Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

A personalized prediction of longitudinal growth using People-Like-Me methods

Xin Jin [a], Elizabeth Juarez-Colunga [a,g,*], Stef van Buuren [b], Kathryn Colborn [c], Margaret Rosenfeld [d], Jeremy Graber [e,f], Jennifer Stevens-Lapsley [f,g]

[a] Department of Biostatistics and Informatics, Colorado School of Public Health, CO, United States
[b] Department of Methodology and Statistics, Utrecht University, The Netherlands
[c] School of Medicine, University of Colorado, CO, United States
[d] Department of Pediatrics, Seattle Children's Hospital in University of Washington, WA, United States
[e] Denver/Seattle Center of Innovation for Veteran-Centered and Value-Driven Care, VA Eastern Colorado Health Care System, CO, United States
[f] Department of Physical Medicine and Rehabilitation, University of Colorado, CO, United States
[g] VA Eastern Colorado Geriatric Research, Education, and Clinical Center (GRECC), VA Eastern Colorado Health Care System, CO, United States

ARTICLE INFO

Keywords:
Longitudinal predictive methods
Brokenstick model
Curve matching
Children's growth
Mahalanobis distance

ABSTRACT

Conventional prediction models typically emphasize overall trends, often missing the variability in individual responses, which limits their effectiveness for personalized predictions. The People-Like-Me (PLM) methodology was developed to address this issue by employing curve-matching techniques to generate individualized predictions. PLM is a data-driven algorithm that selects similar matching trajectories to estimate the trajectory for an out-of-sample target. In this study, we extend the PLM methods by introducing the Mahalanobis distance as a new metric for selecting matches, allowing for the consideration of correlations between time points in longitudinal data. We assess the performance of this enhanced PLM across various scenarios using clinical growth data from children with cystic fibrosis and simulated datasets. Our analysis compares (i) different match selection strategies, and (ii) PLM predictions with those from linear mixed models (LMM). The results consistently show that Mahalanobis-based PLM outperforms both the standard PLM and LMM. This establishes Mahalanobis-based PLM as a more accurate and flexible method for personalized prediction of longitudinal trajectories.

1. Introduction

Personalized prediction has received increasing attention in clinical research over the past few decades [1–3]. This approach has been applied across various medical fields, including nephrology [4], psychiatry [5], and physical therapy [6]. Our research group has utilized personalized prediction, based on a data-driven algorithm referred to as People-Like-Me (PLM). Intuitively, the idea of PLM is to build a personalized prediction model using only matched individuals. The PLM method finds the matches by calculating the distance (absolute difference) between two individuals at one time point and identifying individuals with the smallest distance. Then, in the spirit of predictive mean matching, the PLM computes predictions with only observed data of the identified matches. Our clinical collaborators are interested in testing this approach in clinical practice across multiple clinics in the United States. We believe that given that this approach is based on a data-driven algorithm that depends on several user-selected characteristics, such as at which time point to calculate the distance, it is important to improve and to assess how these user-selected choices impact its performance and how this approach compares to a more standard statistical model.

PLM has been used as a decision-support tool to monitor functional recovery following total knee arthroplasty [6,7]. PLM is based on curve matching, first introduced conceptually by [8], which involves selecting matches or donors whose trajectories closely resemble the target individual, and using these similar trajectories to generate personalized predictions. The concept is derived from the robust method of predictive mean matching (PMM), commonly employed for multiple imputation [9]. PMM is a data-driven approach, which selects a subset of observations with similar predictive means to the missing data, and uses those values to impute the missing data points. Curve matching leverages the principles of PMM from the multiple imputation framework to guide individualized predictions. A key decision in curve matching is the method used to select matches, which involves

* Corresponding author.
E-mail address: elizabeth.juarez-colunga@cuanschutz.edu (E. Juarez-Colunga).

https://doi.org/10.1016/j.compbiomed.2025.111281
Received 2 December 2024; Received in revised form 26 September 2025; Accepted 3 November 2025
Available online 13 November 2025

considerations such as the choice of distance metric for comparing curves and the potential inclusion of covariates. The standard PLM found matches based on a single time point. In this paper, we propose a novel, enhanced PLM data-driven algorithm that incorporates multiple-time matching and the application of Mahalanobis distance in the matching process. While taking into account the correlation between multiple longitudinal time points, the Mahalanobis distance measures the distance between a specific point and a distribution, or between a point and the mean of the distribution [10]. We hypothesize that the enhanced PLM with multiple-time matching and Mahalanobis distance will improve predictive performance. We apply the proposed PLM methods to a growth study of children with cystic fibrosis.

Cystic fibrosis (CF) is a genetic disorder that leads to significant damage in the lungs, digestive system, and other organs throughout the body. CF is associated with a broad spectrum of diseases related to exocrine dysfunction, including chronic respiratory bacterial infections and reduced life expectancy [11]. Children with CF often experience growth deficiencies resulting from malabsorption, reduced caloric intake, increased resting energy expenditure, and other metabolic challenges [12,13]. These growth-related complications can have long-term health implications. To evaluate the utility and robustness of our personalized prediction approach, we demonstrate the performance of the PLM method to predict growth patterns using both real-world growth data from children with CF and simulated datasets that mimic real growth patterns.

Methodological approaches employed for personalized predictions include methods based on a combination of models, including meta-analysis and Bayesian model averaging [14–16]; regression models [17,18]; and algorithmic-based approaches, including random forests and deep learning [2,19]. Machine learning methods encompass both regression models and algorithmic-based approaches [20–22]. However, these methods have several important limitations. First, this often requires an exhaustive search in the feature space or time-consuming expert involvement to generate statistically meaningful features. Second, there is the potential loss of useful information if the wrong features are removed. A mis-specification of the features can result in overlooking important aspects of the data, especially if there are limited available features in the dataset. Third, most of the mentioned approaches do not address out-of-sample personalized prediction, i.e., predictions for individuals not included in the training data.. In particular, it is challenging to provide a personalized prediction for out-of-sample subjects in longitudinal linear mixed models (LMM) unless partial information is available as in the personalized predictions from LMM [17]. When making predictions for a new subject using a longitudinal model, the fitted random effects from the training set provide no useful information. The random effects estimated from the training data cannot directly inform predictions for new subjects, because we cannot infer the new subject's information based on the population-level response (i.e., their conditional random effects). The new subject can deviate from the population-level response. Therefore, the best prediction LMM can offer for a new subject is the population-level response, which limits the model's ability to provide accurate individual predictions.

However, if partial information on the longitudinal trajectory is available, predicted personalized trajectories can be built within the framework of longitudinal LMM [17,18]. The methodology begins by constructing a LMM, typically trained on a reference population or dataset. It will predict the trajectory of a new individual with such a trained model and partial information, such as baseline outcomes and features. The model updates predictions as more data become available, improving accuracy over time. A random effect for the new individual is then estimated, conditional on their partial data and the training model, and this estimate is used to predict their personalized trajectory. Here, we compare the performance of the People-Like-Me (PLM) methods with that of LMM-based personalized prediction, using baseline outcome information from the new individual.

In this study, we propose a novel enhanced PLM data-driven approach that extends the PLM methodology to use the Mahalanobis distance based on multiple time points for matching. We also compare different PLM setups to predictions derived from LMM. We illustrate this comparison through the analysis of the Early Pseudomonas Infection Control (EPIC) study data and simulation studies. This article is organized as follows. In Section 2, we outline the three steps of the PLM method, namely: two-stage model training, selecting matches, and predicting based on a generalized additive model for location, scale, and shape. We report the results from the EPIC analysis in Section 3, and the results from the simulation studies in Section 4. In these two sections, we focus on the performance of PLM under different scenarios, including various sets of times at which distances are calculated, and matching criteria; and we also compare PLM to LMM. We conclude with a discussion in Section 5 of the limitations of PLM, as well as future research directions.

2. Methods

The fundamental idea of PLM is to fit a personalized model using a subset of the data, with only matched individuals, as shown in Fig. 1. The PLM method involves three steps: (i) Training a two-stage model, (ii) Finding matches, and (iii) Predicting with the matching subset. In the first step, stage 1 model uses a LMM to build predictions at a set of predefined time points, referred as *anchor time points*; in stage 2 model, we refine those predictions through a linear model. Then, in step 2, we calculate the distance between the target individual and training set and select the closest matching trajectories. Lastly, in step 3, we fit a prediction model using only the selected matches. In the standard PLM method, we similarly follow three steps, but calculate the distance at a single time point and find matches based on this absolute difference. Below we provide our proposed enhanced PLM with multiple-time prediction and matching criteria based on Mahalanobis distance. At the end of this section, we provide a discussion of alternative choices that can be made within a curve matching approach.

1. Two-stage model. In the first step, in stage 1 we use a piecewise LMM to build predictions at the anchor time points; and then in stage 2 model, we refine those predictions through a linear model that includes baseline covariates. Longitudinal data are often observed at irregular times, with measurements taken at non-uniform times for individuals. Since the PLM requires predictions at specific anchor times, a method is needed to transform these irregular observations into balanced repeated measures. Here we have chosen to use a specific type of LMM, referred to as the *brokenstick model*. The brokenstick model is a linear mixed model that models time as a piece-wise linear function with the same fixed and random structure [23]. This model has been shown to be flexible in identifying non-linear patterns and to provide accurate predictions [23,24]. Here, no covariates are used, and the only predictor is time; this lack of covariates implies that the transformation of irregular data into repeated measures is identical for every subject, but the random effects allow for individual-specific estimation of trajectories.

