OXFORD

Briefings in Bioinformatics, 2023, **24(2)**, 1–11
https://doi.org/10.1093/bib/bbad002
Advance access publication date 18 January 2023
Review

A review on longitudinal data analysis with random forest

Jianchang Hu and Silke Szymczak

Corresponding author. Silke Szymczak, Institute of Medical Biometry and Statistics, University of Lübeck, Ratzeburger Allee 160, 23562, Lübeck, Germany. Tel.: (49)0451-500-50600; Fax: (49)0451-500-50604. E-mail: silke.szymczak@uni-luebeck.de

Abstract

In longitudinal studies variables are measured repeatedly over time, leading to clustered and correlated observations. If the goal of the study is to develop prediction models, machine learning approaches such as the powerful random forest (RF) are often promising alternatives to standard statistical methods, especially in the context of high-dimensional data. In this paper, we review extensions of the standard RF method for the purpose of longitudinal data analysis. Extension methods are categorized according to the data structures for which they are designed. We consider both univariate and multivariate response longitudinal data and further categorize the repeated measurements according to whether the time effect is relevant. Even though most extensions are proposed for low-dimensional data, some can be applied to high-dimensional data. Information of available software implementations of the reviewed extensions is also given. We conclude with discussions on the limitations of our review and some future research directions.

Keywords: machine learning, longitudinal data, repeated measurements, clustered data, multivariate response

Introduction

In many scientific fields including medicine, life sciences and economics, the analysis of longitudinal data plays a vital role. Take precision medicine as an example. The goal of precision medicine is to provide customized treatments to patients based on their characteristics and thus to improve treatment efficiency while avoiding serious side effects [1–3]. With recent technological advances, large-scale genetic and other molecular data can now be collected. Along with demographic and clinical profiles they characterize each patient under different aspects. Many of these measurements, however, change over time, often depending on disease activity, treatment, comorbidities and other environmental factors. Consequently, it is important to measure them for the same patient repeatedly over time, and this leads to longitudinal data, where a single observation captures the measurements at a specific time point for a patient.

Depending on the research question, the study design and the outcome of interest, multiple longitudinal data formats can be envisioned. Predictors might be available for a single time point only, such as at baseline visit, or are time-invariant, which is the case for genetic variants. Alternatively, predictors are measured multiple times during a study. Similarly, the outcome can be determined at a single time point. Examples include response to treatment at the end of therapy or after a prespecified follow-up time. But it might also be of interest to predict the outcome over time such as disease activity or severity. Furthermore, the data format is related to the study design where the same number of measurements at fixed time points is taken for each subject or data from a varying number of irregularly spaced time points are available; the latter is often encountered in observational studies.

Table 1. General structure of longitudinal data

Subject	Time	Responses			Predictors		
1	1	y_{111}	...	y_{11m}	x_{111}	...	x_{11p}
1	2	y_{121}	...	y_{12m}	x_{121}	...	x_{12p}
.	.	.	.	.	.	.	.
1	n_1	y_{1n_11}	...	y_{1n_1m}	x_{1n_11}	...	x_{1n_1p}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
N	1	y_{N11}	...	y_{N1m}	x_{N11}	...	x_{N1p}
N	2	y_{N21}	...	y_{N2m}	x_{N21}	...	x_{N2p}
.	.	.	.	.	.	.	.
N	n_N	y_{Nn_N1}	...	y_{Nn_Nm}	x_{Nn_11}	...	x_{Nn_1p}

In general, a longitudinal data set can be formatted as in Table 1. Here in total, there are N subjects, for each of them, n_i, $i = 1, \ldots, N$ observations are measured, and each observation consists of measurements on m responses and p predictors.

Analyzing longitudinal data is not an easy task. The most distinct feature of longitudinal data is the repeated measurements from the same subject. This inevitably leads to clustered and correlated observations. The clustering effect is due to individual characteristics. For instance, average response to a drug could vary from patient to patient. In the meantime, if repeated measurements are collected over a period of time, then there could be serial correlation among measurements.

Furthermore the observation time for the longitudinal data can be either equally spaced or irregularly spaced, which may affect the approach that can be used for the analysis. Visits at every other month would lead to equally spaced observations, while

Jian chang Hu is a postdoctoral research fellow at the Institute of Medical Biometry and Statistics, University of Lübeck, Lübeck, Germany. He received his Ph.D. degree in Statistics from the University of Wisconsin-Madison. His current research focuses on random forest extensions with applications to multi-omics data. **Silke Szymczak** is professor of Genetic Epidemiology at the Institute of Medical Biometry and Statistics, University of Lübeck, Lübeck, Germany. Her research interests include the development and systematic evaluation of machine learning methods, with a particular focus on random forests in the context of omics data.
Received: August 31, 2022. **Revised:** December 12, 2022. **Accepted:** December 31, 2012

following up at 6 months, 1 year and 2 years after the treatment provides an example of irregularly spaced observations. Additionally, irregular spacing can also occur in observational studies when there are no prespecified follow-up times. Apart from that, missing values can pose great challenges as they can turn an equally spaced observation schedule into irregular. More importantly, they may carry vital information when the missingness could be related to the value of the variable. More discussions on the characteristics of longitudinal data can be found in the classic textbooks [4, 5].

