

Translational Research, Design and Analysis Research Article

Cite this article: Gecili E, Huang R, Khoury JC, King E, Altaye M, Bowers K, and Szczesniak RD. Functional data analysis and prediction tools for continuous glucose-monitoring studies. *Journal of Clinical and Translational Science* 5: e51, 1–10. doi: [10.1017/cts.2020.545](https://doi.org/10.1017/cts.2020.545)

Received: 15 March 2020
Revised: 4 August 2020
Accepted: 14 September 2020

Keywords:

Continuous glucose monitoring; functional data analysis; real-time prediction; functional principal component analysis; glycemic excursion; hyperglycemia; hypoglycemia


Address for correspondence:

E. Gecili, PhD, Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, 45229, USA.
Email: emrah.gecili@cchmc.org

© The Association for Clinical and Translational Science 2020. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike licence (<http://creativecommons.org/licenses/by-nc-sa/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the same Creative Commons licence is included and the original work is properly cited. The written permission of Cambridge University Press must be obtained for commercial re-use.



Functional data analysis and prediction tools for continuous glucose-monitoring studies

Emrah Gecili¹ , Rui Huang^{1,4}, Jane C. Khoury^{1,3,5}, Eileen King^{1,3}, Mekibib Altaye^{1,3}, Katherine Bowers^{1,3} and Rhonda D. Szczesniak^{1,2,3}

¹Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA; ²Division of Pulmonary Medicine, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA; ³Division of Pediatrics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA; ⁴Division of Statistics and Data Science, Department of Mathematical Sciences, University of Cincinnati, Cincinnati, OH, USA and ⁵Division of Endocrinology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

Abstract

Introduction: To identify phenotypes of type 1 diabetes based on glucose curves from continuous glucose-monitoring (CGM) using functional data (FD) analysis to account for longitudinal glucose patterns. We present a reliable prediction model that can accurately predict glycemic levels based on past data collected from the CGM sensor and real-time risk of hypo-/hyperglycemic for individuals with type 1 diabetes. **Methods:** A longitudinal cohort study of 443 type 1 diabetes patients with CGM data from a completed trial. The FD analysis approach, sparse functional principal components (FPCs) analysis was used to identify phenotypes of type 1 diabetes glycemic variation. We employed a nonstationary stochastic linear mixed-effects model (LME) that accommodates between-patient and within-patient heterogeneity to predict glycemic levels and real-time risk of hypo-/hyperglycemic by creating specific target functions for these excursions. **Results:** The majority of the variation (73%) in glucose trajectories was explained by the first two FPCs. Higher order variation in the CGM profiles occurred during weeknights, although variation was higher on weekends. The model has low prediction errors and yields accurate predictions for both glucose levels and real-time risk of glycemic excursions. **Conclusions:** By identifying these distinct longitudinal patterns as phenotypes, interventions can be targeted to optimize type 1 diabetes management for subgroups at the highest risk for compromised long-term outcomes such as cardiac disease or stroke. Further, the estimated change/variability in an individual's glucose trajectory can be used to establish clinically meaningful and patient-specific thresholds that, when coupled with probabilistic predictive inference, provide a useful medical-monitoring tool.

Introduction

With the advent of electronic health records and medical devices, modern longitudinal studies typically feature long sequences of observed data. Variables observed over time on study subjects can provide insight into the clinical course of a given biological system or disease, but often through noisy realizations of the true underlying longitudinal process that exhibit natural variation both between- and within-subjects over time. Here, we consider continuous glucose-monitoring (CGM) data where the daily glucose curves were repeatedly observed for each subject along with the duration of the study, and we consider these observations as repeated functional data (FD).

Characterizing and monitoring CGM data have potential value for the assessment of outcomes in clinical studies [1]. Maintaining glucose control is essential for individuals with diabetes mellitus, particularly with respect to the development of comorbidities and during pregnancy. Variability in glucose has been studied for decades as a proxy for diabetes control, but research studies and clinical decision-making are typically based on summary measures [2]. The most commonly reported summary measures are standard deviation, coefficient of variation, and mean amplitude of glycemic excursion, all of which have been used as clinical indicators for years. With more frequent monitoring using CGM, the percentage of time spent within the target range for a given patient is also utilized for care. Although these summary statistics provide a measure for variation around the mean and duration of time within healthy glycemic ranges, the underlying longitudinal structure, the mean response function, and natural variation over time are ignored [3, 4]. This approach ignores additional information that such FD offers [5] and forces clinicians to rely on summary statistics that are prone to measurement error [6]. Obtaining additional information, such as an indication of glucose levels across certain time points, measures of change, rate of change, and variability, and making it available for clinical application has the potential to advance clinical decision-making around optimizing glucose control [5].



Fig. 1. CGM sensor tracings of four representative (first row – females; second row – males) patients aged 18–46 from type 1 diabetes analysis cohort with glucose readings (y-axis, in mg/dL) against clock time (x-axis). Respective demographic/clinical characteristics are on headers. Data points are colored according to observed day of week. RT-CGM, real-time continuous glucose monitoring.

Moreover, summary measures generally result in misleading outcomes if portions of the longitudinal data are missing for a given individual [7]. Missingness in CGM data can arise due to sensor failure, calibration error, or other reasons. These settings produce unequal numbers of repeated measurements and mistimed measurements, both between-subjects as well as within an individual subject recorded over a long duration of time as in multiple CGM sessions (see Fig. 1). In the statistics literature, having such irregular spacing over time is referred to as sparse longitudinal data. The number of observations per individual can vary considerably. Failing to account for sparse longitudinal data through appropriate estimation methods will lead to biased results [8]. In the clinical context, there is an abundance of data; however, the data are distributed in a sparse fashion over time.

Core and novel FD techniques [9] may be useful tools to address issues that are ignored by traditional methods that report simple summary statistics from CGM data. The first goal is to identify phenotypes of type 1 diabetes based on glucose curves. To accomplish this, goal clustering approaches are needed that account for sparse longitudinal patterns of glucose levels. Clustering curve data with strong temporal correlation, like that of longitudinal glucose levels in type 1 diabetes, can be accomplished using the scores from FD analysis technique known as functional principal components analysis for sparse longitudinal data (FPCA) [10].

There are other methods that can be utilized to cluster glucose curves such as latent trajectory classification [11], hierarchical clustering [12], K-means clustering [13], spectral clustering [14], and deep learning-based clustering [15]. As previously demonstrated [4], FPCA will produce similar information to existing techniques when there is an equal number of measurements for each curve and the curves are measured at regular time points across individuals.