Let the observed data in the training set be denoted by $\boldsymbol{S}_{train} = \{\boldsymbol{T}_i, \boldsymbol{X}_i, \boldsymbol{y}_i;\ i = 1, \ldots, N\}$, where $\boldsymbol{y}_i$ is the $n_i \times 1$ vector with entries y_{ij} denoting the response variable for the ith individual on the jth measurement taken at time point $t_{ij},\ j = 1, \ldots, N$; $\boldsymbol{T}_i$ denotes the $n_i \times 1$ vector consisting of every t_{ij}, where n_i is the number of the time points; $\boldsymbol{X}_i$ is the $p \times 1$ vector of p covariates, such as demographic characteristics, which are fixed in time; lastly, y_{i0} is the first outcome of $\boldsymbol{y}_i$ at the baseline. We use B-spline linear basis of degree 1 to implement the piece-wise linear feature of the model (see Web Supplementary Section S13.1 for more details). Specifically, let $\boldsymbol{\Phi}_L(t) = \left(\Phi_{L,0}(t), \ldots, \Phi_{L,e}(t)\right)^\top$ be the B-spline basis function, where $e = k(\text{knots}) + 1(\text{degree of spline}) + 1$ is the number of basis functions. Let the user-defined internal knots be denoted by s_k such that $min(t_{ij}) = s_0 < s_1 < \cdots < s_k < s_{k+1} = \max(t_{ij})$. Then the brokenstick model is written as:

$$\boldsymbol{y}_i = \boldsymbol{\Phi}_L(t_i)^\top \boldsymbol{\beta} + \boldsymbol{\Phi}_L(t_i)^\top \boldsymbol{b}_i + \boldsymbol{\epsilon}_i = \boldsymbol{\Phi}_L(t_i)^\top (\boldsymbol{\beta} + \boldsymbol{b}_i) + \boldsymbol{\epsilon}_i, \tag{1}$$

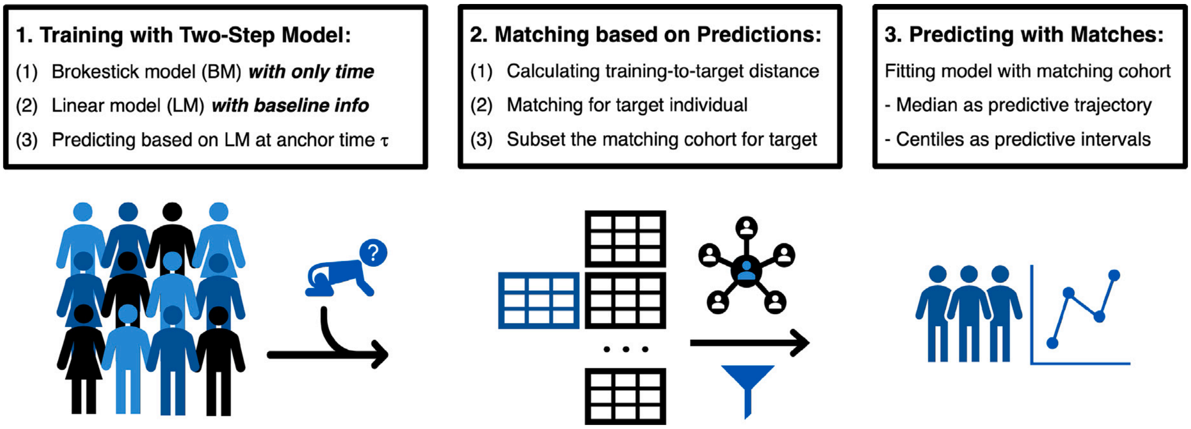
1. Training with Two-Step Model:
(1) Brokestick model (BM) with only time
(2) Linear model (LM) with baseline info
(3) Predicting based on LM at anchor time τ
2. Matching based on Predictions:
(1) Calculating training-to-target distance
(2) Matching for target individual
(3) Subset the matching cohort for target
3. Predicting with Matches:
Fitting model with matching cohort
- Median as predictive trajectory
- Centiles as predictive intervals
?

Fig. 1. Illustration of PLM steps predicting for a new target individual, with only baseline information. Three steps encompass PLM method: 1. Fitting a two-stage model: with training group data as reference, we fit a brokenstick model to impute the outcome at the anchor times from observations at irregular times (note the brokenstick model is only fit to the training dataset); then we fit a linear model to add variables and the outcomes are predictions at the anchor times; we will predict the target outcomes only use the baseline information and the first outcome from this linear model. 2. Matching for the target individual: with the prediction for both training set and target individual, we use Mahalanobis distance at multiple anchor times with to find the matches for the target individual. 3. Predicting: we use a flexible model with selected matches from the step 2. Here we used GAMLSS model. The medians from model are used as predictive trajectories and the centiles are used for predictive intervals. These are presented with the blue colored areas.

where $e \times 1$ vector $\boldsymbol{\beta}$ denotes the fixed effect, and $\boldsymbol{b}_i$ the conditional random effect, with $\boldsymbol{b}_i \sim N(\mathbf{0}, \boldsymbol{G}_0)$, and $\boldsymbol{G}_0$ the covariance matrix for random effects; vector $\boldsymbol{\epsilon}_i$ contains the residuals for each individual i with $\boldsymbol{\epsilon}_i \sim N(0, \sigma_\epsilon^2 \boldsymbol{I}_{n_i})$, where $\boldsymbol{I}_{n_i}$ is a $n_i \times n_i$ identity matrix. Let $\boldsymbol{\tau} = \{\tau_1, \dots, \tau_m\}$ denote the set of anchor times, which will be common to all individuals. Let $\dot{\boldsymbol{y}}_i = (\dot{y}_{i1}, \dots, \dot{y}_{im})$ denote the predictions obtained from the brokenstick model at the anchor times for individual i, $\dot{\boldsymbol{y}}_i = \boldsymbol{\Phi}_L(t_i)^\top(\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{b}}_i)$, $i = 1, \dots, N$. In the standard PLM, these predictions are obtained for all individuals at a single anchor time, say τ_1.

In the second stage, we incorporate time-invariant covariates X_i and the baseline outcome y_{i0} into a linear model to predict the outcome $\dot{\boldsymbol{y}}_i = (\dot{y}_{i1}, \dots, \dot{y}_{im})$. For a given individual, we use the brokenstick predicted value as the response variable as follows:

$$\dot{\boldsymbol{y}}_i = y_{i0}\gamma_0 + \boldsymbol{X}_i \boldsymbol{\gamma}_X + \dot{\boldsymbol{\epsilon}}_i, \tag{2}$$

where $\boldsymbol{X}_i$ is the augmented $m \times q$ design matrix for all q fixed covariates for each anchor time; $\boldsymbol{\gamma} = (\gamma_0, \boldsymbol{\gamma}_X^\top)^\top$, the $(q+1) \times 1$ vector, contains the coefficients for the effect of baseline response and covariates; $\dot{\boldsymbol{\epsilon}}_i$ denotes the residuals, $\dot{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma_{\dot{\epsilon}}^2 \boldsymbol{I}_{m\times m})$, where $\boldsymbol{I}_m$ is $m \times m$ identity matrix.

Let $\ddot{\boldsymbol{y}}_i = (\ddot{y}_{i1}, \dots, \ddot{y}_{im})$ denote the prediction from the stage 2 model, specifically $\ddot{\boldsymbol{y}}_i = \boldsymbol{X}_i \hat{\boldsymbol{\gamma}}_X + y_{i0}\hat{\gamma}_0$. This linear model is used to define a prediction for the target individual. For a new target individual with fixed characteristics $\boldsymbol{X}_*$ and a baseline observation value y_{*0}, we define the predicted outcome vector at the anchor times as $\ddot{\boldsymbol{y}}_* = \boldsymbol{X}_* \hat{\boldsymbol{\gamma}}_X + y_{*0}\hat{\gamma}_0$. Note that each anchor time point is treated as an independent factor variable, as a separate linear model for each anchor time point. These predictions $\ddot{\boldsymbol{y}}_i = (\ddot{y}_{i1}, \dots, \ddot{y}_{im})$ will be used to calculate distance and find matches in the next step.

The choice of a linear model is motivated by three reasons: first, the brokenstick predictions primarily capture local trends that are highly dependent on data quality from adjacent intervals [23]. Second, the brokenstick model cannot account for individual-level characteristics, so adding a linear model allows for more personalized predictions and helps mitigate potential biases inherent in the local estimates from the brokenstick model. Notably, this is the only step in which we incorporate the target individual's information.

