BioData Mining

METHODOLOGY Open Access

Leveraging mixed-effects regression trees for the analysis of high-dimensional longitudinal data to identify the low and high-risk subgroups: simulation study with application to genetic study

Mina Jahangiri[1], Anoshirvan Kazemnejad[1*], Keith S. Goldfeld[2], Maryam S. Daneshpour[3], Mehdi Momen[4], Shayan Mostafaei[5], Davood Khalili[6] and Mahdi Akbarzadeh[3*]

*Correspondence:
Anoshirvan Kazemnejad
kazem_an@modares.ac.ir
Mahdi Akbarzadeh
akbarzadeh.ms@gmail.com
[1]Department of Biostatistics, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran
[2]Division of Biostatistics, Department of Population Health, NYU Grossman School of Medicine, New York, NY, USA
[3]Cellular and Molecular Endocrine Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[4]Department of Surgical Sciences, School of Veterinary Medicine, University of Wisconsin-Madison, Madison, WI, USA
[5]Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm, Sweden
[6]Prevention of Metabolic Disorders Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran

Abstract

Background The linear mixed-effects model (LME) is a conventional parametric method mainly used for analyzing longitudinal and clustered data in genetic studies. Previous studies have shown that this model can be sensitive to parametric assumptions and provides less predictive performance than non-parametric methods such as random effects-expectation maximization (RE-EM) and unbiased RE-EM regression tree algorithms. These longitudinal regression trees utilize classification and regression trees (CART) and conditional inference trees (Ctree) to estimate the fixed-effects components of the mixed-effects model. While CART is a well-known tree algorithm, it suffers from greediness. To mitigate this issue, we used the Evtree algorithm to estimate the fixed-effects part of the LME for handling longitudinal and clustered data in genome association studies.

Methods In this study, we propose a new non-parametric longitudinal-based algorithm called "Ev-RE-EM" for modeling a continuous response variable using the Evtree algorithm to estimate the fixed-effects part of the LME. We compared its predictive performance with other tree algorithms, such as RE-EM and unbiased RE-EM, with and without considering the structure for autocorrelation between errors within subjects to analyze the longitudinal data in the genetic study. The autocorrelation structures include a first-order autoregressive process, a compound symmetric structure with a constant correlation, and a general correlation matrix. The real data was obtained from the longitudinal Tehran cardiometabolic genetic study (TCGS). The data modeling used body mass index (BMI) as the phenotype and included predictor variables such as age, sex, and 25,640 single nucleotide polymorphisms (SNPs).

Results The results demonstrated that the predictive performance of Ev-RE-EM and unbiased RE-EM was nearly similar. Additionally, the Ev-RE-EM algorithm generated smaller trees than the unbiased RE-EM algorithm, enhancing tree interpretability.

Conclusion The results showed that the unbiased RE-EM and Ev-RE-EM algorithms outperformed the RE-EM algorithm. Since algorithm performance varies across datasets, researchers should test different algorithms on the dataset of interest and select the best-performing one. Accurately predicting and diagnosing an individual's genetic profile is crucial in medical studies. The model with the highest accuracy should be used to enhance understanding of the genetics of complex traits, improve disease prevention and diagnosis, and aid in treating complex human diseases.

Keywords Mixed-effects model, Regression tree, Longitudinal data, High-dimensional data, Genetic study

Background

Observations in longitudinal genetic studies are typically collected from individuals over time to represent clustered data. More generally, clustered data reflect a hierarchical structure in which multiple observations are nested within objects (e.g., individuals within families). Each subject in a longitudinal setting represents a single cluster. The traditional parametric model for the analysis of these data sets in genetic studies is a linear mixed-effects model (LME) that simultaneously handles random effects (differences between objects) and fixed effects (population-level associations) [1].

The LME is widely used in genome-wide association studies (GWAS) to account for individual population structure and relatedness. This model helps identify genetic variants associated with complex traits by controlling for confounding factors that can lead to spurious associations. This model incorporates random effects to model the genetic relatedness among individuals, which is crucial in GWAS, where population structure can confound results. By accounting for these relationships, LMMs reduce false positives and improve the accuracy of association tests [2].

Assume a panel of subjects i = 1,..., n at time points t = 1,..., T_i, a vector of K-predictor variables $x_{it} = (x_{it1}, \ldots, x_{itK})\prime$, and a response variable y_{it} for each observation (i, t). Then, the LME is defined as $y_{it} = X_{it}\beta + Z_{it}b_i + \epsilon_{it}$ and $y_{it} = [y_{i1}, ., y_{iT_i}]^T$, where X_{it} is the design matrix of fixed effects, β is the vector of fixed-effects regression coefficients, and Z_{it} is the design matrix of random effects for the t_i observations at i^{th} cluster. Furthermore, b_i represents the random-effects regression coefficients (subject-specific intercept) and follows the normal distribution with mean 0 and variance-covariance matrix D. The $\epsilon_{it} = [\epsilon_{i1}, ., \epsilon_{iT_i}]^T$ are the error terms and independent across subjects. These errors follow a normal distribution with mean 0 and variance-covariance matrix $R_i = \sigma^2 I_{T_i}$ for all i and are uncorrelated with the random effects [1]. This model, incorporating random intercepts to consider the within-subject correlation of response variables, is defined as $y_{it} = X_{it}\beta + b_i + \epsilon_{it}$.

In the standard mixed-effects model, a known linear parametric function $f(x_{it1}, \ldots, x_{itK}) = X\beta$ is assumed, along with an assumption of homoscedasticity. These assumptions are restrictive because using a linear form may not be the best representation for f in all real data applications, and there is often heteroscedasticity ($R_i \neq \sigma^2 I_{T_i}$). In addition, when K is large, using all predictor variables for data modelling may lead to overfitting and poor prediction. Using a non-parametric approach to estimate f can resolve these issues. Several recent studies suggest machine-learning methods like tree-based algorithms can be used to estimate f [3–6]. Tree-based algorithms are supervised non-parametric algorithms and are one of the most popular machine-learning

tools for modelling and prediction [7]. These methods offer several advantages over parametric models, such as being easy to interpret due to the graphical display of results. They can handle high-dimensional and large datasets without requiring assumptions about the functional form of the data. Additionally, they can deal with nonlinear relationships and high-order interactions, extract homogeneous subgroups of observations, and are robust to missing data, outliers, and multicollinearity. Tree algorithms classify observations into finite homogeneous subgroups based on predictor variables by a recursive partitioning process, and then fit a constant or a model for data prediction within these subgroups [8–14]. Often, the target populations for assessing longitudinal changes in the response variable are heterogeneous. Tree-based methods offer a suitable approach to evaluate these changes by extracting homogeneous subgroups.

The CART algorithm is one of the best-known algorithms for data mining [15]. This model suffers from problems such as greediness, instability, and bias in split rule selection [16, 17]. This model generates a tree using a greedy search algorithm, which has disadvantages such as limiting the exploration of tree space, dependence on future splits to previous splits, generating optimistic error rates, and the inability of the search to find a global optimum [18]. CART has an instability problem because resampling or generating bootstrap samples from the dataset may create a tree with different splits [19]. The splitting method in the CART model is biased toward predictor variables with many distinct values [20, 21].

Given these challenges, an alternative approach to parametric models for longitudinal analysis may be warranted. Therefore, this study aims to extend the mixed-effects model to machine-learning methods like tree algorithms. By developing a novel algorithm tailored to the intricacies of longitudinal data in the context of genome association studies, we aim to advance methodologies in the field, offering a solution that aligns with the complexities and nuances inherent in real-world applications. This introduction sets the stage for exploring the proposed Ev-RE-EM new algorithm, which combines the strengths of tree-based methods with the demands of longitudinal data analysis.

Various tree-based models have been proposed to address the challenges of the CART algorithm, and the remedial approaches include ensemble models such as random forests (RF), bagging, boosting, and other ensemble methods to mitigate instability issues [22, 23]. To tackle biases in split rule selection, tree algorithms like CRUISE [24, 25], QUEST [26], GUIDE [27], conditional inference tree (*ctree*) [28], and LOTUS [29] have been introduced. Additionally, the Evtree algorithm is specifically recommended to alleviate the greediness problem associated with CART [18].