Despite the difficulties introduced by longitudinal data, they bring rich information. In the context of precision medicine, with longitudinal data, clinicians can better understand disease progression, especially of chronic diseases, so that patients can be properly stratified and treatment plans can be tailored accordingly [6–8]. Furthermore, repeated measurements allow the patients' treatment responses to be captured more accurately, so that effective therapies can be implemented and evaluated.

The development of prediction models with longitudinal data using statistical or machine learning (ML) approaches is also crucial [9]. One of the state-of-the-art ML methods for the development of prediction models is the random forest (RF) algorithm [10]. It is a nonparametric approach that can accommodate different types of responses such as categorical or quantitative outcomes and survival times [11]. Moreover, it can work with predictors of various scales or distributions and is suited for applications in high-dimensional settings where the number of predictors can be larger than the number of observations [12, 13]. Thus, it is very suitable for analyzing complex data such as omics data, which are often high-dimensional, plus metabolite and protein levels are usually skewed and left censored by limits of detection, and microbiome abundances often exhibit an excess of zeros. Furthermore, tree-based methods form data-driven subgroups of samples, which can be beneficial for patient stratification. Via the so-called variable importance measures, the method can also highlight the relevance of each predictor [10]. This could be especially handy for pharmacogenomics [14, 15], where potential genetic variants associated with drug response phenotypes can be identified. Svetnik *et al.* (2004) [16], for example, demonstrate that RF provides comparable or better prediction for compound's biological activity in drug discovery process compared with conventional methods such as partial least square and support vector machine.

However, as with other ML methods, the RF algorithm assumes that observations are independently sampled from a population. Conducting statistical analysis on longitudinal data without considering the dependency among observations could lead to biased inference due to underestimated standard errors in linear models [17] and spurious subgroup identification and inaccurate variable selection in tree-based methods [18, 19]. Furthermore, classification methods that match the data structure and thus properly handle the correlation due to repeated measurements have better prediction performance [20].

Therefore, in this review, we will present a range of extensions of the standard RF algorithm for the analysis of longitudinal data. Even though we mention the applications of these methods within the context of precision medicine, our review is not a summary on studies analyzing longitudinal data, but rather provides an overview of available methods and their implementations. We limit our attention to RF based on the classification and regression tree (CART, [21]) with a focus on prediction of categorical and quantitative outcomes. In section 2 we consider the univariate response longitudinal scenario. We start with a short review on the standard RF in subsection 2.1. Following that, in subsection 2.2 the case of repeated measurements or clustered data is investigated. Subsection 2.3 then presents methods that incorporate time effect into modeling. Section 3 focuses on extensions of RF algorithm suitable for multivariate responses. We provide information on the currently available implementations of the reviewed methods in section 4. We conclude with a discussion in section 5.

Univariate response longitudinal data

We start with the simple scenario where we have univariate response longitudinal data; that is $m = 1$ in Table 1. We first briefly review the standard RF algorithm as a prediction model. Several RF extension methods are then presented and discussed, which are categorized by their ways of incorporating the time effect.

Standard RF algorithm

RF is an ensemble of decision trees where each tree is built from a bootstrapped version of the training data set. Each tree is grown via the principle of repetitive partition where starting from the root node, the same node splitting procedure is applied repetitively until certain stopping rules are met. Its power in prediction comes from the aggregation of many weaker learners (decision trees). The performance is especially good if the correlations between trees in the forest are low. More detailed descriptions and discussions on RF can be found in [10, 22].

For a binary decision tree such as CART, the node splitting process consists of selecting a splitting variable and determining the splitting rule. The guiding principle for node splitting is to minimize the impurity of response values in each node, which is often measured by the Gini index if the response variable is categorical or by the variance if it is quantitative. The growth of each decision tree ends if the nodes to split are already pure (all samples within the node come from the same class or have the same response value) or other predetermined stopping rules are met. The nodes in the final layer of a tree are called leaves and are used for prediction of new observations. More detailed discussions on CART can be found in [21].

To make prediction with RF, an observation goes through every decision tree in the forest. The final prediction for the observation from the RF is made either by majority voting or averaging, based on results from all decision trees in the forest. Because the RF algorithm uses bootstrap samples to grow each decision tree, some observations are left out in the construction of a given tree. By treating these out-of-bag (OOB) samples as observations needed to be predicted, it can, therefore, provide an estimate of prediction error of the constructed forest.