However, this is not the typical case for CGM data. In this particular CGM application, we employ FPCA as our clustering approach for historical and methodological reasons. FPCA has been effective in clustering tracings from CGM and oral glucose tolerance test (OGTT) data in the literature [3, 4]. FPCA works by extracting key modes of variation from glucose level trajectories. Furthermore, the FPCA approach provides individual predictions of smoothed curves of nonlinear decline/increase of glucose. The sparse FPCA approach that we employ accommodates incomplete longitudinal data in the form of missing at random (known as MAR) [3], and it accounts for temporal correlation and use of profiles with varying numbers of glucose values. Thus, this methodology is well suited to CGM data analyses, where numbers of values and trajectories of glucose change over time. FPCA is useful for dimension reduction by yielding functional principal components (FPCs) scores that represent modes of variation, and we then utilize these components to identify phenotypes of glycemic variability that ultimately will lead to improved clinical action. In addition to the traditional FPCA method, we characterize the daily specific glucose curves and their secular evolution using the double FPCA approach described previously [16].

In prior work [4], FPCA was performed to extract shape information of OGTT curves from 974 healthy pregnant women in their first trimester, which was not identified by simple summary measures. The obtained information (FPC scores) discriminated between women with and without gestational diabetes later in pregnancy. Additionally, the FPC scores in the first trimester were associated with large-for-gestational-age (LGA) birth, while summary measures suggested there was no association. Other previous work [3] characterized the timing and degree of variability in glucose from 147 women with type 1 diabetes who had repeated

monitoring over the course of gestation. The glucose profiles were clustered into three subgroups of high, moderate, or low heterogeneity, relative to the overall mean response. These clusters, referred to as phenotypes, were associated with clinical characteristics of the cohort at the beginning of pregnancy, longitudinal changes in maternal glycohemoglobin (HbA1c), and weight and pregnancy-related outcomes.

The second goal of the current study is to examine the risk of glycemic excursions throughout the monitoring period in real time. Repeated measures of glucose levels from CGM are clearly correlated over time within a subject, and a traditional random intercept model [17] has been commonly considered to fit longitudinal data. Various semiparametric mixed-effects models have been used to fit glucose trajectories [18]. Further, a semiparametric mixed-effects model with penalized regression splines was considered to provide smooth estimates of the longitudinal glycemic profiles for blood glucose data during gestation [19]. Recently, proposed machine learning and time series techniques for predicting glucose levels have been extensively discussed [20]. These authors later compared support vector machine (SVM), random forest (RF), and autoregressive integrated moving average (ARIMA) models for forecasting glucose level and concluded that the prediction model developed using the RF method was the most accurate of those considered. Techniques such as recurrent neural network (RNN) [21] and artificial neural network (ANN) [22] have been also used for predicting glucose trajectories. Training RNN or ANN models can be computationally expensive. Most RNN models are not able to model sparse and irregularly sampled sequential data [23] like that in our CGM application. RNN models automatically learn features with higher complexity and representations, but the learning capacity of ANN is limited due to the model complexity since these models are mostly implemented in fewer than three layers [21]. These layers are neurons within the neural network that process a set of input features or the output of those neurons.

To produce predictions while preserving clinical interpretability, we adapt a nonstationary Gaussian linear mixed-effects model (LME) that is preferable over the traditional random slope-intercept mixed-effects models, especially when the data have long follow-up sequences to fit the glycemic profiles and predict excursions. In line with our prediction goal, this model provides a framework for estimating the real-time risk of hypo-hyperglycemic based on target functions that we define in Sect. 3.3. Unlike most of the machine learning methods, our chosen approach does not require model training; hence, it is computationally quite efficient. The proposed model yields predictive probabilities that are interpretable for risk assessment, and the traditional parameter estimates provide useful information about how different covariates/features are associated with the response variable. Additionally, this prediction model allows further analysis of the derivative of the fitted curves (e.g., computing rate of change for glucose trajectory) which could be useful if one is interested in obtaining rapid decline/increase in glucose. By contrast, the weights from machine learning prediction techniques, which serve as the primary decision-making information, are not directly interpretable without performing transformations that require additional assumptions [24]. Methods stemming from our chosen approach have been mostly utilized for monitoring in clinical trials to compute the predictive probability of success given interim data [25]. Recently, predictive probabilities have been expanded in the context of monitoring progression toward renal failure [26] and rapid progression of lung function in cystic fibrosis [27].

The next section gives the details of the CGM data. Section 3 provides the statistical methods used in our work. Section 4 presents the results for our FDA and stochastic model applications. We conclude in Sect. 5 with a discussion of the statistical methods used in this work. The computer code for our analyses is provided in Supplemental materials.

Description of CGM Data

We downloaded data from the Jaeb Center for Health Research from the Juvenile Diabetes Research Foundation (JDRF) CGM Study [28]. The goal of this randomized trial was to compare unblinded real-time CGM (RT-CGM) to blinded collection (Control). The analysis cohort used for this study consisted of patients with a clinical diagnosis of type 1 diabetes who had used daily insulin therapy for at least 1 year, were at least 8 years of age, had glycohemoglobin A1C (HbA1C) less than or equal to 10% and insulin regimen involving either use of an insulin pump or multiple daily injections of insulin and had been stable for the last 2 months prior to randomization. The approval of the Institutional Review Board at Cincinnati Children's Hospital Medical Center is not required since the data is publicly available at http://publicfiles.jaeb.org/jdrfapp/dataset/RT-CGM_Randomized_Clinical_Trial.zip.

The analysis cohort includes a total of 443 (232 RT-CGM; 211 Controls) participants [28] and we focus on the primary cohort's 7-day CGM sessions at the first week. Of the 443, 55% of participants are female, 94% of them are White, and mean age (range) is 25 (8–72) years. The glucose level of individuals is reported every 5 min between clock time 0:00 and 24:00 over a week, which yielded a total of 610,823 measurements. The average number of CGM measurements per subject is 204/day and 1379/week, indicating an overall missing data rate of 31.6% (assuming 5-min observations over 7 days). The daily curves of glucose data are repeatedly collected for each subject as the study progresses from the beginning to the end. The daily glucose curve was repeatedly observed for each subject across the duration of the study. This data is considered to be repeated FD. There is substantial variation both between individuals and within a given individual over time (see Fig. 1).

Methods

In this section, we describe the approaches undertaken to complete the two goals of our study – clustering and prediction. Although these two goals are related, we pursue each goal independently. For the first goal, we used FPCA to extract the modes of temporal variation between glucose curves (Sect. 3.1). We then used the two-stage FPCA method to characterize the daily specific glucose curves and their secular evolution (Sect. 3.2). To address the second goal, we fit a Gaussian LME to predict periods of glycemic excursions by using CGM data (Sect. 3.3).

