test}} \sum_{t=1}^{50} (Y_i(t) - \hat{Y}_i(t))^2}, \tag{18}$$

where $|I_{test}|$ is number of subjects in the test set. Distribution of the Root Mean Squared Error (RMSE) are shown in Figure 5. Across all cases, the RMSE is small and appears to be decreasing with increasing sample size. Comparing the results across different number covariates, it should be noticed that the variability in the distributional outcome increase with increasing number of covariates as evident from Figure 2. Focusing on the sample size $n = 500$, for the 1 covariate scenario, the expected RMSE from data generation is approximately 0.041, the estimated median RMSE of the non-dependent scenario is 0.042, while for the linear one it is 0.053 and for the non-linear one (the most complex) it is 0.061. Similarly, for the case of 2 covariates ($n = 500$),

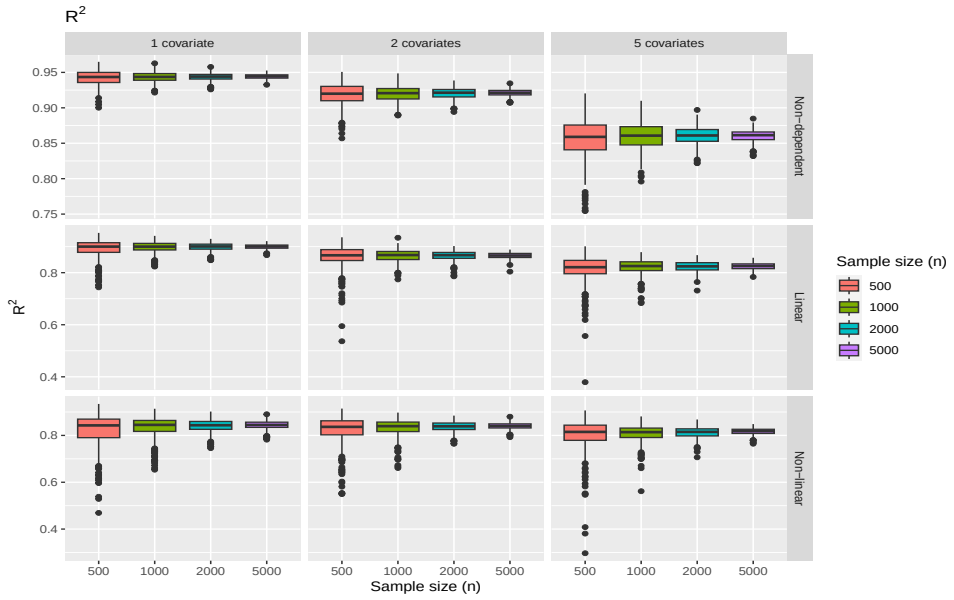


Figure 4. Distribution of the R^2 from the global linear Fréchet model across every simulation scenario.

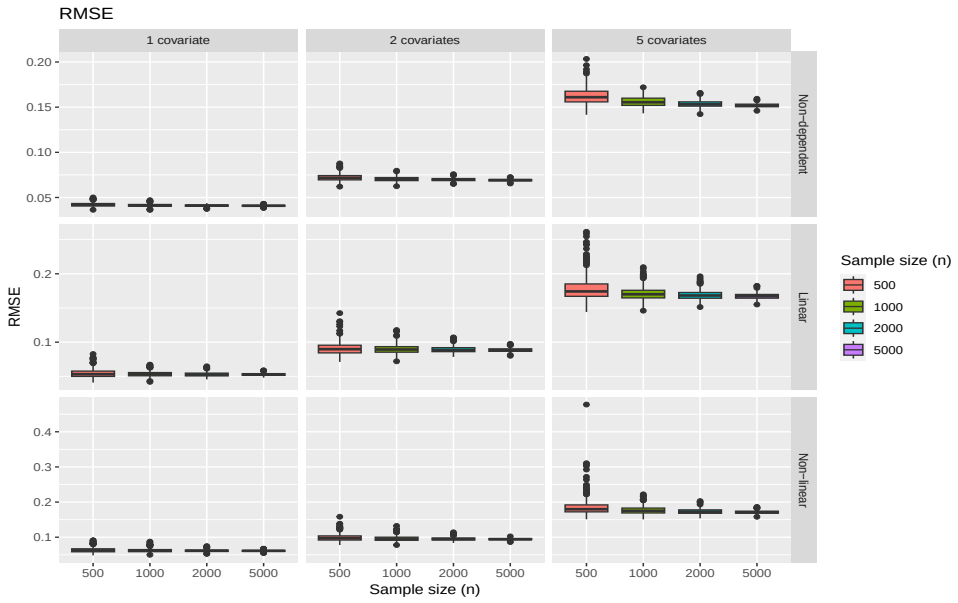


Figure 5. Distribution of the Root Mean Squared Error (RMSE) from the global linear Fréchet model model across every simulation scenario.

the expected RMSE was 0.069, the estimated median RMSE across the non-dependent, linear and non-linear scenario were 0.072, 0.090, 0.097 respectively. Finally, for the case of 5 covariates ($n = 500$), the expected RMSE was 0.069, the estimated median RMSE across the non-dependent, linear and non-linear scenario were 0.151, 0.161, 0.180 respectively. Overall, we see that the error increase with increase in complexity in the data generation mechanism, but, except for a few outliers, the proposed imputation method provide a robust performance.

6 Modeling time to diabetes using distributional CGM information

Diabetes is a complex metabolic disorder where the body struggles to regulate insulin effectively, leading to inconsistent blood sugar levels. It primarily presents in two forms: Type 1 and Type 2. Factors such as genetics, lifestyle, and environmental conditions play crucial roles in the development of diabetes. With an increase in sedentary lifestyles and an aging population, the incidence of diabetes is on the rise. This trend highlights the critical need for innovative public health strategies that focus on early detection and tailored management of the disease.

This section explores the creation of advanced statistical models aimed at predicting diabetes onset by analyzing individual glucose regulation profiles. These profiles are carefully constructed using data from continuous glucose monitoring (CGM) systems over a week. CGM data provides an in-depth analysis of blood sugar variations over time, offering insights beyond conventional health metrics. We utilize a approach called glucodensity, detailed in Section 3.7, to encapsulate CGM data insights. Nevertheless, CGM data was not available for all participants. To address this, we developed a strategy for imputing missing CGM data, described in Section 4.1. This method is further refined with personalized adjustments (Section 4.4), improving the accuracy of our prediction model.

Our analytical pipeline has the following steps. Initially, we use a generalized additive model to evaluate the probability of missing data, leading to the calculation of missing data weights (see Section 6.1). Then, we proceed to the imputation phase, incorporating inverse-probability weighting estimators (Section 6.1.2). After imputing data—combining both imputed and original data—we test the model's effectiveness in predicting diabetes onset using Cox models (Section 6.2). These models blend CGM and non-CGM data to examine the impact of CGM information on prediction precision through time-dependent ROC curves. By applying personalized imputation criteria (Section 6.3), we further enhance the model's predictive performance and explore the potential of CGM data in accurately forecasting diabetes onset.