2. Matching. In this step, we calculate the distances between the target $\ddot{\boldsymbol{y}}_*$ and every individual $\ddot{\boldsymbol{y}}_i$ in the training set $\boldsymbol{S}_{train}$, $i = 1, \dots, N$. Let D_M denote the Mahalanobis distance between predictions of the target individual $\ddot{\boldsymbol{y}}_*$ and individual i in the training set $\ddot{\boldsymbol{y}}_i$, specifically $D_M = \sqrt{(\ddot{\boldsymbol{y}}_i - \ddot{\boldsymbol{y}}_*)^\top \boldsymbol{\Sigma}^{-1} (\ddot{\boldsymbol{y}}_i - \ddot{\boldsymbol{y}}_*)}$; $\boldsymbol{\Sigma}$ corresponds to a covariance matrix of the predicted values $\ddot{\boldsymbol{y}}_i$. Assuming independence between these predictions $\ddot{\boldsymbol{y}}_i$, the squared Mahalanobis distance follows a chi-squared distribution, i.e. $D_M^2 \sim \chi_m^2$. In our case, given that $\ddot{\boldsymbol{y}}_i$ have been obtained as predictions from a model, they are not independently distributed.

In the PLM methodology, this assumption is applied solely for the purpose of computing distances between candidate matches, not for inference, but to calculate predictions. While the predicted values are not strictly independent, the Mahalanobis distance remains a practical metric for multivariate similarity, as it accounts for correlations between variables through the covariance matrix. For the purposes of selecting candidate matches, we assume independence and use chi-squared quantiles to tune this parameter. Then a matching set A_α of matches, selected from the training dataset for $\boldsymbol{y}_*$ is defined as:

$$A_\alpha = \{i : D_M^2 = (\ddot{\boldsymbol{y}}_i - \ddot{\boldsymbol{y}}_*)^\top \boldsymbol{\Sigma}^{-1} (\ddot{\boldsymbol{y}}_i - \ddot{\boldsymbol{y}}_*) \leq \chi^2_{m,\alpha}, i \in \boldsymbol{S}_{train}\}$$

with α as the right-tail rejection region of the chi-squared distribution. In this context, a larger α corresponds to a shorter distance with a smaller cutoff, producing a stricter matching criterion and fewer accepted matches. α will be defined by the user according to the level of matching required. A larger α (rejection) value makes the match selection stricter, as demonstrated in Supplementary Figure S3 of the data analysis.

In this paper, we extend the standard PLM from single-time matching into multiple-time matching. Single-time prediction with a fixed number of matches, will be denoted as S_κ. In this case, the closest κ matches are selected based on a single anchor time point. Multiple-time matching with Mahalanobis distance of a critical value α will be denoted as M_α. Multiple anchor points, especially at crucial turning points, capture dynamic trends and subtle trajectory changes that single-time matching can miss. This directly addresses the limitation of relying solely on one time point and allows for the detection of important inflection points. Mahalanobis distance further incorporates the covariance structure across time, capturing multivariate relationships and accounting for both the direction and variability of change. Together, these advantages provide a more robust and informed basis for personalized prediction.

3. Predicting with Subset. In this last step of PLM, we fit a model that provides flexible and robust predictions for the target individual based on the selected matches. With the matching subset from step 2 for the target individual, the user can choose a flexible parametric or nonparametric model for the final prediction. We have chosen models within the framework of Generalized Additive Model for Location, Scale, and Shape (GAMLSS) models [25,26], because these encompass a wide range of distributions and can capture non-linear trends well. Briefly, GAMLSS models go beyond the traditional modeling of mean by modeling the response variable with separate parameters for scale (variance) and shape (skewness and kurtosis)[26]. This framework supports a wide range of distributions and the ability to model multiple distributional parameters simultaneously. Linear and non-linear predictors can be handled in the estimation of each of the parameters including location, scale, skewness and kurtosis (See Supplementary Section S13.2 for more details). As commonly done when using GAMLSS, we use the median as the predicted trajectory for the target individual, and appropriate quantiles as the predicted interval.

A few points are worth noting regarding the M_α or α criterion. First, the number of matches determined by the α criterion is personalized and varies by individual, providing flexibility in match selection. Second, if α is too small, it will include almost all available matches; however, this does not necessarily improve predictive accuracy, as the prediction for each individual would essentially be the overall mean across all individuals. Third, if α is too large, it can result in very few or even no matches, leading to biased or potentially unreliable estimates, or no prediction at all. In this paper, we examined multiple α values in a wide range. This will be demonstrated in the data analysis in Section 3.

2.1. Alternative options in steps of PLM

We acknowledge that when utilizing PLM there are several choices that need to be made by the user: knots for the brokenstick model, anchor time points, covariates to use in the linear model, prediction model based on matches. Although we consider a range of options, we would like to offer some insight on our decisions.

1. Why multiple steps? It is essential to establish a method for selecting curves closely aligned with the target individual, enabling personalized prediction using observed data in the spirit of predictive mean matching. Consequently, the components of the brokenstick model, the anchor time points, and the linear model all get at the question of curve selection. Across these components, we have explored various options including distinct numbers and placements of knots within the brokenstick model, different sets of anchor time points, and a selection of covariates as predictors. To compare with a prediction model that does not contain multiple steps, we compare this data-driven algorithmic approach to personalized predictions from LMM using only the baseline outcome y_{i0} [17].
2. Why not combine the two-stage model into one linear mixed model? We believe the brokenstick model is a good option for capturing the individual trajectories over time, and it would be great to incorporate covariates in a single step rather than performing two steps. The current implementation of the brokenstick model does not support the inclusion of covariates other than time. We are interested in pursuing this extension in future work.
3. Why use GAMLSS for prediction? Other models or approaches are possible too, including LMM with flexible predictors of time, functional models such as generalized additive models, or nonparametric methods such as local regression. We have chosen to use GAMLSS for specific reasons related to our children's growth dataset: (1) GAMLSS is a natural extension of semi-parametric generalized additive models and includes a broad range of distributions; (2) GAMLSS can fit the model with smooth functions for mean, variance, and shape of the distribution providing flexibility in modeling; and (3) GAMLSS allows for inclusion of predictors and interactions.

Algorithm 1 People-Like-Me (PLM) prediction for $\boldsymbol{y}_*$

```
Inputs:
    Training set $S_{\text{train}} = \{(\boldsymbol{T}_i, \boldsymbol{X}_i, \boldsymbol{y}_i) : i = 1, \dots, N\}$, where $\boldsymbol{T}_i = (t_{i1}, \dots, t_{in_i})$
    Target (out-of-sample) $S_{\text{test}} = \{(\boldsymbol{X}_*, y_{*0})\}$ with baseline features $\boldsymbol{X}_*$ and baseline outcome $y_{*0}$
    Anchor times $\boldsymbol{\tau} = (\tau_1, \dots, \tau_m)$
    Quantile set $\mathcal{Q}$ for prediction intervals (e.g., $\{0.025, 0.5, 0.975\}$)
Outputs: $\hat{\boldsymbol{y}}_*(t)$ for any future time $t \in \mathcal{T}$ and point-wise PIs from $\mathcal{Q}$
Step 1: Fitting two-stage model
Model 1: fitting a brokenstick model:
    $\boldsymbol{y}_i = \boldsymbol{\Phi}(\boldsymbol{T}_i)^\top(\boldsymbol{\beta} + \boldsymbol{b}_i) + \boldsymbol{\epsilon}_i, \quad \boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}),\ \boldsymbol{\epsilon}_i \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}).$
Obtain $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{b}}_i$ for all $i$.
Predict each subject's outcomes at anchor time points:
    $\dot{\boldsymbol{y}}_i \leftarrow \boldsymbol{\Phi}(\boldsymbol{\tau})^\top \left(\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{b}}_i\right) \in \mathbb{R}^m.$
Model 2: fitting linear regression at each anchor time point:
    $\dot{\boldsymbol{y}}_i = y_{i0}\,\gamma_0 + \boldsymbol{X}_i \boldsymbol{\gamma}_X + \dot{\boldsymbol{\epsilon}}_i.$
Estimate $\hat{\boldsymbol{\gamma}}_X, \hat{\gamma}_0$ and the covariance $\hat{\boldsymbol{\Sigma}} = \mathrm{Cov}(\dot{\boldsymbol{y}}_i)$.
Compute fitted anchor times for all donors and the target:
    $\ddot{\boldsymbol{y}}_i \leftarrow \boldsymbol{X}_i \hat{\boldsymbol{\gamma}}_X + y_{i0}\,\hat{\gamma}_0, \quad \ddot{\boldsymbol{y}}_* \leftarrow \boldsymbol{X}_* \hat{\boldsymbol{\gamma}}_X + y_{*0}\,\hat{\gamma}_0.$
Step 2: Match selection in anchor space
for $i = 1$ to $N$ do
    $D_M(i, *) \leftarrow \sqrt{\left(\ddot{\boldsymbol{y}}_i - \ddot{\boldsymbol{y}}_*\right)^\top \hat{\boldsymbol{\Sigma}}^{-1} \left(\ddot{\boldsymbol{y}}_i - \ddot{\boldsymbol{y}}_*\right)}$
end for
Find the matching set: $\mathcal{A}_\alpha \leftarrow \{i : D_M^2(i, *) \le \chi^2_{m,\alpha}\}$
Step 3: Trajectory prediction from matched donors
Define a prediction time $\mathcal{T}$ (e.g., future ages or follow-up times of interest)
for each $t \in \mathcal{T}$ do
    Collect donor outcomes $\{y_i(t) : i \in \mathcal{A}_\alpha\}$
    Point prediction: $\hat{y}_*(t) \leftarrow \mathrm{median}\{y_i(t) : i \in \mathcal{A}_\alpha\}$
    for each $q \in \mathcal{Q}$ do
        $\hat{y}_*^{(q)}(t) \leftarrow q$-th empirical quantile of $\{y_i(t) : i \in \mathcal{A}_\alpha\}$
    end for
end for
return $\{\hat{y}_*(t),\ [\hat{y}_*^{(q_\ell)}(t), \hat{y}_*^{(q_u)}(t)]\ \forall t \in \mathcal{T}\}$
```

4. Finally, we note that the PLM methodology entails several user-defined choices, yet there is substantial interest from clinical collaborators in using this model in clinical practice [6,7]. We have attempted to analyze a range of competitive options to assess its performance and its comparison to other approaches. This has been illustrated in the application example in Section 3.