Recently, several studies extended the CART algorithm and its remedial methods (e.g., tree, GUIDE, and random forests) to deal with continuous longitudinal and clustered data. Segal proposed the first extension of regression tree algorithms to longitudinal data (1992) [30]. Then, Abdolell et al. (2002) suggested a longitudinal regression tree algorithm to find homogeneous subgroups based on a continuous predictor variable in the package longRPart [31, 32]. However, these approaches mentioned for longitudinal data have some disadvantages, such as the inability to deal with time-dependent predictor variables, being inappropriate for unbalanced clusters ($T_i \neq T$ for all i), and being unable to predict the response variables for future time points for subjects in the data set.

Several alternative approaches were proposed to improve the previous longitudinal regression tree algorithms, and in the following, we will only review the longitudinal regression tree algorithms available in the software. Hajjem et al. (2011, and 2014) modified the CART algorithm and RF for clustered data with continuous response variables using the Expectation Maximization (EM) method. Then, they proposed the mixed-effects regression tree algorithm (MERT) and the mixed-effects random forest algorithm (MERF), respectively [3, 4].

Sela and Simonoff (2012) proposed a solution similar to the two approaches of Hajjem et al. (2011 and 2014), using CART to estimate (the population-level effects/fixed effects part of the mixed-effects model) to deal with the continuous longitudinal response variable. They alternate between the regression tree estimation for the fixed and the random effects estimation. Each estimate of the fixed effects plugs in the estimated random effects from the prior iteration and essentially assumes that random effects estimates are known. This strategy is reminiscent of the EM method [33], so this method is called the random effect-expectation maximization (RE-EM) tree and is implemented in the REEMtree package [5]. However, RE-EM uses the CART algorithm, which tends to favor predictor variables with many unique values for splitting the tree nodes during the tree-growing step. Fu and Simonoff (2015) suggested an unbiased RE-EM algorithm using the *ctree* algorithm instead of CART to estimate and demonstrated that it has better prediction accuracy than RE-EM [6].

Loh and Zheng (2013) extended the GUIDE tree algorithm to longitudinal data and, like the unbiased RE-EM algorithm, solved the bias problem of the original RE-EM algorithm. However, it cannot incorporate time-dependent predictor variables [34]. Eo and Cho (2014) suggested a longitudinal tree algorithm using the GUIDE algorithm and mixed-effects models. Their algorithm, named MELT, handles unbalanced clusters and different types of predictor variables. Unlike previous longitudinal tree algorithms, MELT assesses trends of response variables over time within homogeneous subsets of subjects rather than predicting those response variables [35].

Fokkema et al. (2018) suggested the generalized linear mixed-effects model tree (GLMM tree) using model-based recursive partitioning (MOB) [36] for clustered or nested datasets [37, 38]. Their algorithm is available in the glmertree package in R software [39]. The GLMM tree algorithm is appropriate for subgroup analysis or detecting the interaction effects between treatment and subgroups in clinical trials.

Longitudinal data in observational studies and clinical trials may be influenced by several variables measured at baseline. The conventional model for analysis of this data is LME, which includes the baseline covariates and their interactions with the time variable in the model. This analysis has some drawbacks, such as overfitting, inability to detect the nonlinear effects, and bias in estimators. To address this, Kundu and Harezlak (2019) proposed the longitudinal classification and regression tree (LongCART) under the *ctree* algorithm by Hothorn et al. [28] for analysis of longitudinal data using baseline covariates as partitioning variables to find the subgroups with differential longitudinal trajectories [40]. The LongCART algorithm is available in the LongCART package in R [41].

As noted previously, the CART algorithm suffers from a greediness problem (in the greedy search, split rules are selected forward stepwise to partition the data into groups recursively. The splitting rule at each internal node is chosen to maximize its child nodes'

homogeneity without considering nodes further down, thus yielding only locally optimal trees). Currently, no algorithm addresses this issue for longitudinal and clustered data. This study proposes a new non-parametric longitudinal algorithm called Ev-RE-EM for modeling a continuous response variable using the *Evtree* algorithm. It compares the predictive performance of this proposed algorithm with previous longitudinal regression such as RE-EM and unbiased RE-EM on simulation and real datasets (we selected these algorithms because they can handle time-dependent predictor variables and the R code is available). The study innovation is shown graphically in Fig. 1.

Methods

The description of the real data set

The data used to motivate the new method and inform the simulation study were obtained from the Tehran cardiometabolic genetic study (TCGS) to assess the simultaneous effect of predictor variables, including sex, age, and single nucleotide polymorphisms (SNPs) of chromosome 16 on the body mass index (BMI) during six follow-up waves of TCGS [42].

We selected chromosome 16 because many studies indicated that the FTO gene on this chromosome was determined as a locus associated with BMI [43, 44]. Chromosome 16 consisted of 231,501 SNPs in TCGS, 25,680 of which were multiallelic SNPs. We

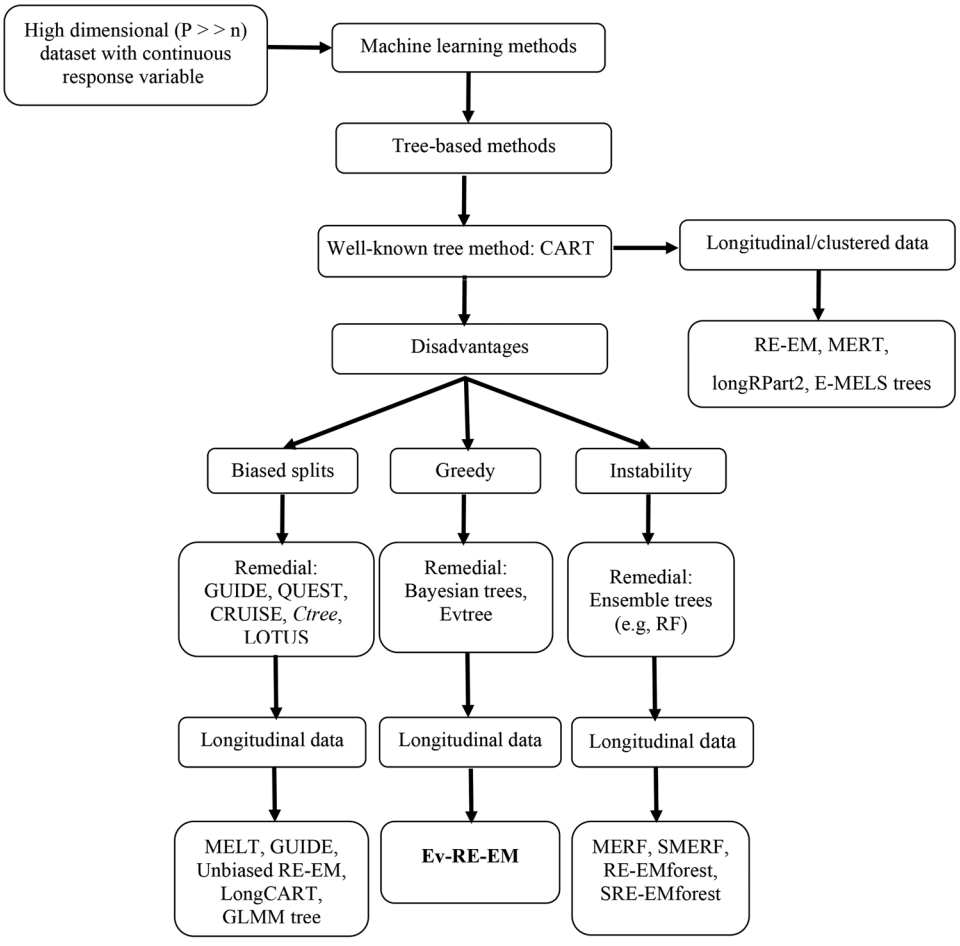

Fig. 1 Study innovation

removed the multiallelic SNPs, leaving 205,821 biallelic SNPs. We used linkage disequilibrium (LD) pruning based on the window size = 50 Kb (kilobase), step size = 5 Kb, and r^2 threshold = 0.1 to select the SNPs in linkage equilibrium (LE); 25,640 SNPs remained in LE and were used in the real dataset.