6.1 Imputing missing Glucodensities using inverse probably weighting estimator

6.1.1 Computing the missing-data weights

As a preliminary step in our analysis, we categorize each participant with a binary label: "0" indicating the absence of Continuous Glucose Monitoring (CGM) data and "1" for those with available CGM data. Interestingly, of the total 1516 subjects in our study, 580 AEGIS-I subjects (*38%*) were equipped with a CGM device at baseline.

Our next objective is to estimate the probability of the availability of CGM data, considering variables such as Age, Sex, HbA1c, FPG (Fasting Plasma Glucose), Smoking habits, Albumin levels, and BMI (Body Mass Index). To accomplish this, we employ a Generalized Additive Model (GAM) with a logistic link function, as proposed by [76], which can be mathematically represented as follows:

$$\log\left(\frac{\mathbb{P}(CGM=1|X=x)}{1-\mathbb{P}(CGM=1|X=x)}\right) \sim s(\text{Age})+\text{Sex}+s(\text{HbA1c})+s(\text{FPG})+\text{Smoking}+s(\text{BMI})+s(\text{Albumin}), \tag{19}$$

where $s(\cdot)$ indicates a smooth function for each variable, estimated using thin-plate regression splines, and $CGM \in \{0,1\}$ represents the binary status of CGM usage. This model demonstrate a moderate predictive power, with a deviance explained of 10.14% and an adjusted R^2 of 0.11. Figure 6 depicts the influence of these variables, highlighting a notable decrease in CGM participation beyond the age of 65. Conversely, participation likelihood increases for individuals

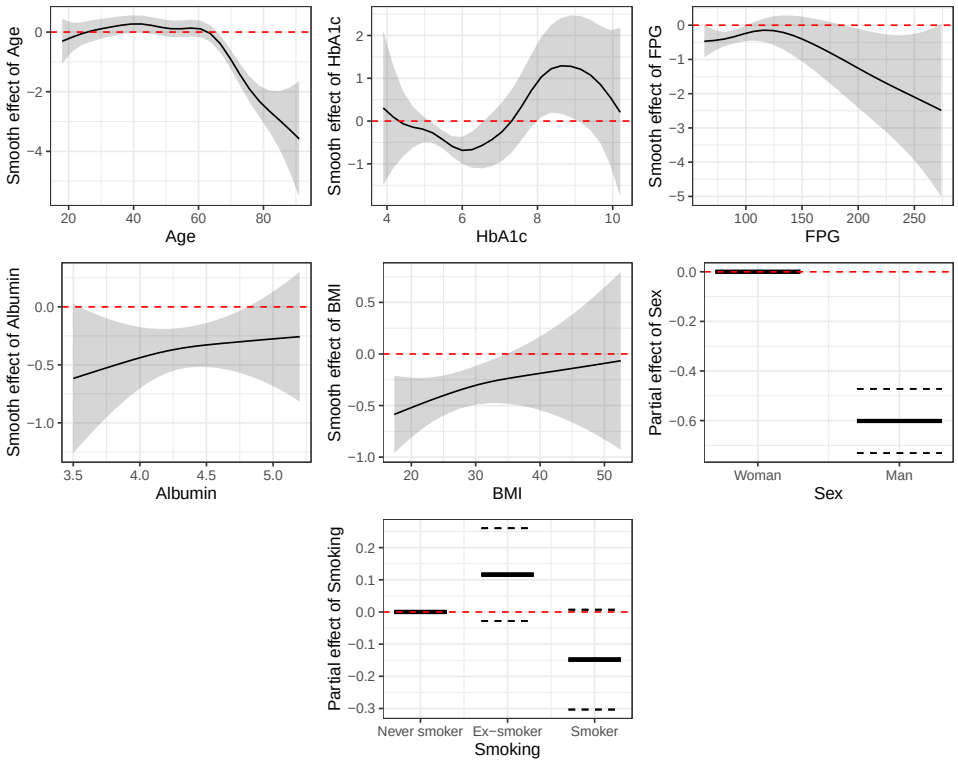


Figure 6. Smooth effect of "Age", "HbA1c", "FPG", "BMI", and "Albumin" on the probability of having CGM.

with elevated HbA1c values and is higher among women. These observations suggest that we are dealing with a "Missing Not At Random" (MNAR) data scenario.

Finally, based on the predictions from the above GAM, we compute the weights for the i-th patient as follows $\omega_i = \frac{\delta_i}{n \cdot \hat{\pi}(X_i)}$, where X_i denotes the characteristics of the i-th patient, and $\hat{\pi}(X_i)$ represents the estimated conditional probability $\mathbb{P}(CGM = 1|X_i)$.

6.1.2 Fitting the global Fréchet model from missing responses

For each participant, labeled as the i-th individual, we calculate their glucose quantile representation, $Y_i(\rho) = \hat{Q}_i(\rho)$, over a spectrum of 101 evenly distributed points, $\tau = \left\{ \frac{j}{100} : j = 0, 1, \ldots, 100 \right\}$. This is achieved by leveraging the empirical distribution, $\hat{F}_i(t) = \frac{1}{n_i} \sum_{j=1}^{n_i} I\{G_{ij} \leq t\}$, which is based on their glucose readings. After establishing the weights and glucodensity quantiles we proceed with the 2-Wasserstein weighted Fréchet regression model based on the covariates: Age, Sex, HbA1c, FPG, Smoking habits, Albumin levels, and BMI. This model demonstrates a coefficient of determination (R^2) of 0.63, indicating a significant portion of the variance is explained. The model's coefficients varying over quantile levels are displayed in Figure 7.

Figure 8 showcases the residual patterns for each participant with CGM data. Specifically, individuals diagnosed with diabetes at the beginning of the study are marked in red, indicating notably higher residuals for this group. This suggests increased uncertainty in predicting glucodensity quantiles among these participants. Following this, in additional analysis presented in supplementary material, we applied a weighted Fréchet regression model to examine the squared residuals, linking them with the same variables previously considered in the Fréchet regression model. The purpose of this analysis was to probe into the conditional variability, and employ a framework of conformal prediction to enhance our understanding of the underlying patterns.