2.2. Predictions from LMM

Individualized predictions from linear mixed models refer to the process of predicting outcomes for a new individual, provided at least one data point is available for the new patient [17]. With this information, individualized predictions are calculated. In this work, we compare the PLM to this personalized LMM prediction methodology. To ensure a fair comparison, we restrict the LMM predictions to use baseline data only, as the PLM model is also limited to baseline information for predicting trajectories. Specifically, we aim to derive predictions for a new individual l who is assumed to follow the same population-average trend as the training data. This individual provides fixed covariate information and a subset of observed longitudinal outcomes, denoted by $\boldsymbol{y}_l = \{y_l(t_{lj}); 0 \le t_{lj} \le t_0\}_j^{n_l}$. In our case, the observed longitudinal outcome sequence includes only baseline data, so $n_l = 1$.

Given the observed baseline outcome $y_l(t_{l0})$, our goal is to predict the outcome at a future time t with $t > t_0$. The LMM for the outcome $y_i(t)$ of individual i, conditional on the random effects vector $\boldsymbol{c}_i$, is specified as: $E(y_i(t)) = X_i'\boldsymbol{\eta} + Z_i'\boldsymbol{c}_i$, where X_i and Z_i are the fixed and random covariate matrices, and $\boldsymbol{\eta}$ and $\boldsymbol{c}_i$ are the fixed and random effect vectors, respectively. Let θ denote the full vector of model parameters, estimated from the training dataset used to develop the model.

A main challenge is to propose an estimator for the individual-specific random effects $\boldsymbol{c}_l$ [27,28]. The conditional distribution of the random effects, $p(\boldsymbol{c}_l \mid \boldsymbol{y}_l, \theta)$, given the observed outcomes $\boldsymbol{y}_l$, covariates, and the training data, will be used for this purpose [17,27,28]. This distribution is derived from: $p(\boldsymbol{c}_l \mid y_l(t_0), \theta) = p(y_l(t_0) \mid \theta, \boldsymbol{c}_l)\, p(\boldsymbol{c}_l \mid \theta)$. Monte Carlo samples of θ are employed to estimate:

$$E[y_l(t) \mid Y_l(t_0), \theta] = \int E[y_l(t) \mid \boldsymbol{c}_l, \theta]\, p(\boldsymbol{c}_l \mid Y_l(t_0), \theta)\, d\boldsymbol{c}_l, \qquad (3)$$

where E denotes the expectation, and $p(\boldsymbol{c}_l \mid Y_l(t_0), \theta)$ represents the posterior predictive conditional distribution of the random effects given the historical data and model parameters.

2.3. Implementation

All analyses were conducted in R (V.4.2.2, available at https://cran.r-project.org). The “brokenstick” package [23] was employed to fit the brokenstick model. Longitudinal data analysis was conducted using the “nlme” package [29]. For modeling flexible distributions of variables, we utilized the “gamlss” package [30] with a wide range of distributional models. The “JMbayes” package [27] was employed to obtain predictions based on LMM. Model specifications can be found in Supplementary Section S1 and detailed code may be obtained from the corresponding author upon request.

3. EPIC analysis

The Early Pseudomonas Infection Control (EPIC) Observational Study is a prospective, multi-center longitudinal study that enrolled children under 12 diagnosed with CF who tested negative, for at least 2 years, for Pseudomonas aeruginosa culture [31]. Between 2004 and 2006, the study enrolled 1,772 participants [32]. Eligible participants for this analysis were 1,370 individuals who enrolled no younger than age 3 with more than 10 study visits and at least 5 years of follow-up. Since we needed a common starting time for all individuals, we aligned all growth trajectories to start at baseline (time 0). We took individuals whose age was between 3 and 4 and aligned them all at 0. Data were split randomly into training (67% of individuals) and testing (33% of individuals) sets. Demographic characteristics of individuals in the training (913 individuals) and testing (457 individuals) sets were similar (Table 1). The testing cohort comprised 51% females, with a median age at the first visit of 3.1 years (First and third quartiles Q1-Q3: 3.05-3.22). The median follow-up period was 9.7 years (Q1-Q3: 6.8-13), and the median height at baseline was 94.2 cm (Q1-Q3: 91.5-97).

In this analysis, we applied PLM methods and LMM prediction to the EPIC study dataset. Our focus is on predicting the growth trajectories of children’s heights over time. Because the PLM method involves several choices such as, knots for brokenstick model, number of anchor time points, and the final model for prediction, we wanted to investigate how different choices at each step can affect the final predictive performance. We specifically followed the process of algorithm 1 step by step, and we made choices as follows:

1. Step 1: The first stage (brokenstick model) was a piece-wise linear model with internal knots at 5, 10, and 15 years; the boundary knots were defined by the data, with the beginning of the study at 0 and the longest follow-up time 20 years. Other choices were explored but yielded no improvements, so results are not reported. As outlined in algorithm 1, the second stage linear model used time as a factor variable, as well as interaction terms between baseline outcome, and gender and genotype. This is the only step in the algorithm where covariates are utilized, and through the predictions generated by the linear model, they influence the subsequent matching step. We used the following sets of anchor times:
    a. 3 anchor-time-point sets: $t(4, 8, \mathbf{12})$ and $t(5, \mathbf{10}, 15)$
    b. 4 anchor-time-point sets: $t(3, 6, 9, \mathbf{12})$ and $t(6, 10, 11, \mathbf{12})$
    c. A single anchor-time-point: based on our previous cross-validation study, we considered a single anchor time point at 10 or 12 years, and a matching number of 10 or 30. In our previous work we showed by cross-validation that the best matching number is around 30 [33]; and we also used 10 as a matching number for comparison. Bolded time points denote the single-time matching PLM scenario.

Table 1
Characteristics of EPIC study data. For the gender category, the reference level is ‘Male’; the reference level of Race is ‘Other’ race groups which includes Black, Asian, etc. The reference of ethnicity is ‘Non-Hispanic.’ Categorical variables are reported as n (%), and continuous variables as median (Q1: first quartile, Q3: third quartile).

Characteristics	Testing, N = 457	Training, N = 913
Genotype		
Two alleles F508del	251 (55%)	492 (54%)
One allele F508del	156 (34%)	326 (36%)
Others or Unknown	50 (11%)	95 (10%)
Female	234 (51%)	459 (50%)
Race/ethnicity		
White	434 (95%)	874 (96%)
Hispanic	15 (3.4%)	28 (3.2%)
Visit number	44 (31, 58)	46 (32, 60)
Age at the first visit (years)	3.13 (3.05, 3.22)	3.13 (3.06, 3.22)
Age at the last visit (years)	12.9 (9.9, 16.1)	12.8 (10.3, 15.9)
Follow-up (years)	9.7 (6.8, 13.0)	9.6 (7.2, 12.6)
Height at baseline (cm)	94.2 (91.5, 97.0)	94.0 (91.3, 96.7)

2. Step 2: For the matching step, we used the predictions from step 2 and considered a range of α values including 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, and 0.95, which determined the probability region defined by the Mahalanobis distance χ^2 criteria (M_α).
3. Step 3: For prediction based on the selected matches, we applied the GAMLSS model with the normal distribution family with mean and variance as parameter functions: the mean function was modeled with a cubic smoothing spline with 3 degrees of freedom; the variance function was modeled with a smoothing spline with 1 degree of freedom.
4. Predictions from LMM. We used a function of time and gender as predictors along with their interaction term. Genotype was also included in the model; see Web Supplementary Table S2 for more details on this. A B-spline basis function with a cubic term and three internal knots at 10, 12, and 15 years was used as a function of time, and interactions to allow for an individual B-spline for each gender. Random effects were also included as with B-spline function with a quadratic term and three equally distributed internal knots; see Web Supplementary Table S2 for more details. The final model was selected based on AIC (see Web Supplementary Table S1 for more details).
5. To evaluate the predictive performance of different PLM and LMM methods, we specifically examined the following metrics: the mean absolute error (MAE) to measure the bias between the observed and predicted values; a smaller MAE indicated more

Table 2
EPIC data analysis comparing different methods including: single value matching with 10 ($S_{\kappa=10}$) and 30 matches ($S_{\kappa=30}$); Mahalanobis distance matching with multiple χ^2 critical values for ($M_{\alpha=0.6}$, $M_{\alpha=0.7}$, $M_{\alpha=0.8}$, $M_{\alpha=0.85}$, $M_{\alpha=0.9}$, and $M_{\alpha=0.95}$); and LMM. The single anchor time point used is in bold font. The mean absolute error (MAE), the root mean squared error (RMSE), and coverage rate of 90% predictive interval (90% CR) are reported.