FPC Analysis

FPCA for sparse longitudinal data was used to extract phenotypes of variation from glucose level trajectories. Similar to traditional principal component analysis (PCA), FPCA utilizes linear combinations of a small number of features to maximize variance across data. FPCA achieves this by extracting the common temporal characteristics of a set of curves. This approach was previously performed to identify phenotypes of type 1 diabetes in pregnancy by finger stick data [3]; similar steps were taken for the observed

glucose levels in this study. Briefly to perform the sparse FPCA, we started by choosing a suitable basis function for representing the eigenfunctions using cubic B-splines. We used a routine from the R package “fPCA” to implement the restricted maximum likelihood estimation through a Newton–Raphson procedure, and estimate the FPCs from the CGM data. This approach addressed both the selection of the number of basis functions, as well as the dimension of the process (i.e., number of nonzero eigenvalues) used in the model by minimizing an approximation of the leave-one-curve-out cross-validation score. We implemented the algorithm for each subject’s monitored glucose collection of longitudinal data to obtain smooth individual functions across study duration. Cubic B-splines were specified with equally spaced knots. The candidate models had different settings of the number of basis functions for the eigenfunctions (M) and the number of nonzero eigenvalues used in the model (r). We examined combinations of $M = (4, 5, 6)$ and $r = (2, 3, 4)$, and report results from the model with the best (i.e., the smallest) cross-validation score. This procedure for selecting basis functions and eigenfunctions has been used in prior studies [3, 29]. After fitting the FPCA model (by trying all the combinations of M and R), the final selected model with convergence was $M = 6$ and $r = 4$, which showed four FPCs. Then the scores from the FPC were used to classify CGM sensor tracings as used for previous work in glucose monitoring [3, 4] and other studies [29] using the first and third quartiles (Q_1 and Q_3 , respectively).

Two-stage FPCA

In this part, we consider the CGM data observed during the first week of the study. The two-stage FPCA method is capable of incorporating the nested design of the data for a whole week. The traditional FPCA cannot accommodate nested data; thus, we used this method to assess variation for a single day from each subject’s profile. The goal of the two-stage FPCA was to characterize the daily specific glucose curves and their secular evolution over a week using the double FPCA approach.

The double FPCA [30] procedure is an extension of the conventional FPCA method. It provides a decomposition of the total variation into the variation within the repeatedly observed functions and the variation between these random functions as the second component. The method has several appealing advantages over the traditional FPCA methods. First, it relies on mild assumptions [30]. Previous models for repeated FD rely on a general hierarchical structure. The multiple functions observed for each subject are modeled as a multilevel ANOVA design, relying on the additive assumption [31]. In contrast, the two-stage FPCA approach assumes the functions are smoothly changing over the times at which the repeated functions are recorded, which leads to a non-parametric model under minimal assumptions. Second, the results provide the variation within the repeatedly observed functions as one component and the variation between these random functions the second component. It’s easy to interpret the patient’s glucose trend within a 24-h routine, as well as the evolution of the glucose levels over a long period. Third, the method adopts a local-linear smoother approach [30] to estimate mean and covariance kernel, thus it is applicable to both dense and sparse observations. Even though our data contain individuals with very sparse observations, such as Fig. 1 (the second plot in the first row), we can still effectively estimate the mean surface by borrowing strength from the entire sample.

Real-time Prediction of Glycemic Excursions

The goal of this section is to develop a reliable prediction model that can accurately predict glycemic levels based on past data collected from the CGM sensor and real-time risk of hypo-hyperglycemic for individuals with type 1 diabetes. Below, we provide details of the Gaussian LME with nonstationary covariance that we used for modeling the trajectories of glucose levels and predicting real-time risk of hypo- and hyperglycemic excursions. This model allows for estimating risk based on target functions, which we specified as glycemic excursions. Previously, this model has been used in studies of renal failure and cystic fibrosis [27]. This model captures between and within patient heterogeneities by a random intercept and stochastic process. The main effects are linearly included in model (1); however, the model is able to predict nonlinear curves. It was shown to outperform traditional random slope-intercept models when data has long sequences of repeated measurements [32, 33] which is the case for our CGM data. Some model details are provided below.

We let Y_{ij} denote the glucose measurements for the i^{th} patient taken at time point t_{ij} , where t_{ij} is time of measurement since midnight based on clock time (in hours) and $i = 1, \dots, N; j = 1, \dots, n_i$.

The nonstationary LME has the following form:

$$Y_{ij} = \mathbf{X}_i(t_{ij})\boldsymbol{\alpha} + U_i + W_i(t_{ij}) + Z_{ij}, \quad (1)$$

where $\mathbf{X}_i(t_{ij})$ is the design matrix that includes covariates and $\boldsymbol{\alpha}$ is the corresponding parameter vector. Between-patient heterogeneity is estimated with a random intercept term U_i , where $U_i \sim N(0, \omega^2)$. The terms Z_{ij} are independent, identically distributed as Gaussian with $N(0, \tau^2)$ and represent measurement error. The term $W_i(t_{ij})$ represents realizations from the zero-mean (which represents change in a patient’s glucose level over time that cannot be explained by the linear regressions), continuous-time integrated Brownian motion process such that $W_i(t) = \int_0^t B_i(v)dv$, where $B_i(v)$ is the rate of change in glucose level at time v depicted as Brownian motion and $B_i(0) = 0$. The integrated Brownian motion process is nonstationary and follows a Gaussian distribution with covariance function for time points (hours) s and t :

$$\begin{aligned} \gamma(s, t) &= \text{Cov}(W_i(s), W_i(t)) \\ &= \sigma^2 \frac{[\min(s, t)]^2}{2} \left(\max(s, t) - \frac{\min(s, t)}{3} \right). \end{aligned} \quad (2)$$

This enables greater flexibility compared to traditional models, in terms of the shape of realizations that have been used to characterize variation in glucose levels.

We construct the target function to predict real-time risk of hypo- and hyperglycemic excursions by using model (1). To predict real-time risk of hypo- and hyperglycemic excursions for i^{th} patient at time t_{ik} , we utilize the predicted glucose level $Y_i(t_{ik})$ of that patient at that time point;

We let the covariate history up to a given time t of each patient be represented as $\mathcal{H}_i(t) = \{\mathbf{X}_i, (t_{ij}, y_{ij}) : t_{ij} \leq t\}$. Based on this history, we can build a predictive probability distribution for $Y_i(t_{ik})$ being below or above given thresholds at time t_{ik} :

$$\begin{aligned}
 p_i^{\text{hypo}}(t_{ik}) &= P(Y_i(t_{ik}) < \delta_1 | \mathcal{H}_i(t_{ik})) \\
 &= P(W_i(t_{ik}) < \delta_1 - \mathbf{X}_i(t_{ij})\alpha - U_i - Z_{ik} | \mathcal{H}_i(t_{ik})), \quad (3)
 \end{aligned}$$

where δ_1 is the threshold (in mg/dL) for identifying hypoglycemia and $p_i^{\text{hypo}}(t_{ik})$ is the predicted risk of hypoglycemia.

Similarly, the risk of hyperglycemic excursions at time t_{ik} can be obtained with the following predictive probability distribution

$$\begin{aligned}
 p_i^{\text{hyper}}(t_{ik}) &= P(Y_i(t_{ik}) > \delta_2 | \mathcal{H}_i(t_{ik})) \\
 &= P(W_i(t_{ik}) > \delta_2 - \mathbf{X}_i(t_{ij})\alpha - U_i - Z_{ik} | \mathcal{H}_i(t_{ik})), \quad (4)
 \end{aligned}$$

where δ_2 is the threshold (in mg/dL) for identifying hyperglycemia and $p_i^{\text{hyper}}(t_{ik})$ is the predicted risk of hyperglycemia.