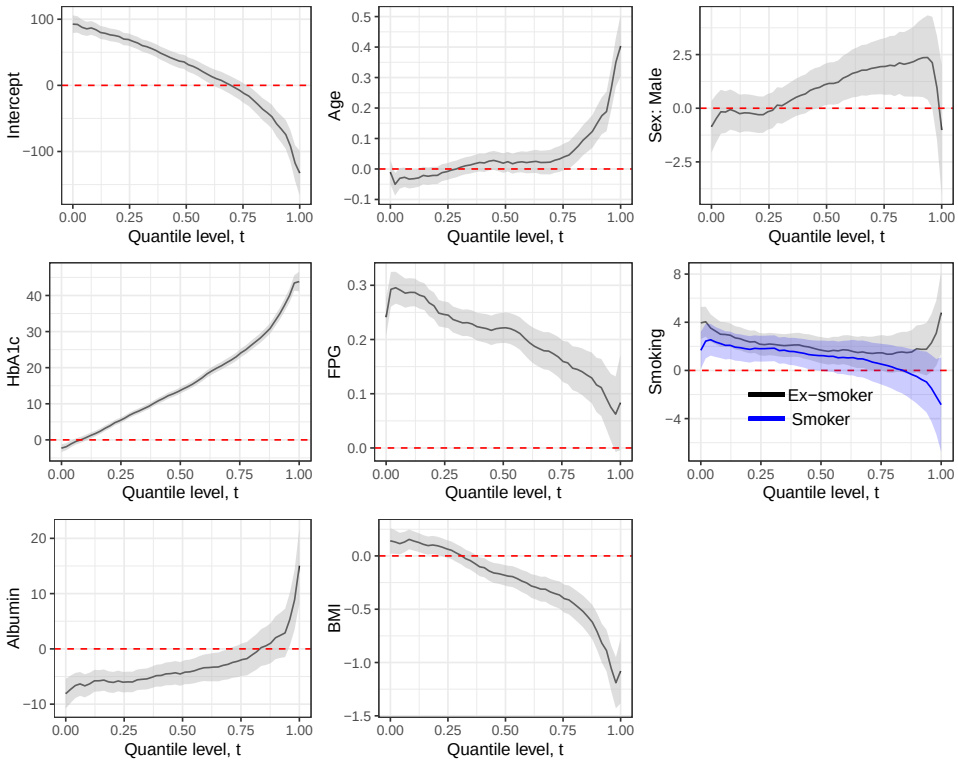


Figure 7. Estimated functional $\beta-$ coefficients of the Fréchet weighting estimator.

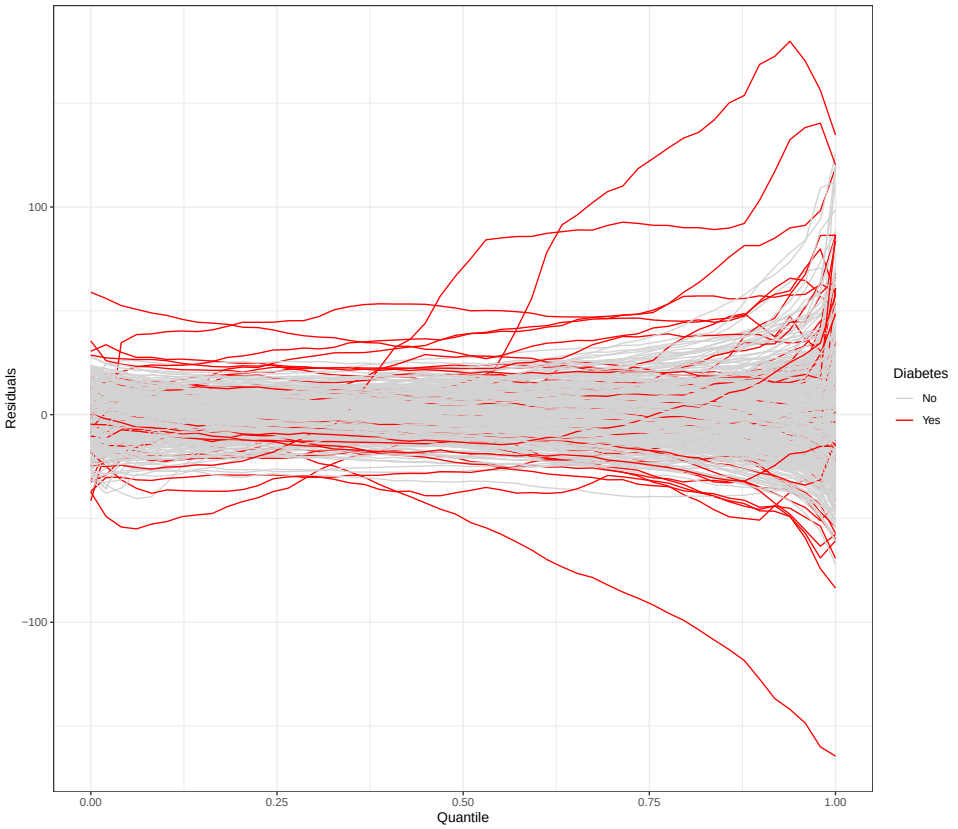


Figure 8. Residuals for the individuals with CGM measurements. Individuals with diabetes are shown in red.

6.2 Survival analysis following imputation in AEGIS

In this Section, we go a step further and take the follow-up time into account, to estimate the probability of time-to-diabetes using survival models. To do so, we are sticking only to the 1293 individuals without diabetes which also have a follow-up time of more than half a year. We used the Fréchet regression model to obtain the CGM-imputed and non-imputed quantile CGM profiles, $Y_i(t)$, for $t \in [0,1]$ and $i = 1, \ldots, n = 1293$ among which 789 individuals (61%) have imputed CGM. The subjects are followed up approximately over a 10 year period, with median follow-up time being 7.4 years. Among the initial sample 75 individuals developed diabetes (5.8%), by the end of the study. The goal of this subsection is to estimate the risk of diabetes over time as a function of the functional principal components (fPC) of the glucodensities, adjusting or not for other covariates, using an additive Cox regression model.

6.2.1 Survival models with all imputed patients and CGM information

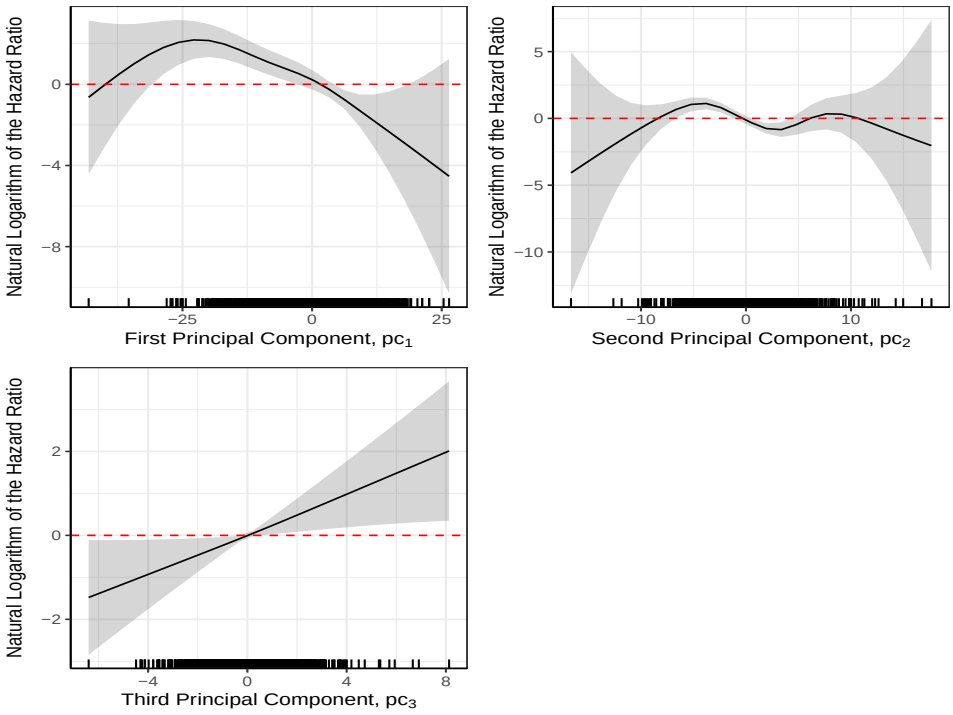


Figure 9. Smooth effect of the covariates of model that only incorporate CGM information.