For data modelling, the co-dominant genetic model was considered for SNP effects. The Beagle version 5.4 was used to impute missing genotypes [45]. The sample size consisted of 3,088 unrelated participants (1,261 (40.8%) of participants were male, and 1,827 (59.2%) were female) based on the genetic relationship matrix (GRM) with a cut point equal to 0.025 [46, 47] and 25,642 predictor variables (25,640 SNPs, gender, and age) to predict BMI as a longitudinal phenotype. This real data includes 3,088 unbalanced clusters; 716, 1,094, and 1,278 were presented in four, five, and six phases of TCGS, respectively. The study plan for selecting participants is shown in Fig. 2.

Estimation method of the Ev-RE-EM algorithm

The operational procedure for running the Ev-RE-EM algorithm can be summarized in three main steps:

1. Set the initial values of $\widehat{b}_i$ to zero.
2. Run the following steps, *a* and *b* repeatedly until the convergence of $\widehat{b}_i$. Convergence is established when the likelihood or restricted likelihood change is less than a predetermined tolerance value (e.g., 0.001).

a) Fit a regression tree to estimate an initial approximation of f using the *Evtree* algorithm, based on the response variable, $y_{it} - \widehat{b}_i$, using predictor variables, $x_{it} = (x_{it1}, \ldots, x_{itK})$, where $i = 1,\ldots,$ I and $t = 1,\ldots, T_i$. This regression tree generates

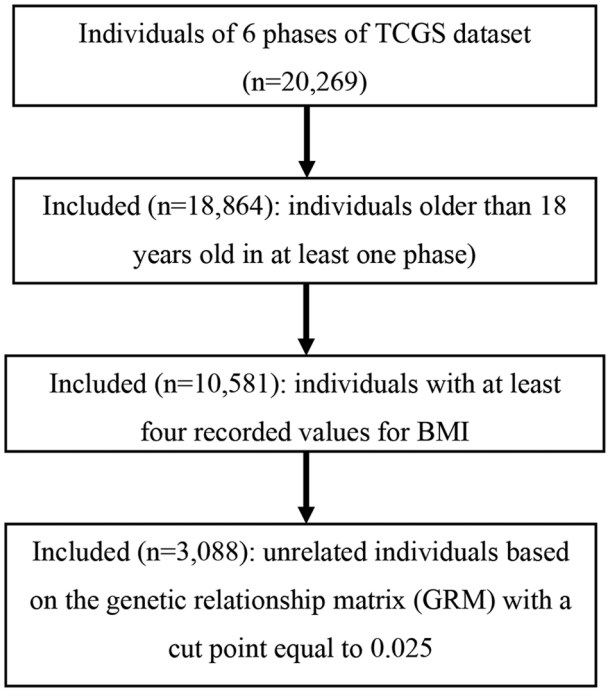

Fig. 2 The study plan for selecting the participants

a set of predictor variables, I($x_{it} \in g_p$) and g_p ranges over all the terminal nodes of the tree.

b) Fit the linear mixed-effects model (LME), $y_{it} = b_i + \sum_p \mathrm{I}\left(x_{it} \in g_p\right) \mu_p + \epsilon_{it}$ and get estimates $\widehat{b_i}$. μ_p is the mean outcome for each of the terminal nodes p.

3. Use the estimated predicted response $\widehat{\mu}_p$ from the fitting of LME in step 2b instead of the predicted response at each tree terminal node.

The *Evtree* algorithm in step 2a is performed using the Evtree R package and as proposed by Grubinger et al. in 2011 [18]. The LME in step 2b is fitted based on the restricted maximum likelihood (REML) instead of maximum likelihood (ML) using the nlme package in R. Studies showed that this estimation method provides unbiased estimates compared to ML.

Simulation study

We designed a simulation experiment to assess the predictive performance of the Ev-RE-EM algorithm and other longitudinal regression trees using 100 simulated datasets containing 3,088 individuals. All parameters in the simulation models were derived from real data to ensure that the results of the simulated data analysis are comparable to those of real data analysis.

In this simulation design, the sex variable was generated from a Binomial distribution (3088, 0.5). The Age variable at the first wave was also generated using a truncated normal distribution with exact minimum = 2, maximum = 81, mean = 39.91, and standard deviation = 15.20 of the real age variable at wave 1. Then, age at other waves was generated as follows:

$\mathrm{Age_{it}} = \mathrm{Age_{i1}} + (\mathrm{t}-1) \times 3$ i = 1,..., 3088, t = 2,..., 6

Finally, the phenotype, BMI at each wave was generated based on the following LME:

$\mathrm{BMI_{it}} = 16.93 + 0.13 \times \mathrm{Age_{ij}} + 2.35 \times \mathrm{Sex_i} + \beta_1 \mathrm{SNP_1} + \beta_2 \mathrm{SNP_2} + \ldots + \beta_{70} \mathrm{SNP_{70}} + \gamma_{0i} + \gamma_{it}$

i = 1,..., 3088, t = 1,..., 6

Where γ_{0i} = N (0, 4.48) is the random intercept, γ_{it} = N (0, 1.74) is the residual error at time *t*, and $\mathrm{SNP_1}, \mathrm{SNP_2}, \ldots, \text{and } \mathrm{SNP_{70}}$ are the causal SNPs. We selected these SNPs from the FTO gene (5,219 SNPs) because different studies indicated this gene as a locus potentially affecting BMI [43, 44]. We used LD pruning on this gene based on the window size = 50 kb, step size = 5 kb, and r^2 threshold = 0.1, and 533 SNPs remained. The pool of causal SNPs was determined from the LD pruned dataset with minor allele frequency (MAF) equal to 0.3, and 85 SNPs satisfied this condition. Fifteen SNPs from a pool of causal SNPs were removed to solve the multicollinearity because these SNPs had variance inflation factor (VIF) of more than 10 in the LME (BMI ~ Age+ Sex + $\mathrm{SNP_1}$ + $\mathrm{SNP_2}$ +... + $\mathrm{SNP_{85}}$+ 1|id). The VIF results before and after removing the SNPs with high multicollinearity are shown in Tables S1 and S2 (supplementary files), respectively. The steps for determining the causal SNPs to simulate the phenotype at each wave are depicted in Fig. 3.

The previous studies generated the regression coefficients of causal SNPs for phenotype simulation either from standard normal distribution or from effect sizes obtained from summary statistics of other studies [46]. In the present study, we generated the regression coefficients of 70 causal SNPs for phenotype simulation at each wave from

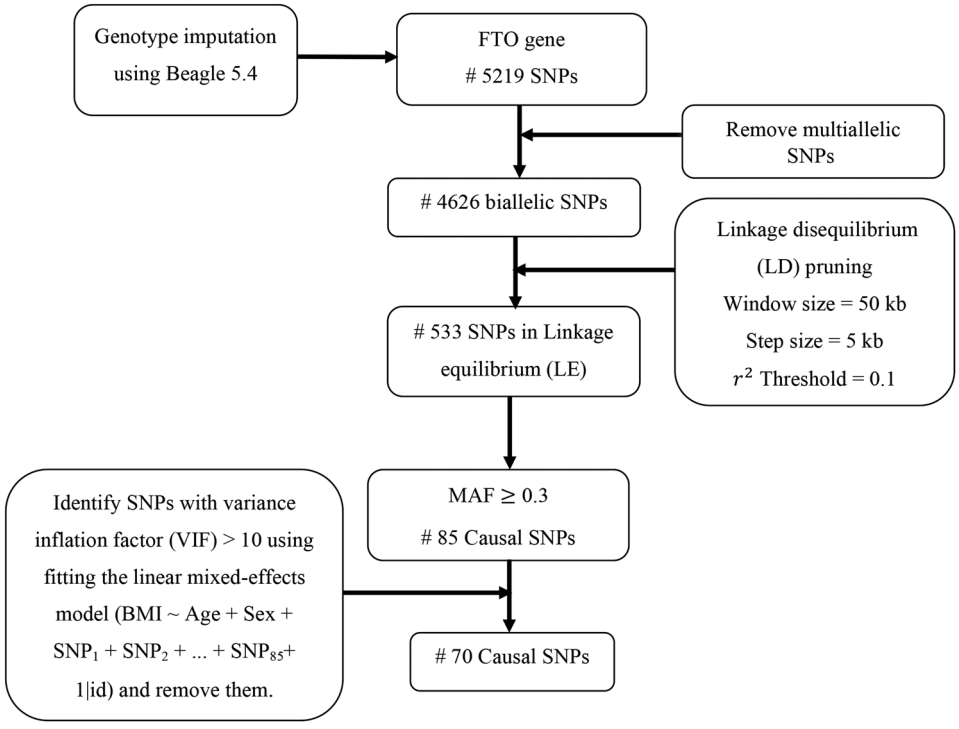

Fig. 3 The study plan for selecting the causal single nucleotide polymorphisms (SNPs) to simulate the body mass index (BMI)