We implemented the model (1) and obtained parameter estimates of the model for our CGM data by using the *ngme* package in R [32]. Prediction performance of model (1) was assessed with predictive accuracy metrics: root-mean-square-error (RMSE) and mean absolute error (MAE). Predicted probabilities (3–4) are also computed by using R and the computer code is available in the Supplemental file.

Results

FPCA Results

Based on FPCA implementation, we found that the first FPC explained 48% of the total proportion of the variation in the CGM readings. The first FPC classifies how each individual patient's glucose level trajectory differs from the mean trajectory. Focusing on the first two FPCs (FPC1–FPC2), which accounted for 73% of the total variation, we created clusters of the trajectories (of patients) according to the first and third quartiles (denoted as Q1 and Q3) of the FPC1 and FPC2 scores. This grouping of individuals across FPC1 and FPC2 quartiles results in nine clusters corresponding to CGM tracings. These classifications are presented in Fig. 2, which illustrates that there are clusters of patients who tend to exhibit hyperglycemic excursions more frequently than the overall cohort (top row). Extreme smoothed values were verified as being similar to the observed CGM values (e.g., top row, third plot from left: the most extreme profile had a peak value of 568 mg/dL (31.5 mmol/L)). Clusters of patients tended to have steadier, normal glucose levels (e.g., middle row, second plot from left); other clusters had lower glucose levels (e.g., bottom row).

Two-Stage FPCA Results

The observed glucose level for a given day was assumed to correspond to a random process, which quantifies glucose level as a function of time of the day. This is considered as the first step of the FPCA. Then we study the subject-specific changes of the functional relationships as weekly time progresses from the beginning of the week to the end of the week as the second step of the FPCA.

The fitted mean surface is visualized in Fig. 3A, which reflects the smoothed CGM tracings (glucose levels) for 10 representative patients from the data, noting higher glucose levels on weekends, compared to weekdays. A sharp increase is seen around Saturday. Fig. 3B provides the first harmonic which provides the “modes of

variation” of the repeated functions sampled on a specific day, is seen to be quite heterogeneous across days of the week. The values below zero in Fig. 3B suggests that profiles tended to be lower, compared to the overall mean CGM profile, during nocturnal hours and on weekends. The second harmonic (Fig. 3C) illustrates higher order variation in the CGM profiles occurred at nighttime throughout the week, although variation was higher on weekends.

Real-Time Prediction Modeling Results

By performing our dynamic FD model, we model the glucose level trajectories in CGM data and provide real-time risk of glycemic excursions. In this part, for illustration, we only consider the data observed on the second day (randomly selected) of first week of the CGM data since the proposed prediction model cannot accommodate nested random effects. We discuss this as a potential extension in concluding remarks. We included time (hours) and group variable related to the clinical trial (1 if subject is from the control group; 0 otherwise) as covariates in model (1), since we do not have information on other features that can be included in the model as covariate such as meals, exercise, insulin regimen. Although the main goal is prediction, the coefficient estimates of intercept, hour, and group variables are 170, -0.632 , and -6.663 , respectively. Additionally, the maximum likelihood estimates of the covariance parameters indicated large between-patient heterogeneity ($\hat{\omega}^2 = 2984$) and residual variance ($\hat{\tau}^2 = 11.14$); estimated variance for the integrated Brownian motion process was $\hat{\sigma}^2 = 0.0052$.

Graphs on the left panel of Fig. 4 provide observed and predicted glucose levels with 95% CI for a 62-year-old White female from the control group with height 160 cm and weight 68 kg, an 8-year-old White female from the control group with height 140 cm and weight 32.8 kg, and a 41-year-old White male from the RT-CGM group with height 168 cm and weight 79 kg, (respectively, from top to bottom). Our prediction model is capable of capturing the observed glucose curves (Fig. 4, left panel). The predicted glucose trajectories are quite similar to the observed glucose trajectories of these patients with small mean predictor of MAE and RMSE: 4.8 (SD = 0.45) and 7.2 (SD = 2); 4 (SD = 0.33) and 6.4 (SD = 2.9); 4.31 (SD = 0.4) and 5.9 (SD = 1.7), respectively (SD: standard deviation; units of both MAE and RMSE are mg/dL). We additionally report overall RMSE and MAE to compare the predictive performance of our model with the traditional random intercept-and-slopes model [17]. Overall RMSE and MAE for our model were 6.3 and 4.1. By contrast, overall RMSE and MAE are 46.4 and 34.8 for the traditional model, which exceeds estimated values from our prediction model.

The graphs on the right panel of Fig. 4 present the predicted risk of hypo- and hyperglycemic excursions for the same three subjects mentioned above. We used $\delta_1 = 60$ mg/dL (3.33 mmol/L) and $\delta_2 = 200$ mg/dL (11.1 mmol/L) for identifying hypoglycemia and hyperglycemia, respectively. The first patient was at high risk of hyperglycemia for the whole day except 6–8 am and 4–12 am. Her risk of hypoglycemia was relatively low except for around 9–12 am in the evening. The second patient was at high risk of hyperglycemia in the afternoon, followed by decreased risk between 6 and 10 pm. Her risk of hypoglycemia was quite high around 9 am in the morning. The third patient was at high risk of hyperglycemia in the afternoon from 1 to 5.30 pm and right before