In our comparative analysis, we start with a time-to-diabetes model, utilizing Continuous Glucose Monitoring (CGM) data exclusively. Specifically, we use both CGM-imputed and non-imputed quantile CGM profiles, $Y_i(t)$, for $t \in [0,1]$ and $i = 1, \ldots, n = 1293$, as functional predictors and extract the first three functional principal component (fPC) scores that account for 98% of the data variability. Subsequently, we adopt an additive Generalized Additive Model (GAM) Cox model as delineated in the 'mgcv' package in R, fitting the time to diabetes onset over a 9-year follow-up. We incorporate the three principal components, pc_j, $j = 1, 2, 3$, in the Cox regression model as follows:

$$h_i(t) = h_0(t) \exp\left\{s(pc_{i1}) + s(pc_{i2}) + s(pc_{i3})\right\},$$

where $h_i(t)$ signifies the hazard function over time for the i-th subject, and $s(\cdot)$ denotes the smooth function of each covariate. The influence of the three principal component scores on the risk model is displayed in Figure 9 reveals non-linear dynamics, with the third score, in particular, showing a linear escalation in diabetes risk. An extended model also incorporates the

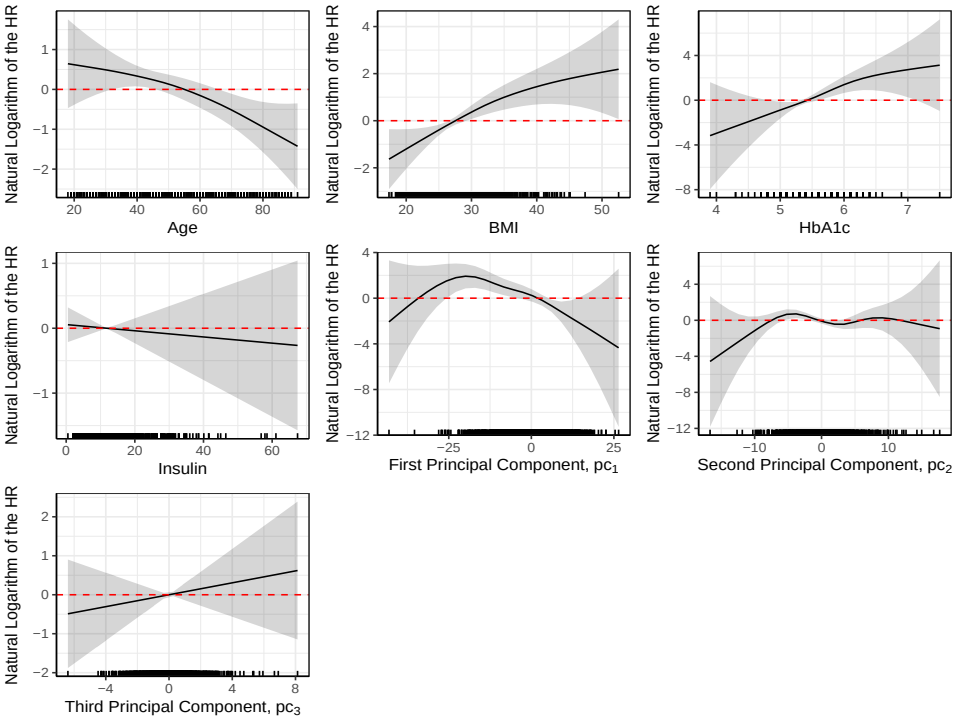


Figure 10. Smooth effect of the continuous covariates of model with CGM and non CGM information).

demographic and clinical covariates Sex, Age, BMI, HbA1c, and Insulin—alongside the PCA scores, and is given by

$$h_i(t) = h_0(t) \exp\left\{Sex_i + s(Age_i) + s(BMI_i) + s(HbA1c_i) + s(Insulin_i) + s(pc_{i1}) + s(pc_{i2}) + s(pc_{i3})\right\}.$$

The estimated effects are displayed in Figure 10 and underscore the nonlinear effects of PCA scores and reveals an increased diabetes risk associated with higher BMI and HbA1c levels. Intriguingly, the risk diminishes with age, suggesting a protective effect against diabetes in individuals maintaining a non-diabetic state and favorable metabolic health over time.

6.2.2 Comparing CGM and non-CGM models in terms of AUC over time

In this section, we evaluate and compare the predictive performance of three distinct models through the lens of the time-dependent Receiver Operating Characteristic (ROC) curve:

1. Employing solely Continuous Glucose Monitoring (CGM) data.
2. Relying exclusively on non-CGM information.
3. Integrating both CGM and non-CGM data.

Figure 11 presents the Area Under the Curve (AUC) for each of these models. A notable enhancement in predictive accuracy is observed when CGM data is combined with traditional biomarkers. To determine the statistical significance of the differences observed between the AUC curves, we performed bootstrap resampling with 1000 iterations to calculate the 95% confidence intervals for the AUC differences. As illustrated in Figure 12, the confidence intervals do not encompass 0, underscoring that the comprehensive model incorporating both CGM and non-CGM information significantly outperforms the model based solely on non-CGM data in terms of AUC. The Concordance Index (C-index) for the comprehensive model, which includes all imputed CGM data as well as non-CGM information, stands at 0.82, highlighting its superior predictive capacity.

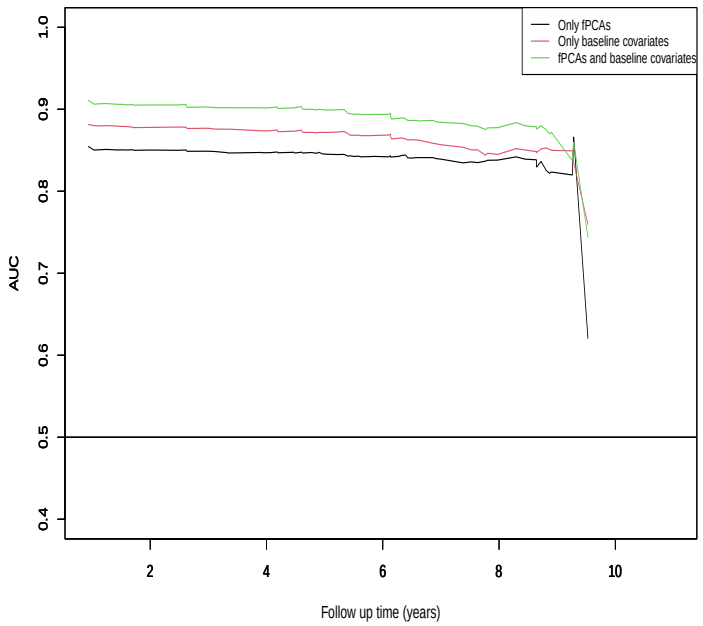


Figure 11. AUC curves for the three survival models considered: i) only CGM-information; ii) only non-CGM information; iii) CGM and non CGM information.