the multivariate normal distribution; the parameters of this generating distribution were selected based on the effect size of these SNPs on the BMI that was observed in our real-world data at each wave, determined through the fitting of the 70 linear regression models ($\mathrm{BMI_i} \sim \mathrm{Age_i} + \mathrm{Sex} + \mathrm{PC_1} + \mathrm{PC_2} + \ldots + \mathrm{PC_{10}} + \mathrm{SNP_{ij}}$, i = 1,2,..., 6, j = 1,2,..., 70) at each wave of TCGS using the first 10 principal components (PC) based on the genetic relationship matrix to adjust the effect of population stratification. Principal component analysis (PCA) is a widely used statistical method in population genetics for inferring the genetic structure of populations. PCA identifies genetic variations among individuals and groups, often associated with geographic or cultural factors. By reducing the dimensionality of genetic data, PCA simplifies complex patterns of genetic variation and facilitates the interpretation of genetic relationships among populations [46, 47]. These effect sizes are shown in Table S3 (supplementary file). In addition, before fitting the linear regression model in each study phase, rank-based inverse normal transformation (RINT) was used for BMI normalization [48]. The density plots for BMI phenotype before and after using RINT in each study phase are shown in Figures S1 to S6 (supplementary files).

The trend plot of effect sizes, provided in Figure S7 (supplementary file), suggests that these SNPs have different effects over time. This variability led us to use different regression coefficients for causal SNPs to simulate BMI at each wave. The correlation plot between the causal SNPs for six waves is shown in Figure S8 (supplementary file); the correlation between the causal SNPs appears to decrease over time. Furthermore, using the multivariate normal distribution to generate the regression coefficients of causal SNPs for BMI simulation at each wave of the study appears reasonable.

Finally, we have a mean vector $\mu_{(70\times 6)\times 1}$ (the long format of effect sizes in Table S3 is considered as the mean vector) and variance-covariance matrix $\Sigma_{(70\times 6)\times(70\times 6)}$ of the multivariate normal distribution to generate the regression coefficients of the SNPs to simulate the phenotype at each wave. We considered a block diagonal matrix for Σ, where each block belongs to a specific SNP and is based on the variance-covariance matrix of effect sizes for the six waves. The block diagonal matrix for Σ is defined as follows:

The matrix R_i (variance-covariance matrix of individual-level noise in LME) can be diagonal when we are assuming independence within each individual. However, a non-diagonal matrix might better represent the noise, as the errors might be correlated over time within an individual. If this is the case, this matrix can have different structures, such as the first-order autoregressive process, compound symmetry structure with a constant correlation, or a general correlation matrix (unstructured). These structures are shown in the supplementary file for a hypothetical longitudinal data set with three subjects (subject 1: present at 4-time points, subject 2: present at 2-time points, and subject 3: present at 3-time points).

Criteria for comparing the predictive performance of longitudinal regression tree algorithms

In the present study, we compared the predictive performance of longitudinal regression tree algorithms under different covariance patterns in the simulated and real datasets. We used the mean of square error (MSE) of residuals, mean absolute difference (MAD), and deviance to evaluate the predictive performance of different longitudinal regression tree algorithms.

$$MSE = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it}-\widehat{y}_{it})^2}{n\times T}$$

$$MAD = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}|y_{it}-\widehat{y}_{it}|}{n\times T}$$

$$Deviance = -2\times log-likelihood$$

y_{it} and $\widehat{y}_{it}$ indicate the observed and predicted response variables for t^{th} observation in the i^{th} individual in the test dataset, respectively.

We anticipated that the best performers would have a smaller MSE, MAD, and deviance. In both simulation and real datasets, the first 70% of observations for *I* subjects were used as a training data set, and the remaining 30% were considered the testing set.

$$\begin{bmatrix} 0.000571 & 0.000495 & 0.000484 & 0.000460 & 0.000513 & 0.000475 & & & & & & \\ 0.000495 & 0.000750 & 0.000584 & 0.000583 & 0.000640 & 0.000596 & & & & & & \\ 0.000484 & 0.000584 & 0.000696 & 0.000605 & 0.000634 & 0.000574 & \cdots & & 0 & & & \\ 0.000460 & 0.000583 & 0.000605 & 0.000741 & 0.000756 & 0.000662 & & & & & & \\ 0.000513 & 0.000640 & 0.000634 & 0.000756 & 0.000904 & 0.000753 & & & & & & \\ 0.000475 & 0.000596 & 0.000574 & 0.000662 & 0.000753 & 0.000804 & & & & & & \\ & & & \vdots & & & \ddots & & \vdots & & & \\ & & & & & & & 0.000571 & 0.000495 & 0.000484 & 0.000460 & 0.000513 & 0.000475 \\ & & & & & & & 0.000495 & 0.000750 & 0.000584 & 0.000583 & 0.000640 & 0.000596 \\ & & & 0 & & & \cdots & 0.000484 & 0.000584 & 0.000696 & 0.000605 & 0.000634 & 0.000574 \\ & & & & & & & 0.000460 & 0.000583 & 0.000605 & 0.000741 & 0.000756 & 0.000662 \\ & & & & & & & 0.000513 & 0.000640 & 0.000634 & 0.000756 & 0.000904 & 0.000753 \\ & & & & & & & 0.000475 & 0.000596 & 0.000574 & 0.000662 & 0.000753 & 0.000804 \end{bmatrix}$$

Software programs

The R 4.1.0 software was used to code the Ev-RE-EM algorithm, fit other longitudinal regression tree algorithms and longitudinal GWAS, and simulate the data. The RE-EM longitudinal tree algorithm was fitted using the RE-EMtree package [49]. The R code was used to fit the unbiased RE-EM tree algorithm, which is available by the http://people.stern.nyu.edu/jsimonof/unbiasedRE-EM link. BMI phenotype normalization based on the RINT method was done with the RNOmni package [50], and the GMMAT package was used to run longitudinal GWAS [51].

In addition, we employed Plink 1.9 software (http://pngu.mgh.harvard.edu/purcell/plink/) for data quality control and LD calculation [52]. GRM and principal components were calculated to adjust for population classification using the rigorous GCTA (version 1.94.0beta) software (https://yanglab.westlake.edu.cn/software/gcta/) [46]. All analyses were conducted in a robust Linux environment and with the versatile Bash Scripting language. The R code for running the Ev-RE-EM algorithm and simulating the data is provided in the supplementary files.

Results

We conducted a GWAS with BMI as a longitudinal phenotype using 10,945,256 genome-wide SNPs after quality control in TCGS. We used the generalized linear mixed model association tests (GMMAT) to find the associated SNPs by adjusting age, sex, and the first ten principal components to account for the population stratification [53]. The RINT was used for BMI normalization. The Manhattan plot and quantile-quantile plot of this analysis are indicated in Figs. 4 and 5, respectively.

The autocorrelation test for different autocorrelation structures was evaluated in the RE-EM tree algorithm using the likelihood ratio (LR) test, and the tree without an autocorrelation structure was considered the null hypothesis. This test demonstrated that there was no statistical difference between the compound symmetric structure with a constant correlation and.

autocorrelation structure under the RE-EM algorithm ($P=0.999$). Additionally, there was a statistical difference between the first-order autoregressive process and the general correlation structure with $\sigma^2 I_{T_i}$ autocorrelation structure ($P<0.001$ and $P<0.001$, respectively). Therefore, in this study, fitting longitudinal tree algorithms should include the autocorrelation structure of the first-order autoregressive process and the general correlation structure between the within-subject errors.

As represented in Table 1, the results of the real data analysis indicated that the RE-EM algorithm with $\sigma^2 I_{T_i}$ and a compound symmetric structure with a constant correlation as the autocorrelation structures have similar and better performance compared to the use of other autocorrelation structures between the errors within the subjects. Additionally, using a first-order autoregressive process as the autocorrelation structure in the RE-EM algorithm shows the weakest predictive performance (Table 1). The RE-EM algorithm with $\sigma^2 I_{T_i}$ or a compound symmetric structure with a constant correlation as an autocorrelation structure has the best performance.