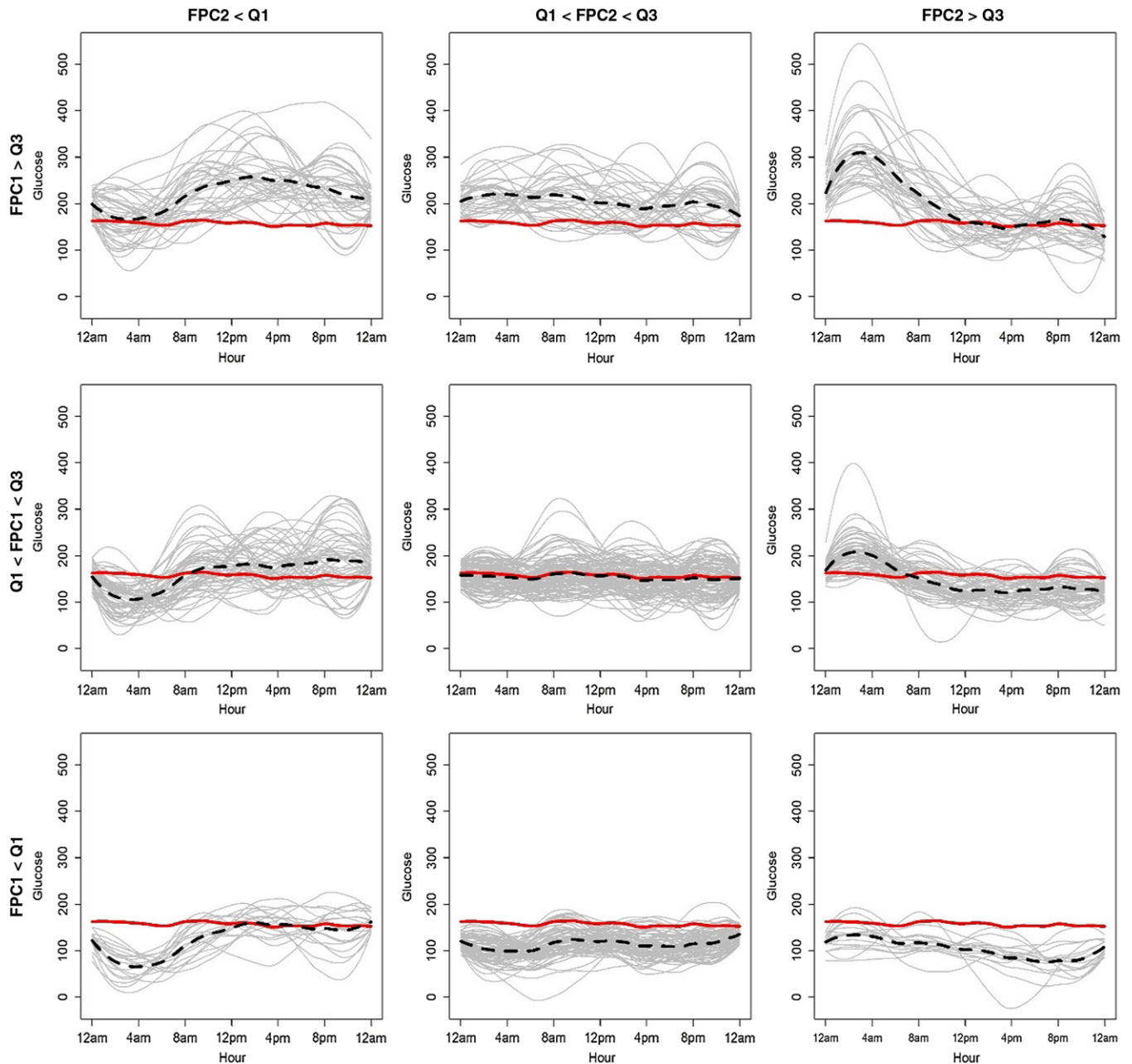


Fig. 2. Phenotypes (clusters) of patients according to glycemic variability over time. Smoothed CGM sensor tracings (gray lines) categorized by quartiles (Q1, Q3) and medians of each of the first two FPCs (FPC1, FPC2) scores in the functional principal components analysis for sparse longitudinal data (FPCA). The solid red line is the mean function of glucose (y-axis) over clock time (x-axis); the dashed black line is the mean function for the specific groups.

midnight from 11 to 12 am. His risk of hypoglycemia was quite low except between 9 and 10 pm in the evening.

Conclusions

Some studies have shown that having glycemic fluctuations is a deterministic factor for hypo- and hyperglycemic excursions. Recently, CGM systems have emerged as an effective technology with an ability to monitor glucose trends over time. Involving large amounts of irregularly observed data, CGM systems provide information, every 5 min, enabling the capture of frequency of fluctuations regarding blood glucose levels [34], so that efficiently utilizing

CGM data would shed light on variation in blood glucose trends over time, which is frequently more difficult to measure. Hence, advanced statistical tools, as we have proposed here, are needed to efficiently study CGM data.

We first presented FD analysis tools for sparse longitudinal patterns of medical-monitoring data to classify glucose curves to identify phenotypes of type 1 diabetes. We considered the temporal information from the CGM FD, and have classified the fitted glucose curves into different clusters. The scores provided by this analysis can be used to examine a range of phenotypes as shown in Fig. 2. The limitation of this method was that it is not accounting for the nested design of the CGM data that was based on daily