6.3 Personalized imputations vs. globally imputed CGM models

Table 2. C statistics for the survival models with the different subsets depending on the maximum radius of the confidence bands of the glucodensities, and the number of subjects.

Radius	Number of individuals	C index
0	504	0.891
80	511	0.891
90	524	0.892
100	568	0.898
110	566	0.899
120	595	0.882
130	633	0.885
140	686	0.881
150	727	0.877
160	776	0.876
170	822	0.865
180	865	0.874
190	914	0.869
200	992	0.875
370	1293	0.781

We develop personalized imputation criteria for glucodensity quantiles, contrasting these findings with outcomes derived from predictive models employing a uniform imputation strategy across the entire patient cohort. For a thorough analysis, we examine various radii within a pre-defined grid. Table 2 displays the Concordance Index (C-index) values obtained for each defined subset, based solely on the imputed glucodensity quantiles. These analyses are conducted under the condition that the maximum confidence interval radius does not surpass the predefined

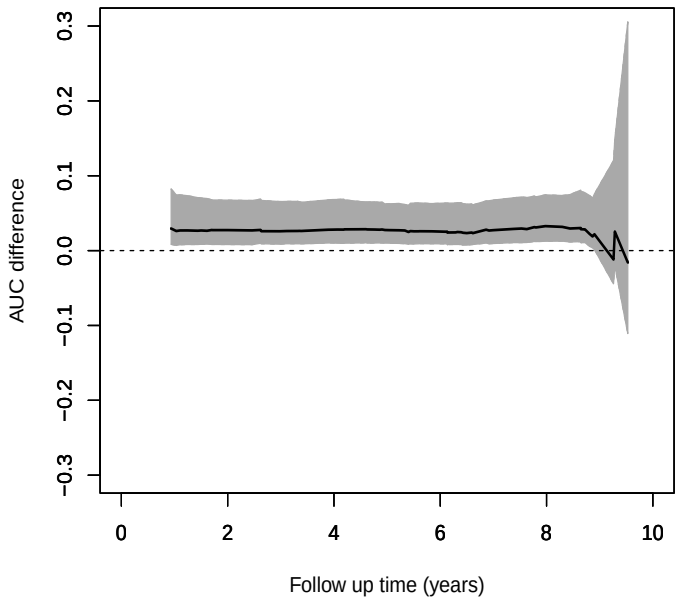


Figure 12. Difference of AUC over time for the model that contain funcional and non functional CGM information. The 95% confidence bands calculated by bootstrap resampling with 1000 samples are shown in grey.

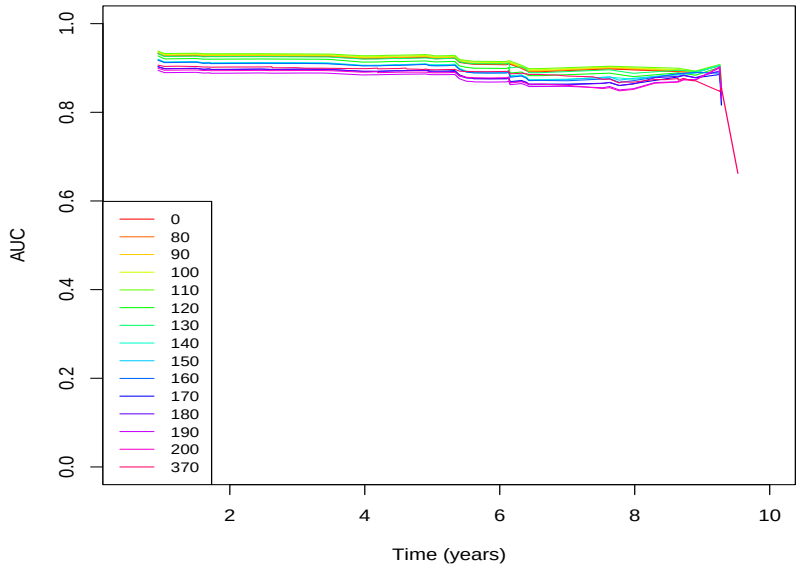


Figure 13. AUC over time for survival models with the different subsets depending on the maximum radius of the confidence bands of the quantiles of the glucodensities.

limit, and the table also lists the number of subjects analyzed at each radius. Importantly, our dataset includes 504 individuals with actual glucodensity measurements. Notably, the optimal C-index (0.90), occurs at a radius of 110 and encompasses 62 subjects with imputed Continuous Glucose Monitoring (CGM) data. Figure 13 illustrates the Area Under the Curve (AUC) over time for all studied radii, indicating that the overall predictive accuracy exceeds traditional CGM risk assessments, a point elaborated upon in our previous discussions. In overall the C-score

with the model with non-functional information is 0.805, there are a improvement of more of ten percent. This improvement highlights the benefit of integrating personalized CGM data into the analysis.

Finally, to elucidate the effectiveness of our personalized imputation approach, Figure 14 displays the conditional Fréchet mean based on the glycemic condition of four representative subjects, showcasing the point-wise results and the predictive bands for the patients, each assigned a distinct maximum radius, and therefore having different levels of uncertainty.