	Time	$S_{\kappa=10}$	$S_{\kappa=30}$	$M_{\alpha=0.6}$	$M_{\alpha=0.7}$	$M_{\alpha=0.8}$	$M_{\alpha=0.85}$	$M_{\alpha=0.9}$	$M_{\alpha=0.95}$	LMM
MAE	t(4, 8, **12**)	3.07	2.96	3.19	3.04	2.91	2.84	2.79	2.74	3.04
	t(5, **10**, 15)	3.12	2.97	3.19	3.04	2.90	2.84	2.78	2.74	
	t(3, 6, 9, **12**)	3.14	2.95	3.48	3.31	3.11	3.03	2.92	2.83	
	t(6, **10**, 11, 12)	3.16	2.94	3.47	3.30	3.12	3.01	2.92	2.82	
RMSE	t(4, 8, **12**)	4.32	4.20	4.32	4.17	4.03	3.99	3.94	3.91	4.17
	t(5, **10**, 15)	4.47	4.22	4.31	4.17	4.04	3.99	3.94	3.90	
	t(3, 6, 9, **12**)	4.44	4.18	4.60	4.44	4.24	4.15	4.05	3.98	
	t(6, **10**, 11, 12)	4.50	4.13	4.60	4.43	4.24	4.15	4.05	3.98	
90%CR	t(4, 8, **12**)	0.84	0.89	0.94	0.94	0.94	0.94	0.94	0.93	0.95
	t(5, **10**, 15)	0.84	0.89	0.94	0.94	0.94	0.94	0.94	0.93	
	t(3, 6, 9, **12**)	0.84	0.90	0.93	0.94	0.94	0.94	0.94	0.94	
	t(6, **10**, 11, 12)	0.83	0.89	0.93	0.94	0.94	0.94	0.94	0.94	

accurate predictions. The root of mean squared error (RMSE) measured the variance of the squared differences between the observed and predicted values, and a smaller RMSE indicated more precise predictive results. Coverage rates (CR) of 50%, 80%, and 90% were included to evaluate the confidence of predictive intervals in capturing true observed values. Coverage rates closer to the given CR indicated better predictive performance; neither under- nor over-coverage was considered ideal.

Table 2 displays the results of the main analyses of the EPIC data. First, we discuss methods with matching criteria selecting the top 10 or 30 smallest distance matches ($S_{\kappa=10}$ & $S_{\kappa=30}$) and then methods that utilize α criterion. Using a larger matching number of κ improves the PLM performance slightly, which is consistent with our previous cross-validation results [33]. When comparing the scenarios varying the α value, ranging from 0.6 to 0.95, the best scenario was $M_{\alpha=0.95}$ with respect to MAE, RMSE and CR and for all anchor time sets. The M_α methods overestimated the 90% CR between 93–94%. With the same number of anchor time points, larger α values yielded better predictive performance of M_α. For instance, as α increased, for the 3-anchor-time sets $t(4, 8, 12)$ and $t(5, 10, 15)$, the MAE decreased from 3.19 ($M_{\alpha=0.6}$) to 2.74 ($M_{\alpha=0.95}$) and the RMSE decreased from 4.3 ($M_{\alpha=0.6}$) to 3.90 ($M_{\alpha=0.95}$). This improvement was attributed to the stricter matching criteria with larger α values, leading to fewer and more similar matches. As we have shown in Supplementary Table S3, the median matching number decreased from 308 ($M_{\alpha=0.6}$) to 74 ($M_{\alpha=0.95}$). We believe the better performance of M_α not only due to the larger number of matches selected but also the flexibility for different matching numbers for different target individuals (seen in Supplementary Table S3).

Comparing M_α to other scenarios, we found that the 3-anchor-time set with $\alpha < 0.65$ or the 4-anchor-time set with $\alpha < 0.8$ generally provided poorer prediction results compared to $S_\kappa = 30$. The performance of M_α was comparable to $S_\kappa = 10$ in MAE and RMSE, at 3-anchor-time set $\alpha < 0.65$ or 4-anchor-time set $\alpha < 0.8$. Both the 3-anchor-time set with $\alpha \approx 0.85$ and the 4-anchor-time set $\alpha \approx 0.95$ provided similar performance and more reliable in terms of MAE and RMSE than both $S_{\kappa=10}$ and $S_{\kappa=30}$. The best results, with respect to MAE, RMSE, and 90% CR, were those with higher values of α, 3-anchor-time set $\alpha > 0.9$ or 4-anchor-time set $\alpha > 0.95$. For M_α, increasing the number of anchor time points, from a 3-anchor-point to a 4-anchor-point, decreased model accuracy and precision, resulting in higher MAE, larger RMSE, and wider coverage intervals. Adding more anchor times increased the degrees of freedom (df) in the χ^2 distribution (see Supplementary Section S3) such that the density function had a heavier tail, indicating an elevated chance of outliers or influential values. This led to a more liberal match selection.

One can use both the number of anchor time points and α to enhance matching selection. More anchor time points can allow for more matches and higher α permits fewer matches. Fig. 2 illustrates the performance of the methods for a particular individual and demonstrates a few important points. The $M_{\alpha=0.8}$ and $M_{\alpha=0.9}$ provided the best prediction followed by $S_{\kappa=30}$. Second, the selection of matches was not symmetric around the target trajectory. For this individual in Fig. 2, the closest matches tended to lie above the target trajectory. Third, for $S_{\kappa=30}$ the way matches are found, by relying on the anchor time point, yielded trajectories that cannot accurately represent the entire target trajectory.

In summary, while increasing the number of anchor times seems like a straightforward way to enhance PLM performance, in practice, this results in poor match selection and reduced PLM prediction performance (in Supplementary Table S4). The α criterion did add flexibility, but it also introduced new complexities and risks, notably the potential for inclusion of fewer similar trajectories. Therefore, careful consideration and validation are needed when choosing the number of anchor time points and the α level.

Comparison to predictions from LMM. Predictions from PLMs with $\alpha \geq 0.85$ outperformed LMM in terms of MAE (3.04 for LMM, and 3.01 for $M_\alpha = 0.85$ with $t(, 6, 10, 11, 12)$) and RMSE (4.17 for LMM, and 4.15 for $M_\alpha = 0.85$ with $t(, 6, 10, 11, 12)$). Similar to M_α methods, LMM overestimated 90% CR with an estimated CR equal to 95%, whereas M_α provided a CR closer to 94%. In comparison to $S_{\kappa=30}$, LMM prediction was slightly worse in terms of MAE (2.97 for $S_\kappa = 30$ for $t(4, 8, 12)$), but had a smaller RMSE in general.

4. Simulation study

We compare the performance of various scenarios of PLMs to LMM through a simulation study of nonlinear trajectories. We simulate data from the Preece–Baines model I (PB1), a popular nonlinear model for representing children's growth [34–36]. PB1 is widely used to fit growth curves of children and adolescents, capturing the complex curve of growth for the steady growth during childhood, and the adolescent growth spurt [36]. PB1 is an example of a parametric model in which the parameters can be interpreted as biological phenomena and thus allows the study of differences between individuals and the populations to which they belong. In practice, some simulated datasets produced unrealistic growth patterns. In this simulation, we filter out unrealistic growth patterns by comparing the simulated data with WHO Growth Standards of children aged 5 to 19 years based on heights at 0, 3, 5, 10, and 20 years [37]; we excluded data outside the 5th to 95th percentiles, but added random variability to resemble real trajectories.

We generated 1000 simulated datasets, with irregularly spaced observations and varying follow-up periods with a sample size of 1000,