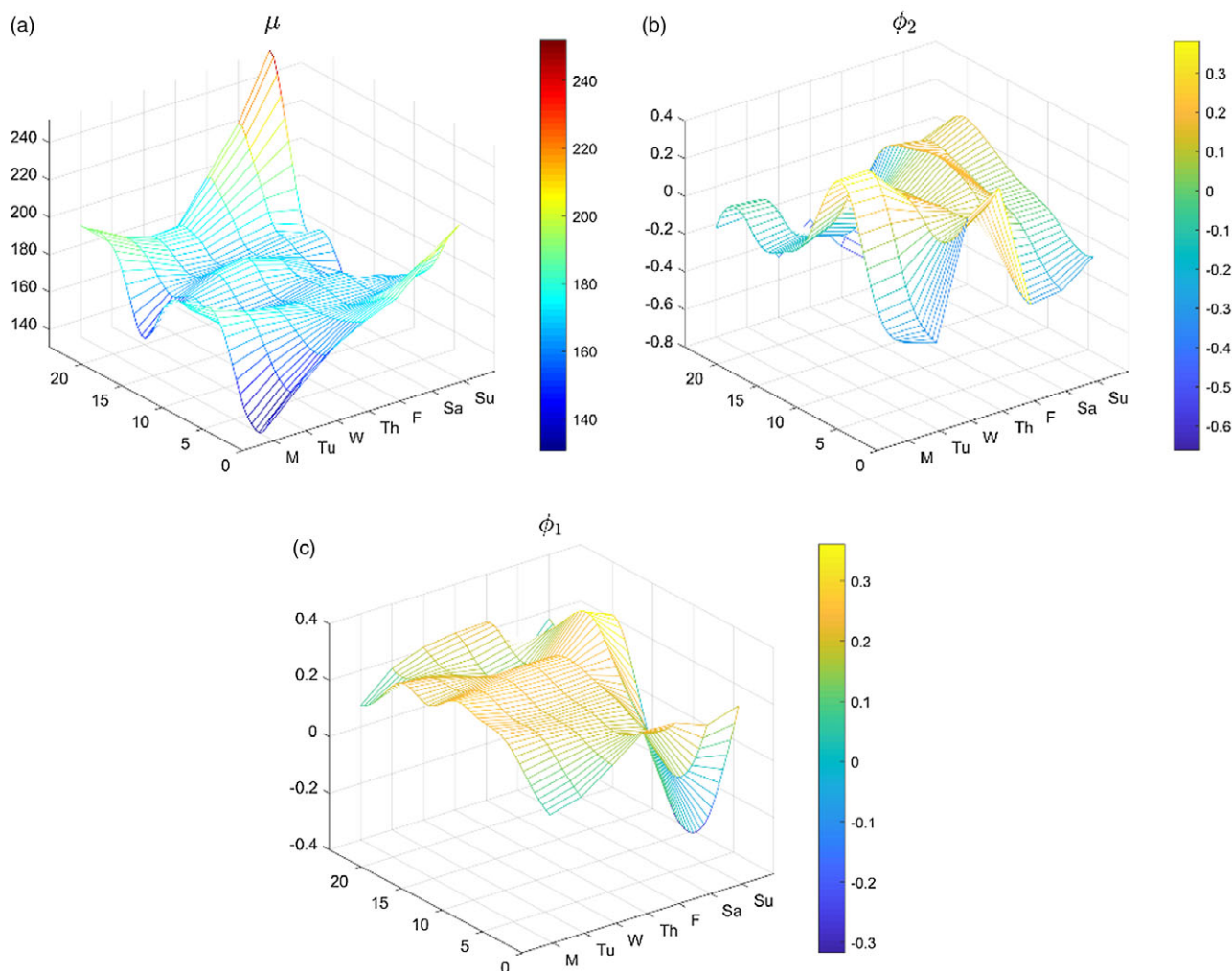


Fig. 3. Two-stage functional principal components analysis for sparse longitudinal data (FPCA) shows poorer glycemic control at nighttime and on weekends (three-dimensional manifold plots of FPCA on the CGM cohort). In each plot, the hour of CGM (0–24 h represents 12–12 am) is on the lower left axis; day of week is on the lower right axis; magnitude is on the upper axis. The vertical axes represent (A) glucose level (mg/dL); (B–C) degree of oscillatory variability in the first and second FPCs, respectively, which are unitless quantities. The vertical heatmap bars depict values ranging from lower magnitudes (blue) to higher magnitudes (red). (A) Smoothed CGM tracings for 10 representative patients, (B) the first harmonic, and (C) the second harmonic.

CGM data observed within a 24-h period. Hence, we then performed two-stage FPCA which incorporates the nested design of the CGM data and enables classification of the GM curves that were repeatedly measured over a week. Performing two-stage FPCA allowed us to compare the glucose curves observed on different days of a week; it was observed that the glucose levels are higher on weekends, compared to weekdays. Further, the glucose profiles tended to be lower, compared to the overall mean CGM profile, during nighttime and on weekends. Moreover, higher order fluctuation in the CGM profiles occurred at nighttime during the week, although variation was higher on weekends which agrees with previous findings [35, 36].

Although the two-stage FPCA has clear advantages over the traditional FPCA approach, we illustrate the utility of each method in CGM analysis for our clustering goal. Our rationale is that, although both approaches have made important but limited appearances in the diabetes literature, the traditional approach has been more frequently applied. However, CGM studies tend to consist of multiple days for a given subject; therefore, our study examined the findings and utility of both traditional FPCA and the

two-stage approach. Further, regular FPCA can be used when one is interested in analyzing only one-day data to cluster daily trajectories (or one week if one is clustering the weekly trajectories). However, the two-stage approach assumes CGM recordings follow a nested design. Although our study goals did not involve comparing the FPCA-based clustering method with other approaches, it may be a worthwhile empirical study to examine our chosen approach alongside the aforementioned statistical and machine learning techniques for clustering.

We addressed our second goal by utilizing a statistical method for predicting and detecting real-time risk of hypo- and hyperglycemic excursions by using long, irregularly observed time series, tailoring the approach to type 1 diabetes. Based on the results in Sect. 4.3, we see that the prediction model that we implemented has very low prediction errors (e.g., low RMSE and MAE), especially compared to the errors obtained with the traditional random intercepts-and-slopes model. The predictive performance of our model is not compared with other previously employed techniques in literature; however, this also represents an important area for future empirical investigation. Additionally, by visually inspecting