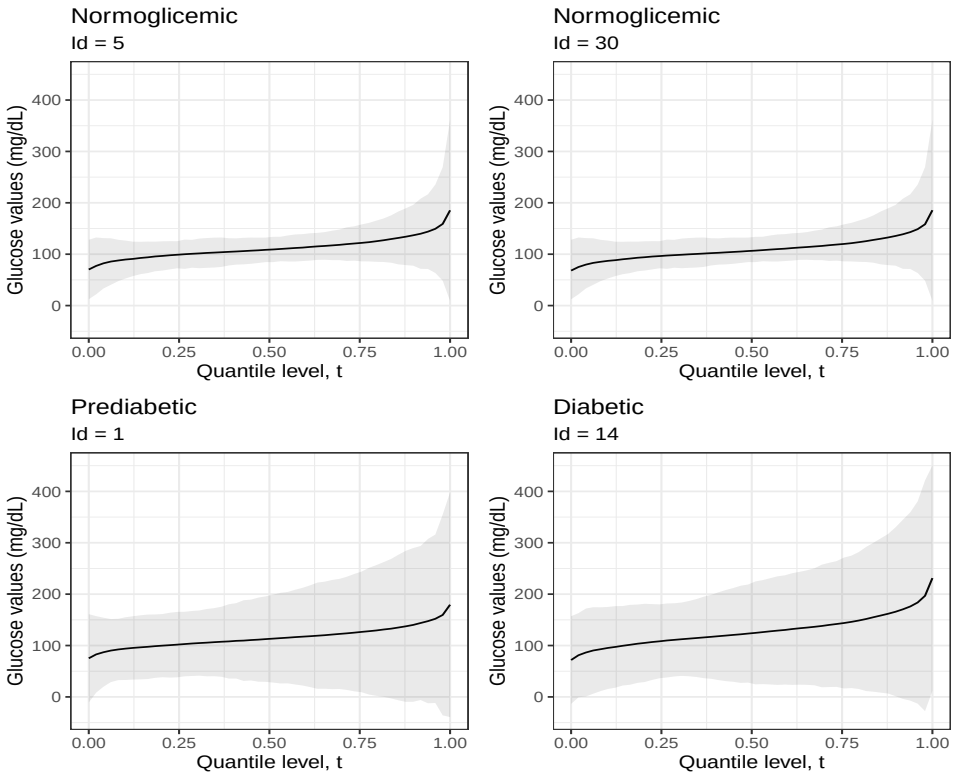


Figure 14. Conditional Fréchet mean and associated prediction regions in Patients with varying glycemic conditions.

7 Discussion

In this study, we delve into a less-explored area of statistical research: personalized imputation strategies for biomedical applications, underscoring the significance of handling missing data, which profoundly impacts both outcomes and predictors [20, 77]. Unlike previous work focusing primarily on optimal sampling techniques and the development of efficient estimators for scalar variables [71, 72], our methodology extends to statistical objects within metric spaces [22], marking a significant advancement in statistical modeling for high-resolution medical data.

Methodologically, this paper introduces several notable contributions: a weighted least squares estimator for linear models in metric spaces [22], specialized imputation methods for these spaces, innovative non-asymptotic inference techniques for conformal prediction algorithms based on the 2-Wasserstein distance, asymptotic theory for conditional mean imputation within a bounded metric space, and a comprehensive personalized imputation method applicable to various clinical outcomes. This approach underscores the importance of balancing imputation accuracy and reliability, especially in predicting time-to-event outcomes, advocating for a direct assessment of improvements in predictive capacity rather than a sole reliance on biomarkers.

Our approach is demonstrated through its application in biomedicine, specifically in predicting the onset of diabetes in a longitudinal study utilizing data from continuous glucose monitors

(CGM). This application not only addresses the widespread use of advanced medical tests in public health initiatives and screening campaigns but also showcases an improvement in model performance by more ten percent when integrating CGM data as a digital biomarker , compared to models relying solely on traditional biomarkers. This improvement in predictive accuracy, validated by the C-score, is consistent with traditional biomarkers and findings from other studies [25, 26]. A key innovation of our model is its ability to integrate high-resolution glucose data through the 'glucodensity' concept [18] using distributional representations [64, 78], offering novel perspectives for the early identification of diabetes risk. Our results suggest the potential of CGM data to create quantifiable methods to assess the glucose homeostasis of the individual in health-populations, a relatively unexplored topic [79]. We explore for the first time the incorporation of CGM information and glucodensity into predicting time to diabetes. In the future, it may be useful in establishing diagnostic thresholds for diabetes from a personalized standpoint based on CGM data.

Multiple research directions remain to be explored based on this current research. For longitudinal or multilevel statistical objects, e.g., distributional profiles, the imputation method would need to carefully account for the correlation present within various sub-clusters [80, 81]. Another interesting direction would be to extend the proposed imputation method to multivariate metric-spaced valued objects, where the distribution of one object could inform another [82, 83].

By addressing the challenges associated with missing data in digital medicine and the statistical treatment of metric spaces, our study highlights the crucial role of personalization in statistical methodologies, evidenced by a substantial real-world application. As the collection of high-resolution longitudinal data becomes more common, the methodologies introduced herein are poised to become increasingly essential in extensive biomedical studies and the integration of data from wearable devices with genetic information [84].

References

[1] Michael S Hughes, Ananta Addala, and Bruce Buckingham. Digital technology for diabetes. *New England Journal of Medicine*, 389(22):2076–2086, 2023.

[2] LBEA Hoeks, WL Greven, and HW De Valk. Real-time continuous glucose monitoring system for treatment of diabetes: a systematic review. *Diabetic Medicine*, 28(4):386–394, 2011.

[3] Dean Ho, Stephen R Quake, Edward RB McCabe, Wee Joo Chng, Edward K Chow, Xianting Ding, Bruce D Gelb, Geoffrey S Ginsburg, Jason Hassenstab, Chih-Ming Ho, et al. Enabling technologies for personalized and precision medicine. *Trends in biotechnology*, 38(5):497–518, 2020.

[4] Jonathan Tyler, Sung Won Choi, and Muneesh Tewari. Real-time, personalized medicine through wearable sensors and dynamic predictive modeling: a new paradigm for clinical medicine. *Current opinion in systems biology*, 20:17–25, 2020.

[5] Marcos Matabuena Rodríguez. Contributions on metric spaces with applications in personalized medicine.

[6] Guido Freckmann. Basics and use of continuous glucose monitoring (cgm) in diabetes therapy. *Journal of Laboratory Medicine*, 44(2):71–79, 2020.

[7] Marcos Matabuena, Marcos Pazos-Couselo, Manuela Alonso-Sampedro, Carmen Fernández-Merino, Arturo González-Quintela, and Francisco Gude. Reproducibility of continuous glucose monitoring results under real-life conditions in an adult population: a functional data analysis. *Scientific Reports*, 13(1):13987, 2023.

[8] Satish K Garg. Past, present, and future of continuous glucose monitors. *Diabetes Technology & Therapeutics*, 25(S3):S–1, 2023.

[9] David C Klonoff, Kevin T Nguyen, Nicole Y Xu, Alberto Gutierrez, Juan C Espinoza, and Alaina P Vidmar. Use of continuous glucose monitors by people without diabetes: an idea whose time has come? *Journal of Diabetes Science and Technology*, 17(6):1686–1697, 2023.

[10] CS Lau and TC Aw. Hba1c in the diagnosis and management of diabetes mellitus: an update. *Diabetes*, 6:1–4, 2020.

[11] Marcos Matabuena, Paulo Félix, Carlos García-Meixide, and Francisco Gude. Kernel machine learning methods to handle missing responses with complex predictors. application in modelling five-year glucose changes using distributional representations. *Computer Methods and Programs in Biomedicine*, page 106905, 2022.

[12] Marcos Pazos-Couselo, Cristina Portos-Regueiro, María González-Rodríguez, Jose Manuel García-Lopez, Manuela Alonso-Sampredro, Raquel Rodríguez-González, Carmen Fernández-Merino, and Francisco Gude. Aging of glucose profiles in an adult population without diabetes. *Diabetes research and clinical practice*, 188:109929, 2022.

[13] Mark Woodward. *Epidemiology: study design and data analysis*. CRC press, 2013.

[14] Yochai Edlitz and Eran Segal. Prediction of type 2 diabetes mellitus onset using simple logistic regression models. *MedArxiv*, 2020.