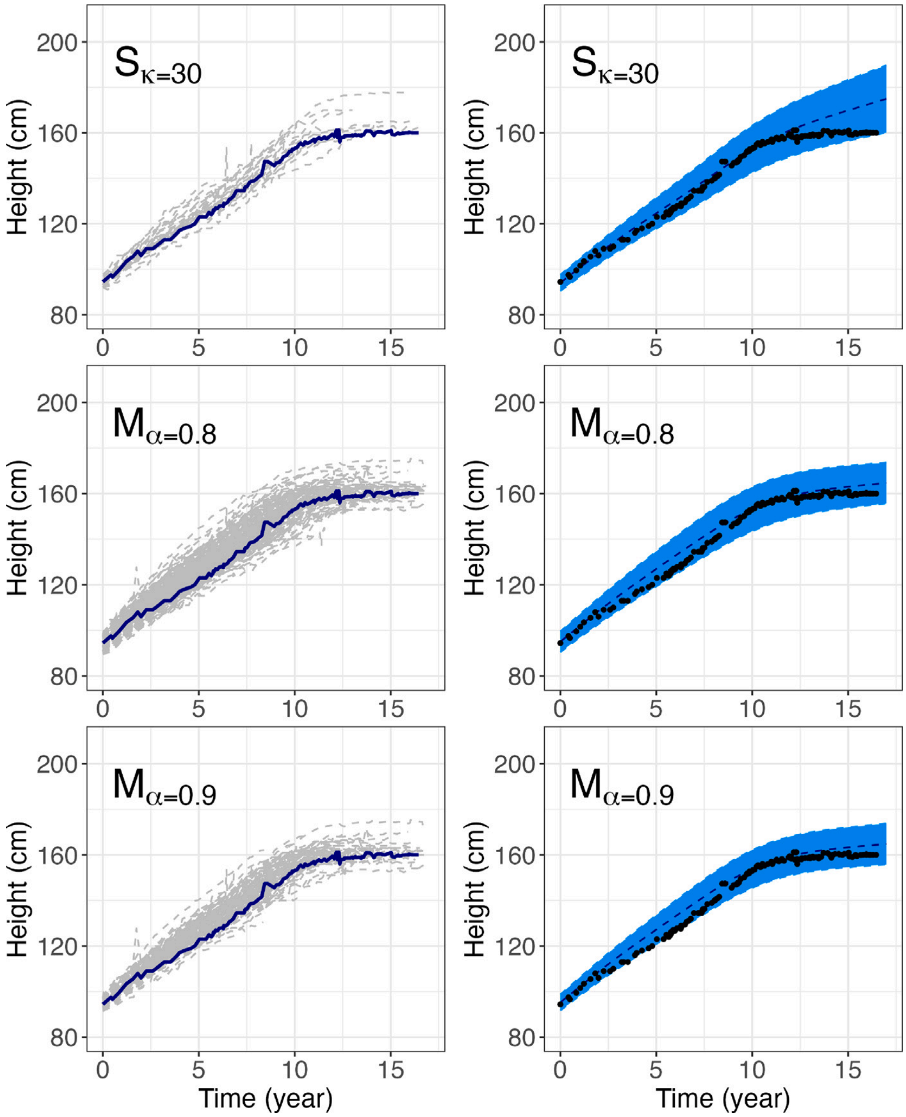

Fig. 2. Trajectory of growth for a sample individual. The left panel displays the target trajectory in dark blue color and the matches in light-grey color. The right panel displays the prediction of target individual, with the shades of blue regions for 90% predictive intervals; the black dots represent the actual observation points. Anchor times are at $t(3, 6, 9, 12)$. Five different methods are presented (from top to bottom): $S_{\kappa=30}$ Single-time (t = 10) with 30 matches; $M_{\alpha=0.8}$ and $M_{\alpha=0.9}$ matching based α criterion with $\alpha = 0.8$ and $\alpha = 0.9$, respectively. The quantity α denotes the value chosen to define the quantile used by the chi-squared distribution criterion.

with training and testing datasets having a ratio of 500:500. Due to the filtering of unrealistic trajectories, we originally generated 10,000 individual trajectories for each gender resulting in a final sample size of 1000. Details of how individual visit times were generated are included in Supplementary Section 3. Children's growth data are simulated based on the following equation.

$$h(t;\theta) = h_{max} - \frac{2(h_{max} - h_\theta)}{e^{s_0(t-\theta)} + e^{s_1(t-\theta)}}, \tag{4}$$

where h_{max} corresponds to the estimated mature or final height in centimeters for each individual; we simulated from normal distributions with different means and standard deviations for males ($N(175, 10^2)$) and females ($N(160, 11^2)$). The age at peak velocity θ was assumed uniformly distributed for males ($U(11, 17)$) and females ($U(10, 15)$). The difference $h_{max} - h_\theta$ was assumed normal for males with $N(20, 4^2)$ and females $N(20, 6.5^2)$. Lastly, the prepubertal and pubertal growth rates controlling growth velocity were represented by s_0 and s_1, respectively.

After simulating longitudinal height from Eq. (4), we added a residual noise component $N(0, 2^2)$. Simulated sample trajectories were displayed in Supplementary Figure S2 which resemble trajectories from the EPIC study data displayed in Supplementary Figure S1. See Supplementary Section S13.3 for more details on the generation of simulated data.

To apply PLM methods, we made similar choices as those for the EPIC analysis. In step 1, we fitted the brokenstick model with internal knots at 5, 10 and 15 years; in step 2, we considered the same scenarios as in the EPIC analysis, which included single-time matching with t = 10, multiple-time matching with Mahalanobis with α criterion; and in step 3, the GAMLSS model was also the same as in the EPIC analysis, namely a normal model with cubic spline for the mean and linear trend of time for the variance. In step 1, for building the linear model, we used gender as a covariate and the following set of anchor times: the 3-anchor-time set $t(4, 8, 12)$, the 4-anchor-time set $t(3, 6, 9, 12)$, and the 6-anchor-time set $t(6, 9, 11, 12, 13, 15)$. The same metrics as in the data analysis were used including MAE, RMSE, as well as 90% interval CR. The overall performance of the PLM method was assessed by taking the average of the metrics across all individuals in each simulated dataset.

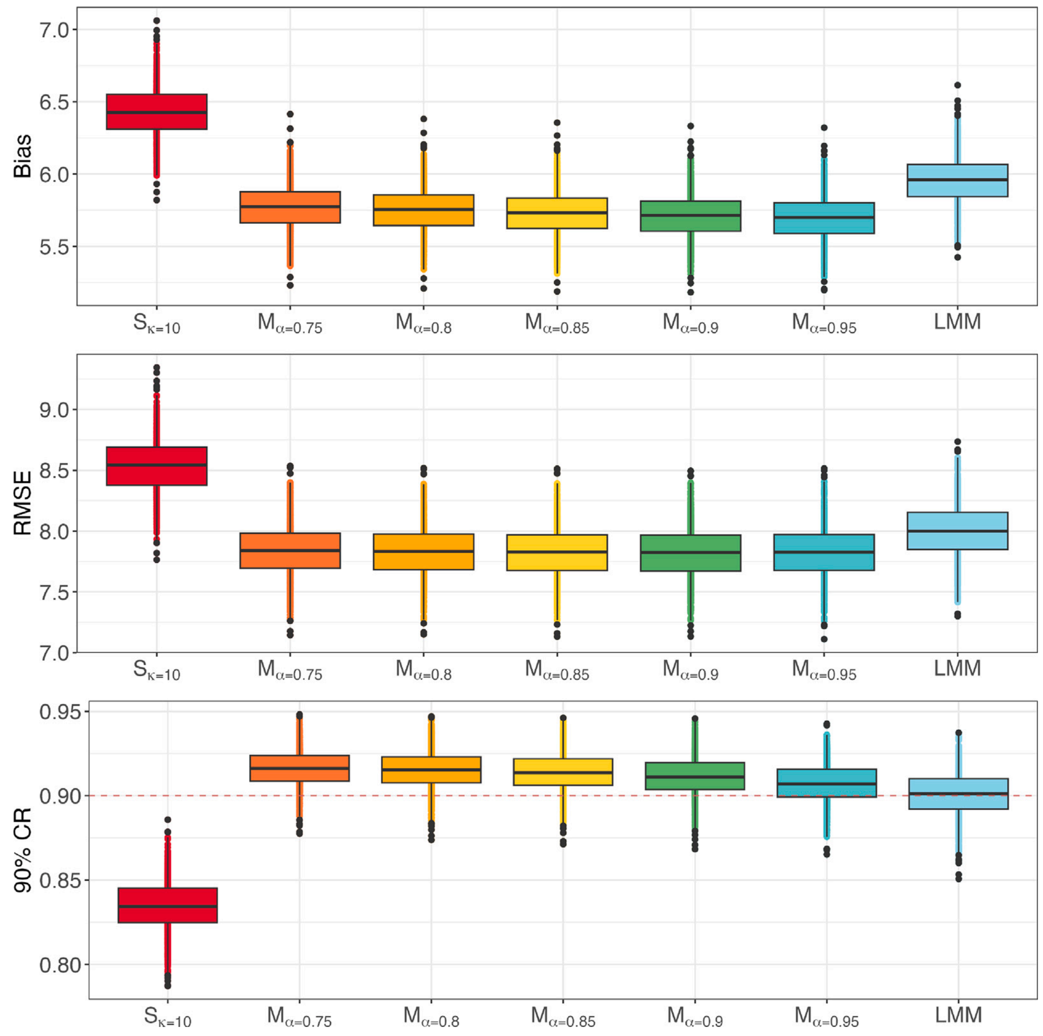

Fig. 3. Simulation study results for different PLM methods. Results of 1000 simulation datasets of sample size of 1000 individual trajectories, with 500 as training set and 500 as testing set: the datasets were simulated based on Preece–Baines model I (PB1) for nonlinear growth curve; the anchor time set were defined at $t(6, 9, 12)$. From the top, the figure presents the mean absolute error (MAE), the root mean squared error (RMSE), and coverage rate of 90% predictive interval (90% CR). For left to right, boxplot include: Single-time point matching $S_{\kappa=10}$, Mahalanobis distance with critical value ($M_{\alpha=0.75}$, $M_{\alpha=0.8}$, $M_{\alpha=0.85}$, $M_{\alpha=0.9}$, and $M_{\alpha=0.95}$), and LMM.

Table 3
Simulation results reporting the mean absolute error (MAE), the root mean squared error (RMSE), and coverage rate of 90% predictive interval (90% CR). The simulated dataset is based on Preece–Baines model I (PB1) and 1000 datasets have been simulated and the anchor times are reported under Time.