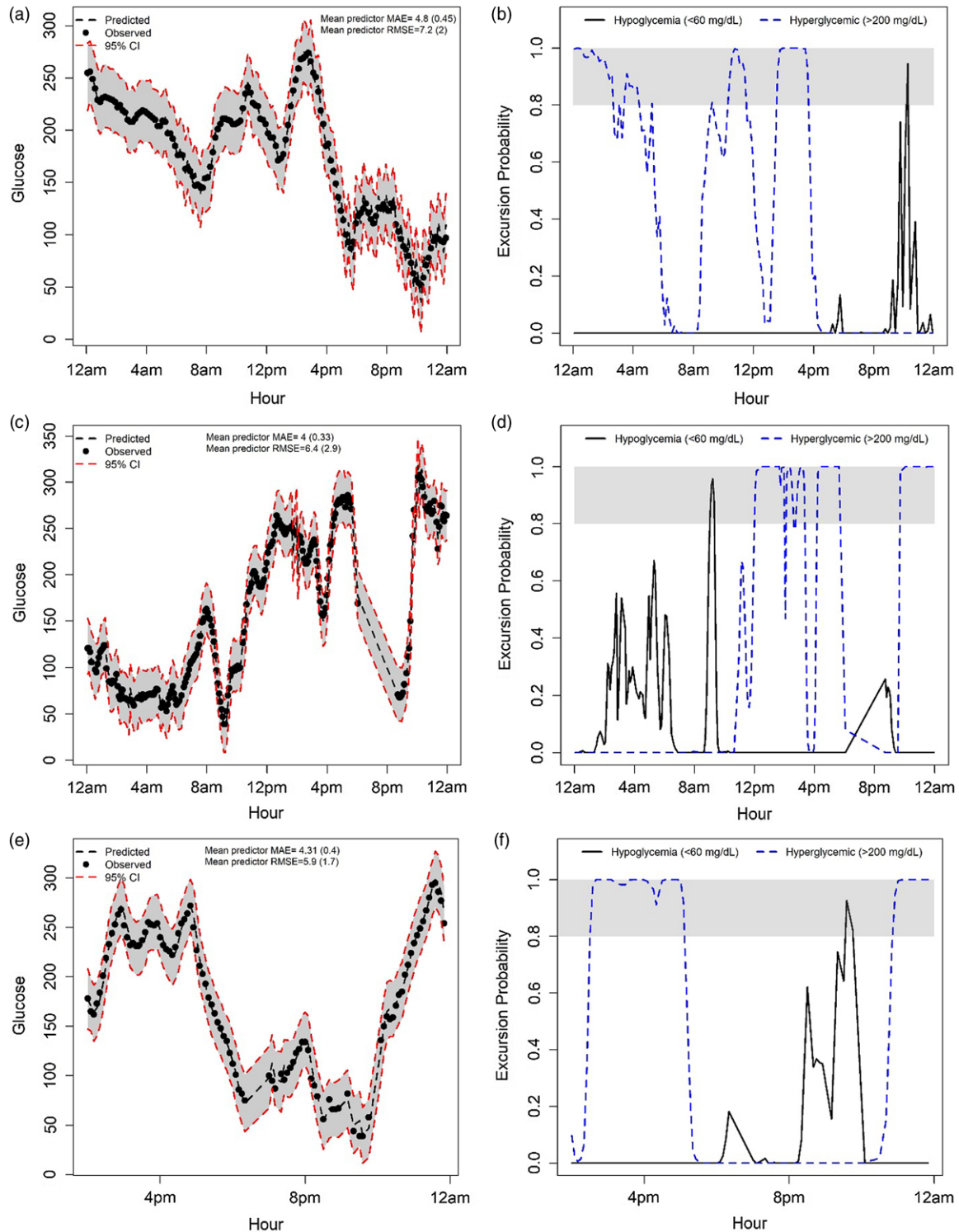


Fig. 4. Observed glucose tracings and model fit/prediction for three different study subjects (one per row). The first row is for a 62-year-old White female from the control group; height: 160 cm; weight: 68 kg. The second row is for a 8-year-old White female from the control group; height: 140 cm; weight: 32.8 kg. The third row is for a 41-year-old White male from the RT-CGM group; height: 168 cm; weight: 79 kg. **Left panel:** Observed glucose readings (y-axis) from CGM (black dots) over clock time (x-axis) are shown with FD prediction (dashed line) and 95% CI (gray band with red dashed lines); **Right panel:** real-time risk for glycaemic excursions (black line is the probability of hypoglycemia; blue line is the probability of hyperglycemic; gray band is the area where probabilities ≥ 0.80 or 80%).

the excursion plots, we can conclude that the predicted real-time risks of hypo-hyperglycemic are highly accurate (the predicted probabilities are in gray band (the area where probabilities

≥ 0.80 or 80%) when the glucose level is over 200 mg/dL (11.1 mmol/L) and below 60 mg/dL (3.33 mmol/L) for hyperglycemia and hypoglycemia, respectively, in the graphs presented in

the right panel of Fig. 4). This paper also shows how estimated change/variability in an individual's glucose trajectory can be used to establish both clinically meaningful and patient-specific thresholds that, when coupled with probabilistic predictive inference provide a useful medical-monitoring tool. Therefore, increasing patients' and clinicians' abilities to take timely actions in a more accurate manner.

This is the first study to apply the two-stage FPCA method to CGM data to identify the daily specific glucose curves and their evolution over a week. Similarly, this is the first study to propose different target functions based on glucose thresholds, tailored to hypo and hyperglycemic. Our novel applications have particular relevance to pregnancy outcome. Specifically, our group has recently published using self-glucose-monitoring data to predict premature delivery in women with type 1 diabetes using a joint modeling approach. Using CGM data would most likely yield even better prediction to a very important obstetrical problem [37].

There are several limitations to the methods provided in our work. Although the original experiment collected daily measurements of glucose over multiple weeks, we only used part of the data for our analysis due to lack of ability to account for the nested design of the whole data (as explained in the previous sections, we considered only 1 day's data for the procedures provided in Sects. 3.1 and 3.3 and used just one week of data for the method provided in Sect. 3.2). Our analyses did not include covariate adjustment since these additional features of the study cohort were not accessible. As aforementioned, we acquired these data from a previously completed study in which there are limited details on the reasons for missing data. However, missingness in CGM data is common and can occur due to various reasons, such as intermittent sensor errors, sensor compression, and user errors [38]. The most common reason for missing data, however, is patient non-compliance in wearing the monitor. Most trials require a minimum number of days to wear the CGM sensor and a minimum number of hours of glucose values, including night-time values. Thus, missing data is inherent in any analysis of CGM data [39].

There are various new developments that could be undertaken based on our application and findings. An important advancement could be extending the prediction model to incorporate the typically nested design of the CGM data measured over weeks/months to improve the prediction accuracy of glucose levels and real-time risk of hypo-hyperglycemic excursions; which would help in better understanding and interpretation of variability in glucose tracings. Although other longitudinal models can accommodate nested designs, the stochastic processes being used in these models typically do not produce the same degree of predictive accuracy [26]. Further, improved accuracy of prediction of real-time risk would provide a timely warning of severe hyperglycemia or hypoglycemia. Additional future work could be adjusting the FPCA and the prediction models presented in Sect. 3 to incorporate covariates, when available.

Supplementary material. To view supplementary material for this article, please visit <https://doi.org/10.1017/cts.2020.545>.

Acknowledgements. This work was supported by the National Institutes of Health (NIH) under Grants R01 DK109956, K25 HL125954, and R01 HL141286; the Cystic Fibrosis Foundation (CFF) under Grant CLANCY15R0. The authors thank the Jaeb Center for Health Research (JCHR) for making available the data to conduct this study. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH, CFF, or JCHR.

Disclosures. The authors have no conflict of interest to declare.

References

1. Beck RW, Calhoun P, Kollman C. Use of continuous glucose monitoring as an outcome measure in clinical trials. *Diabetes Technology & Therapeutics* 2012; **14**(10): 877–882. doi: [10.1089/dia.2012.0079](https://doi.org/10.1089/dia.2012.0079).
2. Service FJ, et al. Mean amplitude of glycemic excursions, a measure of diabetic instability. *Diabetes*. 1970; **19**: 644–655.
3. Szczesniak RD, et al. Longitudinal patterns of glycemic control and blood pressure in pregnant women with type 1 diabetes mellitus: phenotypes from functional data analysis. *The American Journal of Perinatology* 2016; **33**: 1282–1290.
4. Froslic KF, et al. Shape information from glucose curves: functional data analysis compared with traditional summary measures. *BMC Medical Research Methodology* 2013; **13**: 6. doi: [10.1186/1471-2288-13-6](https://doi.org/10.1186/1471-2288-13-6).
5. Law GR, et al. Analysis of continuous glucose monitoring in pregnant women with diabetes: distinct temporal patterns of glucose associated with large-for-gestational-age infants. *Diabetes Care*. 2015; **38**(7): 1319–1325. doi: [10.2337/dc15-0070](https://doi.org/10.2337/dc15-0070).
6. Barrett JK, et al. Estimating the association between blood pressure variability and cardiovascular disease: An application using the ARIC Study. Version 2. *Statistics in Medicine* 2019; **38**(10): 1855–1868. doi: [10.1002/sim.8074](https://doi.org/10.1002/sim.8074).
7. Everitt BS. The analysis of repeated measures: a practical overview with examples. *The Statistician*. 1995; **44**: 113–135.
8. Little RJA, Rubin DB. *Statistical analysis with missing data*. Vol. 2. Hoboken, N.J: Wiley, 2002.
9. Ramsay JO, Silverman BW. *Functional data analysis*. 2. New York: Springer, 2005.
10. James GM, Hastie TJ, Sugar CA. Principal component models for sparse functional data. *Biometrika* 2000; **87**: 587–602.
11. Szczerbinski L, et al. Clusters of glycemic response to oral glucose tolerance tests explain multivariate metabolic and anthropometric outcomes of bariatric surgery in obese patients. *Journal of Clinical Medicine* 2019; **8**(8): 1091. doi: [10.3390/jcm8081091](https://doi.org/10.3390/jcm8081091).
12. Kahkoska AR, et al. Characterizing the weight-glycemia phenotypes of type 1 diabetes in youth and young adulthood. *BMJ Open Diabetes Research and Care* 2020; **8**(1): e000886. doi: [10.1136/bmjdr-2019-000886](https://doi.org/10.1136/bmjdr-2019-000886).
13. Bian J, et al. A survey on trajectory clustering analysis (2018). ArXiv, abs/1802.06971.
14. Hall H, et al. Glucotypes reveal new patterns of glucose dysregulation. *PLOS Biology* 2018; **16**(7): e2005143. doi: [10.1371/journal.pbio.2005143](https://doi.org/10.1371/journal.pbio.2005143).
15. de Jong J, et al. Deep learning for clustering of multivariate clinical patient trajectories with missing values. *Gigascience* 2019; **8**(11): giz134. doi: [10.1093/gigascience/giz134](https://doi.org/10.1093/gigascience/giz134).
16. Chen K, Muller, HG. Modeling repeated functional observations. *Journal of the American Statistical Association* 2012; **117**(500): 1599–1609.
17. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982; **38**: 963–974. doi: [10.2307/2529876](https://doi.org/10.2307/2529876).
18. Vandyke RD, et al. Characterizing maternal glycemic control: a more informative approach using semiparametric regression. *Journal of Maternal-Fetal and Neonatal Medicine* 2012; **25**(1): 15–19.
19. Gupta R, et al. Glycemic excursions in type 1 diabetes in pregnancy: a semi-parametric statistical approach to identify sensitive time points during gestation. *Journal of Diabetes Research* 2017: 1–7.
20. Rodríguez-Rodríguez I, et al. Utility of big data in predicting short-term blood glucose levels in type 1 diabetes mellitus through machine learning techniques. *Sensors (Basel)* 2019; **19**(20): 4482. doi: [10.3390/s19204482](https://doi.org/10.3390/s19204482).
21. Li K, et al. Convolutional recurrent neural networks for glucose prediction. *IEEE Journal of Biomedical and Health Informatics* 2020; **24**(2): 603–613. doi: [10.1109/jbhi.2019.2908488](https://doi.org/10.1109/jbhi.2019.2908488).
22. Pérez-Gandía C, et al. Artificial neural network algorithm for online glucose prediction from continuous glucose monitoring. *Diabetes Technology & Therapeutics* 2010; **12**(1): 81–88. doi: [10.1089/dia.2009.0076](https://doi.org/10.1089/dia.2009.0076).