[15] Nitzan Shalom Artzi, Smadar Shilo, Eran Hadar, Hagai Rossman, Shiri Barbash-Hazan, Avi Ben-Haroush, Ran D Balicer, Becca Feldman, Arnon Wiznitzer, and Eran Segal. Prediction of gestational diabetes based on nationwide electronic health records. *Nature medicine*, 26(1):71–76, 2020.

[16] Ayya Keshet, Smadar Shilo, Anastasia Godneva, Yeela Talmor-Barkan, Yaron Aviv, Eran Segal, and Hagai Rossman. Cgmap: Characterizing continuous glucose monitor data in thousands of non-diabetic individuals. *Cell metabolism*, 35(5):758–769, 2023.

[17] Heather Hall, Dalia Perelman, Alessandra Breschi, Patricia Limcaoco, Ryan Kellogg, Tracey McLaughlin, and Michael Snyder. Glucotypes reveal new patterns of glucose dysregulation. *PLoS Biology*, 16(7):1–23, 07 2018.

[18] Marcos Matabuena, Alexander Petersen, Juan C Vidal, and Francisco Gude. Glucodensities: a new representation of glucose profiles using distributional data analysis. *Statistical Methods in Medical Research*, 30(6):1445–1464, 2021. PMID: 33760665.

[19] Serena C. Y. Wang, Grace Nickel, Kaushik P. Venkatesh, Marium M. Raza, and Joseph C. Kvedar. Ai-based diabetes care: risk prediction models and implementation concerns. *npj Digital Medicine*, 7(1):36, Feb 2024.

[20] Anastasios Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.

[21] Tailen Hsing and Randall Eubank. *Theoretical foundations of functional data analysis, with an introduction to linear operators*, volume 997. John Wiley & Sons, 2015.

[22] Alexander Petersen and Hans-Georg Müller. Fréchet regression for random objects with euclidean predictors. *The Annals of Statistics*, 47(2):691–719, 2019.

[23] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

[24] Anastasios A Tsiatis. *Dynamic Treatment Regimes: Statistical Methods for Precision Medicine*. CRC press, 2019.

[25] K Makrilakis, S Liatis, S Grammatikou, D Perrea, C Stathi, P Tsiligros, and N Katsilambros. Validation of the finnish diabetes risk score (FINDRISC) questionnaire for screening for undiagnosed type 2 diabetes, dysglycaemia and the metabolic syndrome in greece. *Diabetes & Metabolism*, 37(2):144–151, 2011.

[26] Kristin Mühlenbruch, Rebecca Paprott, Hans-Georg Joost, Heiner Boeing, Christin Heidemann, and Matthias B Schulze. Derivation and external validation of a clinical version of the german diabetes risk score (GDRS) including measures of HbA1c. *BMJ Open Diabetes Research and Care*, 6(1):e000524, 2018.

[27] Marcos Matabuena. *Contributions on metric spaces with applications in personalized medicine*. PhD thesis, Universidade de Santiago de Compostela, 2022.

[28] Marcos Matabuena and Alex Petersen. Distributional data analysis with accelerometer data in a nhanes database with nonparametric survey regression models. *arXiv*, 2021.

[29] Rahul Ghosal, Vijay R Varma, Dmitri Volfson, Jacek Urbanek, Jeffrey M Hausdorff, Amber Watts, and Vadim Zipunnikov. Scalar on time-by-distribution regression and its application for modelling associations between daily-living physical activity and cognitive functions in alzheimer,Äôs disease. *Scientific reports*, 12(1):11558, 2022.

[30] Rahul Ghosal and Marcos Matabuena. Multivariate scalar on multidimensional distribution regression, 2023.

[31] Jianing Fan and Hans-Georg Müller. Conditional Wasserstein barycenters and interpolation/extrapolation of distributions. *arXiv preprint arXiv:2107.09218*, 2021.

[32] Yaqing Chen, Zhenhua Lin, and Hans-Georg Müller. Wasserstein regression. *Journal of the American Statistical Association*, (just-accepted):1–40, 2021.

[33] Alexander Petersen, Xi Liu, and Afshin A Divani. Wasserstein f-tests and confidence bands for the fréchet regression of density response curves. *The Annals of Statistics*, 49(1):590–611, 2021.

[34] Danielle C Tucker. *Modeling Non-Euclidean Data via Fréchet Regression*. PhD thesis, University of Illinois at Chicago, 2022.

[35] Aritra Ghosal, Wendy Meiring, and Alexander Petersen. Fréchet single index models for object response regression. *Electronic Journal of Statistics*, 17(1):1074–1112, 2023.

[36] Alexander Petersen and Hans-Georg Müller. Functional data analysis for density functions by transformation to a Hilbert space. *The Annals of Statistics*, 44(1):183–218, 2016.

[37] Yidong Zhou and Hans-Georg Müller. Network regression with graph laplacians. *Journal of Machine Learning Research*, 23(320):1–41, 2022.

[38] Paromita Dubey and Hans-Georg Müller. Modeling time-varying random objects and dynamic networks. *Journal of the American Statistical Association*, 117(540):2252–2267, 2022.

[39] Jeong Min Jeon, Young Kyung Lee, Enno Mammen, and Byeong U. Park. Locally polynomial Hilbertian additive regression. *Bernoulli*, 28(3):2034 – 2066, 2022.

[40] Daisuke Kurisu and Taisuke Otsu. Model averaging for global fréchet regression.

[41] Han Chen and Hans-Georg Müller. Sliced wasserstein regression. *arXiv preprint arXiv:2306.10601*, 2023.

[42] Russell Lyons. Second errata to distance covariance in metric spacess. *The Annals of Probability*, 49(5):2668 – 2670, 2021.

[43] Paromita Dubey and Hans-Georg Müller. Fréchet analysis of variance for random objects. *Biometrika*, 106(4):803–821, 2019.

[44] Alex Fout and Bailey K Fosdick. Fréchet covariance and manova tests for random objects in multiple metric spaces. *arXiv preprint arXiv:2306.12066*, 2023.

[45] Danielle C Tucker, Yichao Wu, and Hans-Georg Mülller. Variable selection for global fréchet regression. *Journal of the American Statistical Association*, 0(0):1–15, 2021.

[46] Satarupa Bhattacharjee and Hans-Georg Müller. Geodesic mixed effects models for repeatedly observed/longitudinal random objects. *arXiv preprint arXiv:2307.05726*, 2023.

[47] Qi Zhang, Bing Li, and Lingzhou Xue. Nonlinear sufficient dimension reduction for distribution-on-distribution regression. *arXiv preprint arXiv:2207.04613*, 2022.

[48] Satarupa Bhattacharjee and Hans-Georg Müller. Single index fréchet regression. *arXiv preprint arXiv:2108.05437*, 2021.

[49] Aritra Ghosal, Marcos Matabuena, Wendy Meiring, and Alexander Petersen. Predicting distributional profiles of physical activity in the nhanes database using a partially linear single-index fréchet regression model. *arXiv preprint arXiv:2302.07692*, 2023.