As shown in Table 1, the predictive performance of unbiased RE-EM and Ev-RE-EM is better under $\sigma^2 I_{T_i}$ and a compound symmetric structure with a constant. Conversely, these algorithms showed the weakest predictive performance under the first-order autoregressive process as the autocorrelation structure. The tree structures of

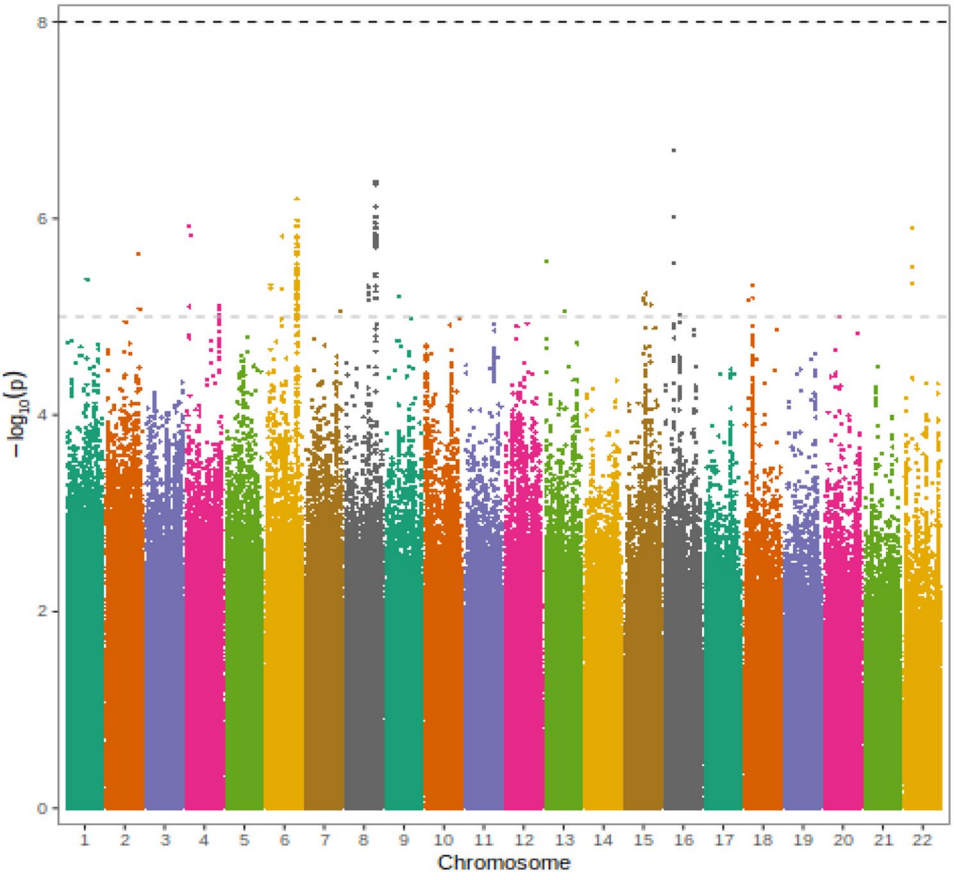

Fig. 4 The Manhattan plot of the BMI as a longitudinal phenotype by adjusting age, sex, and the first 10 principal components in the TCGS using GMMAT

unbiased RE-EM and Ev-RE-EM algorithms with the best predictive performance are shown in Figs. 6, 7, and 8. These algorithms extracted the homogeneous subgroups of observations. The low-risk subgroups based on the unbiased RE-EM, and Ev-RE-EM algorithms are individuals aged ≤ 8 years and ≤ 11 years, respectively. In addition, the women aged > 39, rs112865060= {CC, CG}, and rs11075406 = CC are identified as high-risk subgroups by the unbiased RE-EM tree algorithm (Fig. 6). On the other hand, the high-risk subgroups based on the Ev-RE-EM tree algorithm are women aged > 39, rs75740786 = GA, rs917188199 = AT, and rs184919069 = GC (Fig. 8). tree sizes of RE-EM, unbiased RE-EM, and Ev-RE-EM are 149, 58, and 47, respectively. Thus, Ev-RE-EM generated a smaller, more interpretable tree.

The simulation study results indicated that three longitudinal regression tree algorithms under the AR (1) autocorrelation structure have shown poor predictive performance. On the other hand, the predictive performance of these algorithms under the autocorrelation structure, such as $\sigma^2 I_{T_i}$ and a compound symmetric structure with a constant correlation is almost similar (Table 2).

Discussion

The conventional LME model for analyzing longitudinal or clustered data relies on a parametric linear function that requires a series of assumptions, limiting its applicability in real-world scenarios [5]. Thus, proposing an alternative to parametric models for

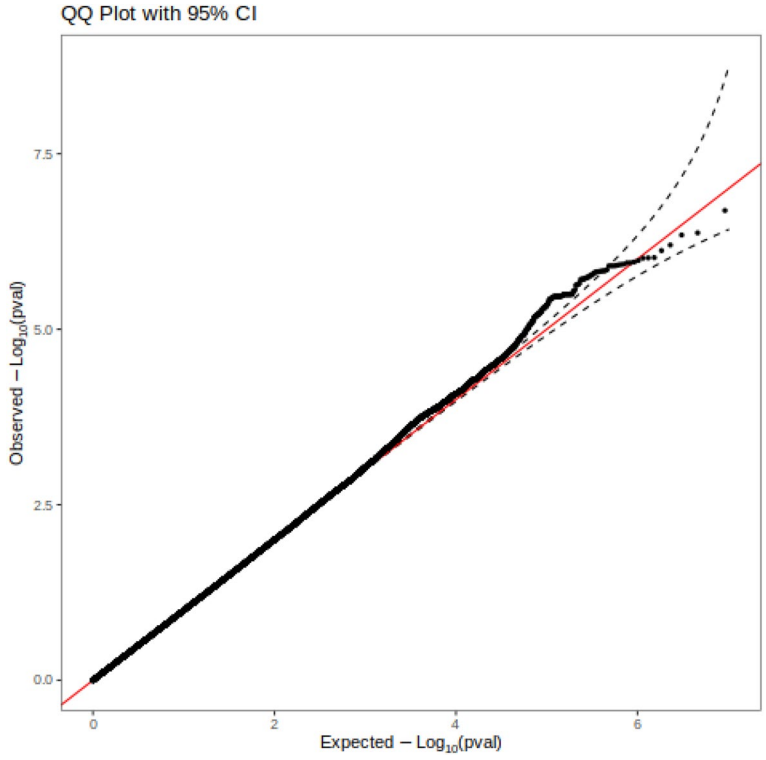

Fig. 5 The quantile-quantile plot of the BMI as a longitudinal phenotype by adjusting age, sex, and the first 10 principal components in the TCGS using GMMAT

Table 1 The predictive performance of the longitudinal regression tree algorithms with and without the autocorrelation structure between within-subject errors for predicting the BMI in TCGS

Longitudinal regression tree algorithm	Autocorrelation structure	MSE	MAD	Deviance
RE-EM	$\sigma^2 I_{T_i}$	5.323	1.728	52094.98
	AR (1)	6.330	1.880	51094.46
	CS	5.323	1.728	52094.98
	C	5.538	1.767	51483.54
Unbiased RE-EM	$\sigma^2 I_{T_i}$	4.023	1.113	51017.86
	AR (1)	5.028	1.195	50236.74
	CS	4.023	1.113	51017.86
	C	4.155	1.165	51211.12
Ev-RE-EM	$\sigma^2 I_{T_i}$	4.073	1.117	51021.67
	AR (1)	5.058	1.20	50256.91
	CS	4.073	1.117	51021.67
	C	4.175	1.169	51218.14

$\sigma^2 I_{T_i}$: variance-covariance diagonal matrix of errors, AR (1): the first-order autoregressive process, CS: compound symmetry structure with a constant correlation, and C: general correlation matrix (unstructured)

longitudinal data analysis is essential. Given the advantages of tree-based methods over parametric models, extending tree algorithms for longitudinal data can address the challenges of analyzing high-dimensional longitudinal data.