Statistics	Time	$S_{\kappa=10}$	$M_{\alpha=0.75}$	$M_{\alpha=0.80}$	$M_{\alpha=0.85}$	$M_{\alpha=0.90}$	$M_{\alpha=0.95}$	LMM
MAE	$t(4, 8, \mathbf{12})$	6.43	5.77	5.75	5.73	5.71	5.70	*5.96*
	$t(3, 6, 9, \mathbf{12})$	6.43	5.91	5.87	5.83	5.79	5.75	
	$t(6, 9, 11, \mathbf{12}, 13, 15)$	6.43	5.92	5.88	5.84	5.80	5.75	
RMSE	$t(4, 8, \mathbf{12})$	8.54	7.84	7.83	7.83	7.82	7.83	*8.00*
	$t(3, 6, 9, \mathbf{12})$	8.54	7.91	7.89	7.87	7.85	7.83	
	$t(6, 9, 11, \mathbf{12}, 13, 15)$	8.54	7.92	7.90	7.87	7.85	7.83	
90%CR	$t(4, 8, \mathbf{12})$	0.83	0.92	0.92	0.91	0.91	0.91	*0.90*
	$t(3, 6, 9, \mathbf{12})$	0.83	0.92	0.92	0.92	0.92	0.91	
	$t(6, 9, 11, \mathbf{12}, 13, 15)$	0.83	0.92	0.92	0.92	0.92	0.91	

Fig. 3 displays simulation results for the three anchor time set $t(6, 9, \mathbf{12})$ scenario with $S_{\kappa=10}$, $M_{\alpha=0.85}$, and LMM; Table 3 displays the corresponding numerical results. The results are consistent with our analysis in Section 3. Specifically, PLMs with a Mahalanobis distance α criterion are usually superior to $S_{\kappa=10}$, with lower MAE and RMSE, and closer 90% CR. The performance of M_α improves with increasing α in terms of smaller MAE and RMSE, as well as closer coverage rate. The 90% CR are consistent across different α values and they are estimated close to 90%. Supplementary Figure S4 shows the results for other anchor-time sets; the numeric results for the simulation study are presented in Table 3. As we had noted in the EPIC analysis in Section 3, when the number of time points increases, the selection criteria become more liberal which can affect the quality of matches, ultimately impacting the performance of PLM. The overall performance of $M_{\alpha>0.80}$ PLM is better than the predictions of LMM with smaller MAE

($M_{\alpha=0.95}$ 5.70 to LMM 5.96) and RMSE ($M_{\alpha=0.95}$ 7.83 to LMM 8.00), as well as comparable coverage rate close to 90% predictive intervals.

5. Discussion

This paper introduces an enhanced People-Like-Me (PLM) method for personalized prediction and compares its performance with prediction from linear mixed models (LMM) across various scenarios. PLM selects curves similar to the target individual and uses them to predict future trajectories. Our proposal of an enhanced approach is based on using the Mahalanobis distance using multiple time points to find matches. We developed and applied PLM methods to children's growth data from the EPIC study and simulated datasets. The PLM methods include single-time-point matching, multiple-time matching using Mahalanobis distance with an α criterion. We also compared PLM predictions to LMM.

Our results show that the PLM using Mahalanobis distance with multiple-time matching and the α criterion (M_α) outperforms single-time matching as well as the LMM prediction methods. For both the real data and simulations of nonlinear trajectories, Mahalanobis distance with the α criterion performed best. In the EPIC data analysis, PLM provided personalized predictions that captured nonlinear changes in individual trajectories (see Fig. 2). This method offers clinical providers with tools to predict individual growth trajectories based on baseline information, allowing for personalized interventions. Moreover, current PLM is a data-driven algorithm, which highly depends on the dataset. In practice, the performance might be affected by the training set sample size and the data quality.

The availability of data to characterize longitudinal trajectories plays an important role in the applicability and performance of the proposed PLM method. A key requirement is that trajectories be aligned at a common time zero—for example, in the EPIC study, trajectories were aligned at a starting age between 3 and 4 years. However, this alignment introduces potential selection bias. Children who tested positive for Pseudomonas aeruginosa (PA+) earlier in life were excluded from the cohort, as eligibility required being PA-negative for at least three months. Similarly, children recruited later in life, such as at age 6, were also omitted because their trajectory start times fell outside the designated window. As a result, some of children were excluded due to their starting age, which can bias the analysis and limit its generalizability to the broader population of children with cystic fibrosis (CF). More broadly, the definition of time zero — such as the time of diagnosis — can significantly impact the availability of longitudinal data. If diagnosis occurs very early or late in life, the amount of usable data for curve matching can be limited, especially for very young individuals. Therefore, the choice of trajectory start time is a critical factor that can impact the effectiveness of the PLM method.

The PLM algorithm requires several user-defined steps, which can significantly affect performance. These include selecting anchor time points, defining matching criteria, and choosing the final prediction model. As shown in the EPIC analysis, selecting the right matches is critical for accurate predictions. We recommend careful validation to refine these steps. Specifically, more anchor points should be used where trajectories change sharply. For instance, in the EPIC data, trajectories were similar before adolescence but diverged after. More anchor points after puberty, where trajectories differ the most by sex, would be sensible. If single-time matching is used, the anchor should be placed where differences are most pronounced. Using more anchor points can capture subtle changes and trends in trajectories, however it also increases the degrees of freedom in the chi-squared distribution. For a fixed α right-tail rejection value, this leads to a larger distance between the target and accepted matches. Consequently, the broader acceptance region can increase the chance of including less similar donors, particularly when outliers or atypical patterns are present. This can reduce matching precision and introduce greater variability in predictions. Limiting anchor points to a small set can help balance the ability to detect meaningful changes as well as to select high-quality matches.

While PLM is flexible, it has limitations. Strict criteria can result in no matches for certain individuals. This could also occur for outliers (unusual trajectories), which raises the question of whether predictions are sensible for such cases. Additionally, multiple-time-point prediction can be computationally intensive, requiring significant resources for distance calculation, matching, and prediction. This complexity increases during validation to fine-tune parameters. We aim to extend this work to improve the PLM methods to address some of these limitations. First, fit a single model instead of a two-stage model. Second, develop a method that uses all calculated distances but weights them to give more influence to trajectories more similar to the target, possibly through weighted models or penalization in the prediction model.

CRediT authorship contribution statement

Xin Jin: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Elizabeth Juarez-Colunga:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Stef van Buuren:** Writing – review & editing, Writing – original draft, Methodology, Investigation. **Kathryn Colborn:** Writing – review & editing, Writing – original draft. **Margaret Rosenfeld:** Writing – review & editing, Writing – original draft. **Jeremy Graber:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jennifer Stevens-Lapsley:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Investigation, Funding acquisition, Conceptualization.

Ethics approval and consent to participate

Study data collection was conducted in accordance with the Declaration of Helsinki and was approved by the ethics committee at the University of Colorado (COMIRB 13-2383). All data collected was retrospective in nature and de-identified, so individual consent forms were not needed and administrative data privileges were not needed.

Funding information

This work was supported by the Agency for Healthcare Research and Quality, United States (R01 HS025692). Dr. Juarez-Colunga's work was partially supported by the National Institute Of Neurological Disorders And Stroke of the National Institutes of Health under Award Number R01NS133040. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Declaration of competing interest

The authors have declared no conflict of interest.

Acknowledgments

This research was done using services provided by the OSG Consortium [38,39], which is supported by the National Science Foundation, United States awards #2030508 and #1836650 [40,41].

Appendix A. Supplementary data

Web Supplementary Materials, referenced in Section 2, Section 3 and Section 4, are available with this paper online version.

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compbiomed.2025.111281.

References

[1] R. Snyderman, Personalized health care: from theory to practice, Biotechnol. J. (ISSN: 1860-7314) 7 (8) (2012) 973–979, http://dx.doi.org/10.1002/biot.201100297.

[2] K. Ng, J. Sun, J. Hu, F. Wang, Personalized predictive modeling and risk factor identification using patient similarity, AMIA Jt. Summits. Transl. Sci. Proc. (ISSN: 2153-4063) 2015 (2015) 132–136.

[3] B.A. Goldstein, A.M. Navar, M.J. Pencina, J.P.A. Ioannidis, Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review, J. Am. Med. Inform. Assoc. : JAMIA (ISSN: 1067-5027) 24 (1) (2017) 198–208, http://dx.doi.org/10.1093/jamia/ocw042, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5201180/.

[4] K. Liu, X. Zhang, W. Chen, A.S.L. Yu, J.A. Kellum, M.E. Matheny, S.Q. Simpson, Y. Hu, M. Liu, Development and validation of a personalized model with transfer learning for acute kidney injury risk estimation using electronic health records, JAMA Netw. Open (ISSN: 2574-3805) 5 (7) (2022) e2219776, http://dx.doi.org/10.1001/jamanetworkopen.2022.19776.

[5] G. Salazar de Pablo, E. Studerus, J. Vaquerizo-Serrano, J. Irving, A. Catalan, D. Oliver, H. Baldwin, A. Danese, S. Fazel, E.W. Steyerberg, D. Stahl, P. Fusar-Poli, Implementing precision psychiatry: A systematic review of individualized prediction models for clinical practice, Schizophr. Bull. (ISSN: 0586-7614) 47 (2) (2021) 284–297, http://dx.doi.org/10.1093/schbul/sbaa120.

[6] A.J. Kittelson, T.J. Hoogeboom, M. Schenkman, J.E. Stevens-Lapsley, N.L. van Meeteren, Person-centered care and physical therapy: a “people-like-me” approach, Phys. Ther. 100 (1) (2020) 99–106.