23. **Ngufor C, et al.** Mixed effect machine learning: A framework for predicting longitudinal change in hemoglobin A1c. *Journal of Biomedical Informatics* 2019; **89**: 56–67. doi: [10.1016/j.jbi.2018.09.001](https://doi.org/10.1016/j.jbi.2018.09.001).
24. **Connolly B, et al.** A nonparametric Bayesian method of translating machine learning scores to probabilities in clinical decision support. *BMC Bioinformatics* 2017; **18**(1): 361. doi: [10.1186/s12859-017-1736-3](https://doi.org/10.1186/s12859-017-1736-3).
25. **Dmitrienko A, Wang MD.** Bayesian predictive approach to interim monitoring in clinical trials. *Statistics in Medicine* 2006; **25**: 2178–2195. doi: [10.1002/sim.2204](https://doi.org/10.1002/sim.2204).
26. **Diggle PJ, Sousa I, Asar O.** Real-time monitoring of progression towards renal failure in primary care patients. *Biostatistics* 2015; **16**: 522–536. doi: [10.1093/biostatistics/kxu053](https://doi.org/10.1093/biostatistics/kxu053).
27. **Szczesniak RD, et al.** Dynamic predictive probabilities to monitor rapid cystic fibrosis disease progression. *Statistics in Medicine* 2020; **39**(6): 740–756. doi: [10.1002/sim.8443](https://doi.org/10.1002/sim.8443).
28. **Group JCS.** JDRF randomized clinical trial to assess the efficacy of real-time continuous glucose monitoring in the management of type 1 diabetes: research design and methods. *Diabetes Technology & Therapeutics* 2008; **10**(4): 310–321.
29. **Szczesniak RD, et al.** Phenotypes of rapid cystic fibrosis lung disease progression during adolescence and young adulthood. *American Journal of Respiratory and Critical Care Medicine* 2017. doi: [10.1164/rccm.201612-2574OC](https://doi.org/10.1164/rccm.201612-2574OC).
30. **Yao F, Muller HG, Wang JL.** Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* 2005; **100**: 577–590.
31. **Greven S, et al.** Longitudinal principal component analysis. *Electronic Journal of Statistics* 2010; **4**: 1022–1054.
32. **Asar Ö, Bolin D, Diggle PJ, Wallin J.** Linear mixed effects models for non-Gaussian continuous repeated measurement data. *Journal of the Royal Statistical Society: Series C* 2020; **69**: 1015–1065. doi: [10.1111/rssc.12405](https://doi.org/10.1111/rssc.12405).
33. **Taylor-Robinson D, et al.** Understanding the natural progression in % FEV1 decline in patients with cystic fibrosis: a longitudinal study. *Thorax* 2012; **67**: 860–866. doi: [10.1136/thoraxjnl-2011-200953](https://doi.org/10.1136/thoraxjnl-2011-200953).
34. **Chen R, et al.** Continuous glucose monitoring for the evaluation and improved control of gestational diabetes mellitus. *Journal of Maternal-Fetal and Neonatal Medicine* 2003; **14**(4): 256–260.
35. **Monzon AD, et al.** Associations between objective sleep behaviors and blood glucose variability in young children with type 1 diabetes. *Annals of Behavioral Medicine: A Publication of the Society of Behavioral Medicine*. 2020. doi: [10.1093/abm/kaaa040](https://doi.org/10.1093/abm/kaaa040).
36. **Raj S, et al.** My blood sugar is higher on the weekends: Finding a role for context and context-awareness in the design of health self-management technology. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ACM. 119.
37. **Gupta R, et al.** Assessing the relationship between gestational glycemic control and risk of preterm birth in women with type 1 diabetes: A joint modeling approach. *Journal of Diabetes Research* 2020; **2020**: 3,074,532. doi: [10.1155/2020/3,074,532](https://doi.org/10.1155/2020/3,074,532).
38. **Fonda SJ, Lewis DG, Vigersky RA.** Minding the gaps in continuous glucose monitoring: a method to repair gaps to achieve more accurate glucometrics. *Journal of Diabetes Science and Technology* 2013; **7**(1): 88–92. doi: [10.1177/193,229,681,300,700,110](https://doi.org/10.1177/193,229,681,300,700,110).
39. Juvenile Diabetes Research Foundation Continuous Glucose Monitoring Study Group, **et al.** Continuous glucose monitoring and intensive treatment of type 1 diabetes. *The New England Journal of Medicine* 2008; **359**(14): 1464–1476. doi: [10.1056/NEJMoa0805017](https://doi.org/10.1056/NEJMoa0805017).