Several tree algorithms have been developed to analyze longitudinal or clustered data by extending the well-known CART algorithm [3–6, 35, 40, 54]. Additionally, further refinements have been proposed to address the disadvantages of the CART algorithm. One disadvantage of the CART method is its greedy approach, which limits the exploration of tree space, induces dependence of future splits on previous splits, generates

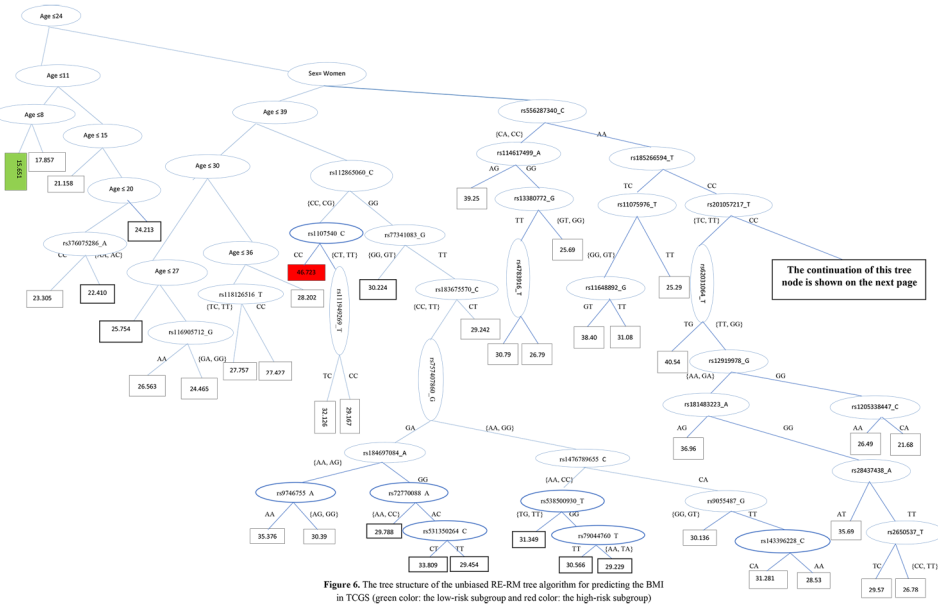

Fig. 6 The tree structure of the unbiased RE-RM tree algorithm for predicting the BMI in TCGS (green color: the low-risk subgroup and red color: the high-risk subgroup)

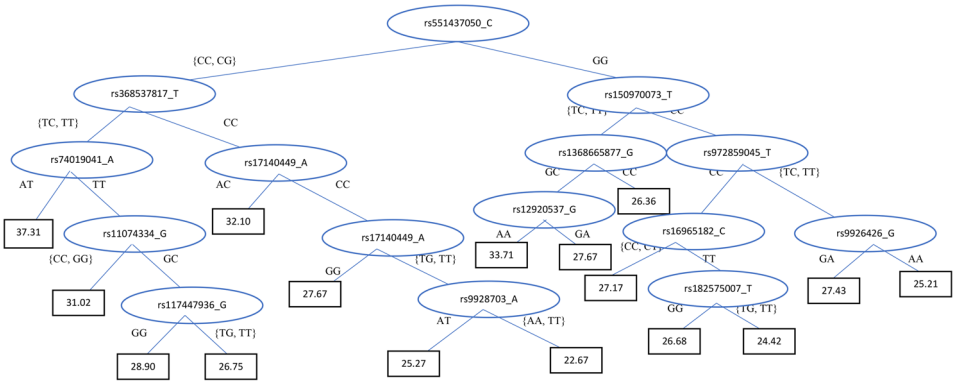

Fig. 7 The continuation of the tree structure in Fig. 6

optimistic error rates, and prevents the search from finding a global optimum [18]. To address these issues, Bayesian tree approaches [55–62] and the Evtree algorithm [18] were proposed as solutions to the limitations of the CART algorithm.

In this study, we address the need for a solution to the greedy problem in analyzing longitudinal or clustered data in GWAS-based studies. Recognizing the significance of this issue, we propose a novel non-parametric algorithm, Ev-RE-EM, which utilizes the Evtree algorithm to estimate the fixed part of the LME model for high-dimensional longitudinal data analysis.

This new longitudinal regression tree algorithm was used for the first time in longitudinal genome-wide association studies (GWAS) to investigate genetic markers affecting BMI as a longitudinal phenotype and to predict this phenotype in TCGS. The predictive performance of this algorithm was compared with previously proposed longitudinal regression tree algorithms such as RE-EM and unbiased RE-EM, under different autocorrelation structures (e.g., first-order autoregressive process, compound symmetric

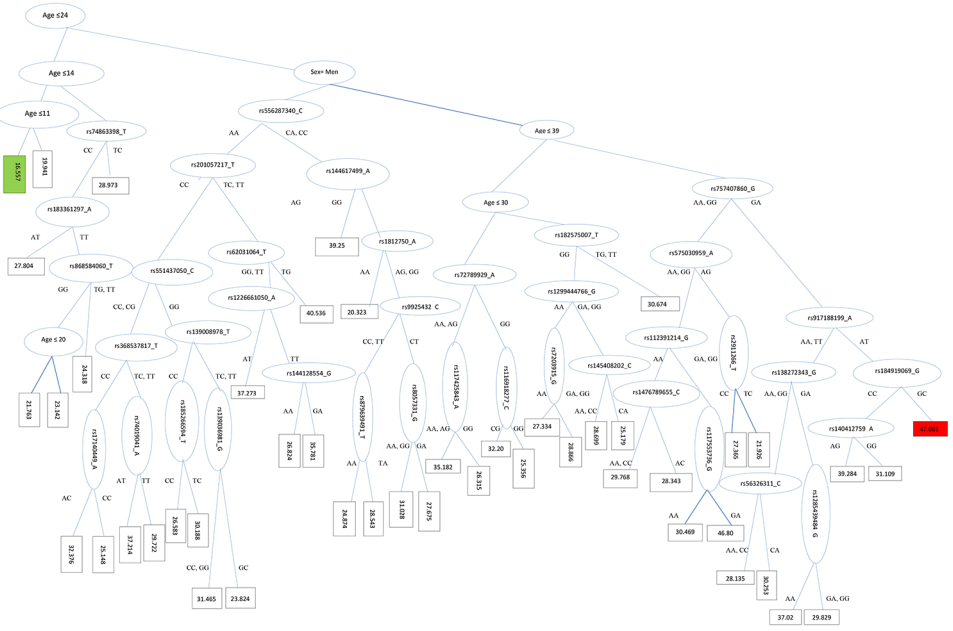

Fig. 8 The tree structure of the Ev-RE-RM tree algorithm for predicting the BMI in TCGS (green color: the low-risk subgroup and red color: the high-risk subgroup

Table 2 The predictive performance of the longitudinal regression tree algorithms with and without the autocorrelation structure between within-subject errors in the simulated dataset

Longitudinal regression tree algorithm	Autocorrelation structure	MSE	MAD	Deviance
RE-EM	$\sigma^2 I_{T_i}$	0.3727947	0.4524195	28636.44
	AR (1)	24.52032	3.943993	29912.64
	CS	0.3782275	0.4584203	28750.63
RE-EM Unbiased	$\sigma^2 I_{T_i}$	0.2657871	0.4305183	25739.56
	AR (1)	17.03199	2.813750	27807.73
	CS	0.2669871	0.4354071	25946.75
Ev-RE-EM	$\sigma^2 I_{T_i}$	0.2735940	0.4317192	25847.34
	AR (1)	17.57367	2.835460	27821.97
	CS	0.2747831	0.4384521	25997.92

$\sigma^2 I_{T_i}$: variance-covariance diagonal matrix of errors, AR (1): first-order autoregressive process, CS: compound symmetry structure with a constant correlation (unstructured)

structure, general correlation matrix) using criteria such as MSE, MAD, and deviation, in both real and simulated datasets.

The predictive performance of the longitudinal regression tree algorithms on the real dataset showed that the RE-EM, unbiased RE-EM, and Ev-RE-EM algorithms under $\sigma^2 I_{T_i}$ and a compound symmetric structure with a correlation constant have similar performance, as the LR test did not show a statistically significant difference between these two structures under the RE-EM model. Additionally, the predictive performance of the longitudinal tree algorithms under the different structures mentioned was similar in both real and simulated datasets.