[7] J. Graber, A. Kittelson, E. Juarez-Colunga, X. Jin, M. Bade, J. Stevens-Lapsley, Comparing “people-like-me” and linear mixed model predictions of functional recovery following knee arthroplasty, J. Am. Med. Inf. Assoc. (ISSN: 1527-974X) 29 (11) (2022) 1899–1907, http://dx.doi.org/10.1093/jamia/ocac123.

[8] S. Van Buuren, Curve matching: A data-driven technique to improve individual prediction of childhood growth, Ann. Nutr. Metab. (ISSN: 02506807) 65 (2) (2014) 227–233, http://dx.doi.org/10.1159/000365398.

[9] D.B. Rubin, The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The SIR algorithm, J. Amer. Statist. Assoc. 82 (398) (1987) 543–546.

[10] H. Zhang, Y. Zhou, D. Feng, Mahalanobis distance similarity measure based distinguisher for template attack: MDSM-based distinguisher for template attack, Secur. Comm. Netw. (ISSN: 19390114) 8 (5) (2015) 769–777, http://dx.doi.org/10.1002/sec.1033, URL https://onlinelibrary.wiley.com/doi/10.1002/sec.1033.

[11] T. Ong, B.W. Ramsey, Cystic fibrosis: a review, Jama 329 (21) (2023) 1859–1871.

[12] K.A. Mason, A.D. Rogol, Trends in growth and maturation in children with cystic fibrosis throughout nine decades, Front. Endocrinol. 13 (2022) 935354.

[13] E. Owen, J.E. Williams, G. Davies, C. Wallis, R.L. Grant, M.S. Fewtrell, Growth, body composition, and lung function in prepubertal children with cystic fibrosis diagnosed by newborn screening, Nutr. Clin. Pr. 36 (6) (2021) 1240–1246.

[14] T. Hastie, R. Tibshirani, Varying-coefficient models, J. R. Stat. Soc. Ser. B Stat. Methodol. 55 (4) (1993) 757–779.

[15] E.W. Steyerberg, D. Nieboer, T.P. Debray, H.C. van Houwelingen, Assessment of heterogeneity in an individual participant data meta-analysis of prediction models: An overview and illustration, Stat. Med. 38 (22) (2019) 4290–4309.

[16] T.M. Fragoso, W. Bertoli, F. Louzada, Bayesian model averaging: A systematic review and conceptual classification, Int. Stat. Rev. 86 (1) (2018) 1–28.

[17] D. Rizopoulos, Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data, Biometrics 67 (3) (2011) 819–829.

[18] B.S. Finkelman, B. French, S.E. Kimmel, The prediction accuracy of dynamic mixed-effects models in clustered data, BioData Min. (ISSN: 1756-0381) 9 (1) (2016) 5, http://dx.doi.org/10.1186/s13040-016-0084-6.

[19] A. Johnson, P.K.-S. Ng, M. Kahle, J. Castillo, B. Amador, V. Holla, T. Vu, L. Huang, F. Su, S. Kim, et al., Patient-specific, tiered, variant-level actionability correlates with functional effect in growth survival assay, Cancer Res. 81 (13_Supplement) (2021) 392.

[20] T. Zhu, K. Li, P. Herrero, J. Chen, P. Georgiou, A deep learning algorithm for personalized blood glucose prediction, in: KDH@ IJCAI, 2018, pp. 64–78.

[21] S. Bose, C.C. Kenyon, A.J. Masino, Personalized prediction of early childhood asthma persistence: a machine learning approach, PloS One 16 (3) (2021) e0247784.

[22] J.R.A. Solares, F.E.D. Raimondi, Y. Zhu, F. Rahimian, D. Canoy, J. Tran, A.C.P. Gomes, A.H. Payberah, M. Zottoli, M. Nazarzadeh, et al., Deep learning for electronic health records: A comparative review of multiple deep neural architectures, J. Biomed. Inform. 101 (2020) 103337.

[23] S. Van Buuren, Broken stick model for irregular longitudinal data, J. Stat. Softw. (ISSN: 1548-7660) 106 (2023) 1–51, http://dx.doi.org/10.18637/jss.v106.i07.

[24] I. Eekhout, S. van Buuren, B. Visser, M.C.A.M. Bink, A. Huisman, Longitudinal individual predictions from irregular repeated measurements data, Sci. Rep. (ISSN: 2045-2322) 13 (1) (2023) 952, http://dx.doi.org/10.1038/s41598-022-26933-1, URL https://www.nature.com/articles/s41598-022-26933-1, Number: 1 Publisher: Nature Publishing Group.

[25] D. Stasinopoulos, R.A. Rigby, F. De Bastiani, GAMLSS: a distributional regression approach, Stat. Model. (ISSN: 1477-0342) 18 (3) (2018) 248–273, URL https://journals.sagepub.com/doi/abs/10.1177/1471082X18759144?journalCode=smja, Number: 3-4 Publisher: SAGE Publications.

[26] D.M. Stasinopoulos, R.A. Rigby, Generalized additive models for location scale and shape (GAMLSS) in R, J. Stat. Softw. (ISSN: 1548-7660) 23 (2008) 1–46, http://dx.doi.org/10.18637/jss.v023.i07.

[27] D. Rizopoulos, The R package JMbayes for fitting joint models for longitudinal and time-to-event data using MCMC, J. Stat. Softw. 72 (7) (2016) 1–45, http://dx.doi.org/10.18637/jss.v072.i07.

[28] J. Lin, K. Li, S. Luo, Functional survival forests for multivariate longitudinal outcomes: Dynamic prediction of Alzheimer's disease progression, Stat. Methods Med. Res. 30 (1) (2021) 99–111.

[29] J.C. Pinheiro, D.M. Bates, Mixed-Effects Models in S and S-PLUS, Springer, New York, 2000, http://dx.doi.org/10.1007/b98882.

[30] R.A. Rigby, D.M. Stasinopoulos, Generalized additive models for location, scale and shape,(with discussion), Appl. Stat. 54 (2005) 507–554.

[31] M.M. Treggiari, M. Rosenfeld, N. Mayer-Hamblett, G. Retsch-Bogart, R.L. Gibson, J. Williams, J. Emerson, R.A. Kronmal, B.W. Ramsey, EPIC Study Group, Early anti-pseudomonal acquisition in young patients with cystic fibrosis: rationale and design of the EPIC clinical trial and observational study', Contemp. Clin. Trials. (ISSN: 1559-2030) 30 (3) (2009) 256–268, http://dx.doi.org/10.1016/j.cct.2009.01.003.

[32] M. Rosenfeld, J. Emerson, S. McNamara, K. Joubran, G. Retsch-Bogart, G.R. Graff, H.H. Gutierrez, J.F. Kanga, T. Lahiri, B. Noyes, et al., Baseline characteristics and factors associated with nutritional and pulmonary status at enrollment in the cystic fibrosis EPIC observational cohort, Pediatr. Pulmonol. (ISSN: 1099-0496) 45 (9) (2010) 934–944, http://dx.doi.org/10.1002/ppul.21279.

[33] X. Jin, E. Juarez-Colunga, S. van Buuren, M. Rosenfeld, J. Graber, M. Bade, J. Stevens-Lapsley, Comparative analysis of people-like-me methods and linear mixed models: Personalized prediction of longitudinal growth in children, TBD TBD (2025) TBD, URL https://github.com/EJC-Lab/jin_scientific_report_2024.git.

[34] B.S. Zemel, F.E. Johnston, Application of the Preece-Baines growth model to cross-sectional data: Problems of validity and interpretation, Am. J. Hum. Biol. 6 (5) (1994) 563–570.

[35] T. Brown, G. Townsend, Adolescent growth in height of Australian Aboriginals analysed by the Preece-Baines function: a longitudinal study, Ann. Hum. Biol. 9 (6) (1982) 495–505.

[36] J.M. Tanner, T. Hayashi, M. Preece, N. Cameron, Increase in length of leg relative to trunk in Japanese children and adults from 1957 to 1977: comparison with British and with Japanese Americans, Ann. Hum. Biol. 9 (5) (1982) 411–423.

[37] World Health Organization, et al., WHO Child Growth Standards: Length/height-For-Age, Weight-For-Age, Weight-For-Length, Weight-For-Height and Body Mass Index-For-Age: Methods and Development, World Health Organization, 2006.

[38] OSG, OSPool, 2006, http://dx.doi.org/10.21231/906P-4D78, URL https://osg-htc.org/services/open_science_pool.html.

[39] OSG, Open science data federation, 2015, http://dx.doi.org/10.21231/0KVZ-VE57, URL https://osdf.osg-htc.org/.

[40] R. Pordes, D. Petravick, B. Kramer, D. Olson, M. Livny, A. Roy, P. Avery, K. Blackburn, T. Wenaus, F. Wurthwein, et al., The open science grid, J. Phys.: Conf. Ser. 78 (2007).

[41] I. Sfiligoi, D.C. Bradley, B. Holzman, P. Mhashilkar, S. Padhi, F. Wurthwein, The pilot way to grid resources using glideinWMS, in: 2009 WRI World Congress on Computer Science and Information Engineering, Vol. 2, IEEE, 2009, pp. 428–432.