[50] Christof Schötz. *The Fréchet Mean and Statistics in Non-Euclidean Spaces*. PhD thesis, 2021.

[51] Steve Hanneke. Universally consistent online learning with arbitrarily dependent responses. In *International Conference on Algorithmic Learning Theory*, pages 488–497. PMLR, 2022.

[52] Matthieu Bulté and Helle Sørensen. Medoid splits for efficient random forests in metric spaces. *arXiv preprint arXiv:2306.17031*, 2023.

[53] Satarupa Bhattacharjee, Bing Li, and Lingzhou Xue. Nonlinear global fr\'echet regression for random objects via weak conditional expectation. *arXiv preprint arXiv:2310.07817*, 2023.

[54] James H. Ware, David Harrington, David J. Hunter, and Ralph B. D'Agostino. Missing data. *New England Journal of Medicine*, 367(14):1353–1354, 2012.

[55] Russell Lyons. Distance covariance in metric spaces. *The Annals of Probability*, 41(5):3284–3305, 09 2013.

[56] Russell Lyons. Strong negative type in spheres. *Pacific Journal of Mathematics*, 307(2):383–390, 2020.

[57] Marcos Matabuena, Paulo Félix, Marc Ditzhaus, Juan Vidal, and Francisco Gude. Hypothesis testing for matched pairs with missing data by maximum mean discrepancy: An application to continuous glucose monitoring. *The American Statistician*, 0(0):1–13, 2023.

[58] Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional data analysis for sparse longitudinal data. *Journal of the American statistical association*, 100(470):577–590, 2005.

[59] Cristian Preda, Gilbert Saporta, and Mohamed Hadj Mbarek. The nipals algorithm for missing functional data. *Revue roumaine de mathématiques pures et appliquées*, 55(4):315–326, 2010.

[60] Jeng-Min Chiou, Yi-Chen Zhang, Wan-Hui Chen, and Chiung-Wen Chang. A functional data approach to missing value imputation and outlier detection for traffic flow data. *Transportmetrica B: Transport Dynamics*, 2(2):106–129, 2014.

[61] Christophe Crambes and Yousri Henchiri. Regression imputation in the functional linear model with missing values in the response. *Journal of Statistical Planning and Inference*, 201:103–119, 2019.

[62] Yulei He, Recai Yucel, and Trivellore E Raghunathan. A functional multiple imputation approach to incomplete longitudinal data. *Statistics in medicine*, 30(10):1137–1156, 2011.

[63] Aniruddha Rajendra Rao and Matthew Reimherr. Modern multiple imputation with functional data. *Stat*, 10(1):e331, 2021.

[64] Rahul Ghosal, Vijay R Varma, Dmitri Volfson, Inbar Hillel, Jacek Urbanek, Jeffrey M Hausdorff, Amber Watts, and Vadim Zipunnikov. Distributional data analysis via quantile functions and its application to modeling digital biomarkers of gait in alzheimer,Äôs disease. *Biostatistics*, 24(3):539–561, 2023.

[65] Margaux Zaffran, Aymeric Dieuleveut, Julie Josse, and Yaniv Romano. Conformal prediction with missing values. *arXiv preprint arXiv:2306.02732*, 2023.

[66] Ziwei Zhu, Tengyao Wang, and Richard J Samworth. High-dimensional principal component analysis with heterogeneous missingness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(5):2000–2031, 2022.

[67] Tiantian Liu and Yair Goldberg. Kernel machines with missing covariates. *Electronic Journal of Statistics*, 17(2):2485–2538, 2023.

[68] Yongshi Deng and Thomas Lumley. Multiple imputation through xgboost. *Journal of Computational and Graphical Statistics*, 0(0):1–19, 2023.

[69] Boris Muzellec, Julie Josse, Claire Boyer, and Marco Cuturi. Missing data imputation using optimal transport. In *International Conference on Machine Learning*, pages 7130–7140. PMLR, 2020.

[70] Mia S Tackney, Derek G Cook, Daniel Stahl, Khalida Ismail, Elizabeth Williamson, and James Carpenter. A framework for handling missing accelerometer outcome data in trials. *Trials*, 22(1):379, 2021.

[71] Thomas Lumley and Tong Chen. Choosing good subsamples for regression modelling. *arXiv preprint arXiv:2203.10701*, 2022.

[72] Tong Chen and Thomas Lumley. Optimal sampling for design-based estimators of regression models. *Statistics in Medicine*, 41(8):1482–1497, 2022.

[73] Francisco Gude, Pablo Díaz-Vidal, Cintia Rúa-Pérez, Manuela Alonso-Sampedro, Carmen Fernández-Merino, Jesús Rey-García, et al. Glycemic variability and its association with demographics and lifestyles in a general adult population. *Journal of diabetes science and technology*, 11(4):780–790, Jul 2017. 28317402[pmid].

[74] Rahul Ghosal, Sujit K Ghosh, Jennifer A Schrack, and Vadim Zipunnikov. Distributional outcome regression and its application to modelling continuously monitored heart rate and physical activity. *arXiv preprint arXiv:2301.11399*, 2023.

[75] Alexander Petersen and Hans-Georg MUller. Frechet regression for random objects with Euclidean predictors. *Annals of Statistics*, 47(2):691–719, 04 2019.

[76] Simon N Wood. *Generalized additive models: an introduction with R*. CRC press, 2017.

[77] Dimitris Bertsimas, Colin Pawlowski, and Ying Daisy Zhuo. From predictive methods to missing data imputation: an optimization approach. *Journal of Machine Learning Research*, 18(196):1–39, 2018.

[78] Marcos Matabuena and Alexander Petersen. Distributional data analysis of accelerometer data from the nhanes database using nonparametric survey regression models. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 72(2):294–313, 2023.

[79] Heather Hall, Dalia Perelman, Alessandra Breschi, Patricia Limcaoco, Ryan Kellogg, Tracey McLaughlin, and Michael Snyder. Glucotypes reveal new patterns of glucose dysregulation. *PLoS biology*, 16(7):e2005143, 2018.

[80] Jeff Goldsmith, Vadim Zipunnikov, and Jennifer Schrack. Generalized multilevel function-on-scalar regression and principal component analysis. *Biometrics*, 71(2):344–353, 2015.

[81] Erjia Cui, Ruonan Li, Ciprian M Crainiceanu, and Luo Xiao. Fast multilevel functional principal component analysis. *Journal of Computational and Graphical Statistics*, 32(2):366–377, 2023.

[82] Jeng-Min Chiou and Hans-Georg Müller. Linear manifold modelling of multivariate functional data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(3):605–626, 2014.

[83] Wenlin Dai and Marc G Genton. Multivariate functional data visualization and outlier detection. *Journal of Computational and Graphical Statistics*, 27(4):923–934, 2018.

[84] Ting Li, Yang Yu, J. S. Marron, and Hongtu Zhu. A partially functional linear regression framework for integrating genetic, imaging, and clinical data. *The Annals of Applied Statistics*, 18(1):704 – 728, 2024.