The results from both real and simulated datasets showed that the unbiased RE-EM and Ev-RE-EM algorithms perform better than the RE-EM algorithm, likely due to the strategy of the RE-EM algorithm being based on the CART algorithm, which has weaknesses such as biased splits and greedy problems. The predictive performance of

unbiased RE-EM and Ev-RE-EM was almost similar on real and simulated datasets. However, Ev-RE-EM explored the tree space more thoroughly, creating a smaller, more interpretable tree. Moreover, this method randomly selects splitting variables and rules, eliminating bias in selecting splitting rules. Therefore, compared to the RE-EM algorithm, Ev-RE-EM is a good alternative for longitudinal data analysis.

Additionally, the simulation study results showed that longitudinal regression tree algorithms do not converge under the general correlation structure. This non-convergence is common under this structure because it requires a general correlation matrix and converges with difficulty.

The result of the longitudinal GWAS for the BMI phenotype showed no significant SNP with a p-value less than 10^{-8}. Longitudinal tree algorithms can be used to determine effective SNPs for longitudinal phenotypes of interest. Unlike GWAS, these algorithms do not have a univariate view and can discover interactions between SNPs. The target population is usually heterogeneous when examining longitudinal phenotype changes. Thus, it is essential to extract homogeneous subgroups using tree algorithms, as the values of predictions or estimates differ between homogeneous subgroups [40]. Furthermore, GWAS results depend on sample size, and the sample size used in this study is small compared to other GWAS. Thus, tree algorithms, unaffected by sample size, can be a good alternative.

Despite the strengths of this study, a weakness is that more research is needed to confirm the findings, as no similar studies have been done in the Iranian population to determine effective SNPs on BMI longitudinal phenotype. However, the results did show that women are in the high-risk subgroup for obesity, a finding supported by other studies [63, 64]. Another weakness is that using longitudinal regression tree algorithms with SNPs of all autosomal chromosomes requires a long time to run.

Capitaine et al. (2019) extended random forests to analyze high-dimensional longitudinal datasets [54]. They used the semi-parametric stochastic mixed effects model to account for the correlation structure between repeated measurements instead of the LME model. They proposed stochastic RE-EM (SRE-EM), SRE-EMforest, stochastic mixed effects regression trees (SMERT), and stochastic mixed effects random forests (SMERF). These algorithms are available in the LongituRF package in R. We also used the algorithms in the LongituRF package and boosted multivariate trees [65] for data analysis but these algorithms could not run on high-dimensional data due to difficulties in achieving convergence under our current computational framework. Louis Capitaine et al. (2021) ran longitudinal random forests to analyze high-dimensional data containing 20,000 gene transcripts [54]. These predictor variables were continuous, but in our study, the SNPs are qualitative variables, which causes a computational burden.

The present study has some limitations, such as the inability to analyze whole genome data and the time computationally required by the RE-EM algorithm, which is smaller than that of unbiased RE-EM and Ev-RE-EM. On the other hand, the RE-EM algorithm suffers from problems such as biased splits and greedy search. So, future works are proposed to improve longitudinal tree-based methods for analyzing whole genome data.

Some decision trees, such as CART [15], CRUISE [24, 25], QUEST [26], and GUIDE [27], can deal with missing data. Among longitudinal tree algorithms, only RE-EM can deal with missing values in the response variables [5]. There is no remedial in other longitudinal regression tree algorithms to impute these values, which is a limitation of the

present study. We propose that future work be done to solve this problem. Jahangiri et al. (2023) assessed a wide range of missing imputation approaches in longitudinal data, and researchers can use these approaches before using the unbiased RE-EM and EV-RE-EM algorithms to analyze longitudinal data [66]. On the other hand, there are no missing values in the SNPs because they are obtained from the imputation of the chip dataset of the genetic study.

In addition, future research can extend other regression tree algorithms under the framework of Capitaine et al. (2019) for longitudinal data analysis. Stegmann et al. (2018) extended the longRPart algorithm and suggested a nonlinear longitudinal recursive partitioning (nLRP) to predict change trajectories using the nonlinear LME based on cluster-level covariates [67]. This algorithm is available in longRPart2. Future research also can focus on extending nonlinear LME under the framework of the conditional inference tree by Hothorn et al. (2006) [28].

Nestler and Humberg (2021) combined an extended mixed-effect location scale (E-MELS) with the CART algorithm of Breiman et al. (1984) to propose E-MELS trees for hierarchical data analysis [68]. The code for this algorithm can be found at https://osf.io/53scf/. Future research can work on extending E-MELS trees based on different tree algorithms.

Conclusion

The results showed that the unbiased RE-EM and Ev-RE-EM algorithms outperformed the RE-EM algorithm. Additionally, the Ev-RE-EM algorithm produced smaller and more interpretable trees by uncovering more tree structures. Since algorithm performance varies across datasets, researchers should test different algorithms on the dataset of interest and select the best-performing one.

Accurately predicting and diagnosing an individual's genetic profile is crucial in medical studies. The model with the highest accuracy should be used to enhance understanding of the genetics of complex traits, improve disease prevention and diagnosis, and aid in treating complex human diseases. Identifying genetic factors, which are immutable through therapeutic intervention, can help screen and prevent at-risk individuals. This also enables effective, personalized treatment based on individuals' genetic conditions at the clinical level.

Abbreviations

BMI	Body mass index
SNPs	Single nucleotide polymorphisms
GWAS	Genome-wide association studies
LME	Linear mixed-effects model
RE	EM-Random effects expectation-maximization
TCGS	Tehran cardiometabolic genetic study

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13040-025-00437-w.

Supplementary Material 1

Acknowledgements

The authors would like to express their gratitude to the staff and participants in the TCGS project.

Author contributions

Conceptualization: Mina Jahangiri and Anoshirvan Kazemnejad; Formal analysis: Mina Jahangiri, Mahdi Akbarzadeh, and Shayan Mostafaei; Methodology: Mina Jahangiri, Mahdi Akbarzadeh, Anoshirvan Kazemnejad, Maryam S Daneshpour,

and Davood Khalili; Medical consultant: Maryam S Daneshpour; Software and data simulation: Mina Jahangiri, Keith Goldfeld, Mahdi Akbarzadeh, and Mehdi Momen; Writing-original draft: Mina Jahangiri and Keith Goldfeld; Supervision: Anoshirvan Kazemnejad and Mahdi Akbarzadeh; All authors reviewed and accepted the manuscript.

Funding
No funding was received for this study.

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
The ethical committee approved this study at the Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences (Research Approval Code: 28778 & Research Ethical Code: IR.SBMU.ENDOCRINE.REC.1400.084). All participants provided written informed consent. This study was performed according to the Declaration of Helsinki.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 8 January 2025 / Accepted: 3 March 2025
Published online: 19 March 2025

References

1. Fitzmaurice GM, Laird NM, Ware JH. Applied longitudinal analysis: Wiley. 2012.
2. Jiang L, Zheng Z, Qi T, Kemper KE, Wray NR, Visscher PM et al. A resource-efficient tool for mixed model association analysis of large-scale data. 2019;51(12):1749–55.
3. Hajjem A, Bellavance F, Larocque D. Mixed effects regression trees for clustered data. Stat Probab Lett. 2011;81(4):451–9.
4. Hajjem A, Bellavance F, Larocque D. Mixed-effects random forest for clustered data. J Stat Comput Simul. 2014;84(6):1313–28.
5. Sela RJ, Simonoff JS. RE-EM trees: a data mining approach for longitudinal and clustered data. Mach Learn. 2012;86(2):169–207.
6. Fu W, Simonoff JS. Unbiased regression trees for longitudinal and clustered data. Comput Stat Data Anal. 2015;88:53–74.
7. Geurts P, Irrthum A, Wehenkel L. Supervised learning with decision tree-based methods in computational and systems biology. Mol Biosyst. 2009;5(12):1593–605.
8. De'ath G, Fabricius KE. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology. 2000;81(11):3178–92.
9. Lemon SC, Roy J, Clark MA, Friedmann PD, Rakowski W. Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. Ann Behav Med. 2003;26(3):172–81.
10. Feldesman MR. Classification trees as an alternative to linear discriminant analysis. Am J Phys Anthropology: Official Publication Am Association Phys Anthropologists. 2002;119(3):257–75.
11. Malehi AS, Jahangiri M. Classic and bayesian Tree-Based methods. Enhanced Expert Syst. 2019:27.
12. Jahangiri M, Khodadi E, Rahim F, Saki N, Saki Malehi A. Decision-tree-based methods for differential diagnosis of β-thalassemia trait from iron deficiency anemia. Expert Syst. 2017;34(3):e12201.
13. Rahim F, Kazemnejad A, Jahangiri M, Malehi AS, Gohari K. Diagnostic performance of classification trees and hematological functions in hematologic disorders: an application of multidimensional scaling and cluster analysis. BMC Med Inf Decis Mak. 2021;21(1):1–13.
14. Jahangiri M, Rahim F, Saki N, Saki Malehi A. Application of Bayesian Decision Tree in Hematology Research: Differential Diagnosis of β-Thalassemia Trait from Iron Deficiency Anemia. Computational and Mathematical Methods in Medicine. 2021;2021.
15. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees: CRC. 1984.
16. Gray JB, Fan G. Classification tree analysis using TARGET. Comput Stat Data Anal. 2008;52(3):1362–72.
17. Fan G, Gray JB. Regression tree analysis using TARGET. J Comput Graphical Stat. 2005;14(1):206–18.
18. Grubinger T, Zeileis A, Pfeiffer K-P, evtree. Evolutionary learning of globally optimal classification and regression trees in R. Working Papers in Economics and Statistics. 2011.
19. Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat Sci. 2001;16(3):199–231.
20. Loh WY. Tree-structured classifiers. Wiley Interdisciplinary Reviews: Comput Stat. 2010;2(3):364–9.
21. Loh WY. Classification and regression trees. Wiley Interdisciplinary Reviews: Data Min Knowl Discovery. 2011;1(1):14–23.
22. Breiman L. Bagging predictors. Mach Learn. 1996;24(2):123–40.
23. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
24. Kim H, Loh W-Y. Classification trees with bivariate linear discriminant node models. J Comput Graphical Stat. 2003;12(3):512–30.
25. Kim H, Loh W-Y. Classification trees with unbiased multiway splits. J Am Stat Assoc. 2001;96(454):589–604.
26. Loh W-Y, Shih Y-S. Split selection methods for classification trees. Statistica sinica. 1997:815–40.
27. Loh W-Y. Improving the precision of classification trees. Annals Appl Stat. 2009:1710–37.
28. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: A conditional inference framework. J Comput Graphical Stat. 2006;15(3):651–74.

29. Chan K-Y, Loh W-Y. An algorithm for Building accurate and comprehensible logistic regression trees. J Comput Graphical Stat. 2004;13(4):826–52.
30. Segal MR. Tree-structured methods for longitudinal data. J Am Stat Assoc. 1992;87(418):407–18.
31. Abdolell M, LeBlanc M, Stephens D, Harrison R. Binary partitioning for continuous longitudinal data: categorizing a prognostic variable. Stat Med. 2002;21(22):3395–409.
32. Stewart S, Abdolell M, Stewart MS. Package 'longRPart'.
33. Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics. 1982:963–74.
34. Loh W-Y, Zheng W. Regression trees for longitudinal and multiresponse data. Annals Appl Stat. 2013;7(1):495–522.
35. Eo S-H, Cho H. Tree-structured mixed-effects regression modeling for longitudinal data. J Comput Graphical Stat. 2014;23(3):740–60.
36. Zeileis A, Hothorn T, Hornik K. Model-based recursive partitioning. J Comput Graphical Stat. 2008;17(2):492–514.
37. Fokkema M, Edbrooke-Childs J, Wolpert M. Generalized linear mixed-model (GLMM) trees: A flexible decision-tree method for multilevel and longitudinal data. Psychother Res. 2021;31(3):329–41.
38. Fokkema M, Smits N, Zeileis A, Hothorn T, Kelderman H. Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. Behav Res Methods. 2018;50(5):2016–34.
39. Fokkema M, Zeileis A, Fokkema MM. Package 'glmertree'. 2019.
40. Kundu MG, Harezlak J. Regression trees for longitudinal data with baseline covariates. Biostatistics Epidemiol. 2019;3(1):1–22.
41. Kundu MG, Kundu MMG. Package 'LongCART'. 2022.
42. Daneshpour MS, Akbarzadeh M, Lanjanian H, Sedaghati-Khayat B, Guity K, Masjoudi S, et al. Cohort profile update: Tehran cardiometabolic genetic study. Eur J Epidemiol. 2023;38(6):699–711.
43. Peng S, Zhu Y, Xu F, Ren X, Li X, Lai M. FTO gene polymorphisms and obesity risk: a meta-analysis. BMC Med. 2011;9:1–15.
44. Liu C, Mou S, Cai Y. FTO gene variant and risk of overweight and obesity among children and adolescents: a systematic review and meta-analysis. PLoS ONE. 2013;8(11):e82133.
45. Browning BL, Zhou Y, Browning SR. A one-penny imputed genome from next-generation reference panels. Am J Hum Genet. 2018;103(3):338–48.
46. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88(1):76–82.
47. Jiang L, Zheng Z, Qi T, Kemper KE, Wray NR, Visscher PM, et al. A resource-efficient tool for mixed model association analysis of large-scale data. Nat Genet. 2019;51(12):1749–55.
48. McCaw ZR, Lane JM, Saxena R, Redline S, Lin X. Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. Biometrics. 2020;76(4):1262–72.
49. Sela R, Simonoff J, Sela MR, Suggests A. Package 'REEMtree'. 2023.
50. McCaw Z. RNOmni: rank normal transformation omnibus test. R Package. 2019;861.
51. Chen H, Conomos M, Pham D, Gilly A, Gentleman R, Ihaka R. Package 'GMMAT'. 2023.
52. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.
53. Chen H, Wang C, Conomos MP, Stilp AM, Li Z, Sofer T, et al. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. Am J Hum Genet. 2016;98(4):653–66.
54. Capitaine L, Genuer R, Thiébaut RJ. Random forests for high-dimensional longitudinal data. Smimr. 2021;30(1):166–84.
55. Chipman HA, George EI, McCulloch RE. Bayesian treed models. Mach Learn. 2002;48(1–3):299–320.
56. Chipman H, McCulloch RE. Hierarchical priors for bayesian CART shrinkage. Stat Comput. 2000;10(1):17–24.
57. O'Leary RA, Murray JV, Low Choy SJ, Mengersen KL. Expert elicitation for bayesian classification trees. J Appl Probab Stat. 2008;3(1):95–106.
58. O'Leary RA. Informed statistical modelling of habitat suitability for rare and threatened species. Queensland University of Technology; 2008.
59. Denison DG, Mallick BK, Smith AF. A bayesian CART algorithm. Biometrika. 1998;85(2):363–77.
60. Chipman H, George E, McCulloch R. Bayesian treed generalized linear models. Bayesian Stat. 2003;7:323–49.
61. Chipman HA, George EI, McCulloch RE. Bayesian CART model search. J Am Stat Assoc. 1998;93(443):935–48.
62. Wu Y, Tjelmeland H, West M, Bayesian CART. Prior specification and posterior simulation. J Comput Graphical Stat. 2007;16(1):44–66.
63. Ng M, Fleming T, Robinson M, Thomson B, Graetz N, Margono C, et al. Global, regional, and National prevalence of overweight and obesity in children and adults during 1980–2013: a systematic analysis for the global burden of disease study 2013. Lancet. 2014;384(9945):766–81.
64. Abarca-Gómez L, Abdeen ZA, Hamid ZA, Abu-Rmeileh NM, Acosta-Cazares B, Acuin C, et al. Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in 128· 9 million children, adolescents, and adults. Lancet. 2017;390(10113):2627–42.
65. Pande A, Li L, Rajeswaran J, Ehrlinger J, Kogalur UB, Blackstone EH, et al. Boosted Multivar Trees Longitud Data. 2017;106:277–305.
66. Jahangiri M, Kazemnejad A, Goldfeld KS, Daneshpour MS, Mostafaei S, Khalili D et al. A wide range of missing imputation approaches in longitudinal data: a simulation study and real data analysis. 2023;23(1):161.
67. Stegmann G, Jacobucci R, Serang S, Grimm KJJM. Recursive Partitioning Nonlinear Models Change. 2018;53(4):559–70.
68. Nestler S, Humberg SJ. A Lasso and a regression tree mixed-effect model with random effects for the level, the residual variance, and the autocorrelation. 2022;87(2):506–32.

Publisher's note