In addition, the so-called variable importance measure can be obtained for each predictor, which measures its relevance to prediction. Thus, for high-dimensional dataset such as omics data, variable selection procedures based on variable importance measure are possible (see [23] and the reference therein for a description and comparison of various variable selection procedures based on variable importance measure).

Although it is possible to directly utilize the standard RF algorithm for longitudinal data analysis, it may suffer from several problems. As shown in Figure 1, bootstrapped samples for different decision trees may have a high chance to include observations from every subject. This may cause correlated or even homogeneous trees to deteriorate the prediction performance. In addition, the estimated prediction error based on OOB samples is often too

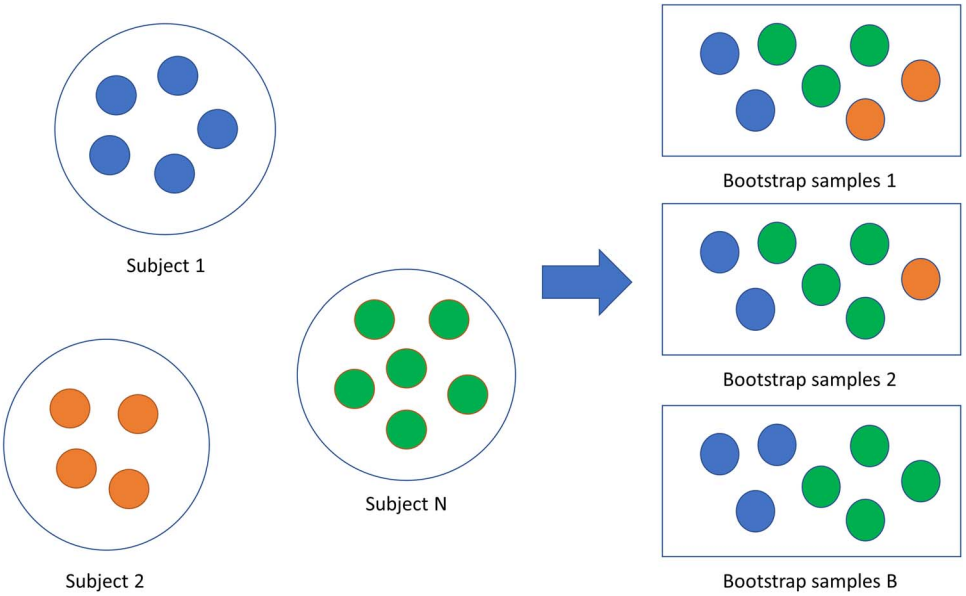

Figure 1. Illustration of bootstrap samples used to construct decision trees in standard RF when it is applied to clustered data.

optimistic due to the high similarity between the observations from the same subject [24]. Therefore, there is a need to build extensions of standard RF for longitudinal data analysis.

Clustered data

In some applications, the observations are made repetitively on the same subject as duplications. This results in clustered data setting. In such setting, the data still follow the general format shown in Table 1, but there is hardly any time effect. In other words, the ordering of the observations from the same subject can be ignored and is not considered in training the prediction model. Hence, one model for clustered data can be written as follows.

$$y_{ij} = \mu_i + \varepsilon_{ij}, \tag{1}$$

where μ_i reflects the mean value of subject i, and ε_{ij} are random fluctuations with mean 0 and independent from each other for all $j = 1, \ldots, n_i$ and across all $i = 1, \ldots, N$. The clustering effect, therefore, is the consequence of the shared mean value for observations from the same subject.

Averaging

One intuitive approach to deal with the aforementioned clustering effect of repeated measurements is to take the average of replicated data for each subject. This then brings the data structure back to the usual one-subject-one-observation scenario and retain the needed independence for standard RF algorithm. Vlahou *et al.* (2004) [25] takes this approach to analyze mass spectrometry data for protein profiling in urine.

Despite the simplicity, this approach suffers from a loss of information. The intrasubject variation is averaged out. Moreover, this approach also masks the imbalance design. Different subjects could contribute different numbers of observations in the original data set, as in Table 1, n_i could be different for $i = 1, \ldots, N$, which may be due to some characteristics of the subjects and may carry underlying distributional information. However, after averaging out the repeated measurements, each subject now makes equal contribution to the training data set. This can have potential effects on the prediction efficiency and variable selection. Karpievitch *et al.* (2009) [24] showed that this approach, when compared with the standard RF, is more sensitive to the total number of subjects N; reduction in N leads to poorer prediction and variable selection. The averaging approach may also be difficult to use when classification and categorical predictors are concerned. For a given patient, his/her cholesterol level based on different blood samples may vary that could lead to different categorization; one observation falls into normal level and another belongs to high level. The averaging approach needs to average all observations of this patient to end up with a subject-level measurement for analysis. However, when different observation-level measurements from the same subject fall in different categories, this averaging would be impossible.

Subject-level bootstrapping

To overcome the disadvantages that averaging approach have, extensions that can utilize all observations are designed. Karpievity *et al.* (2009) [24] proposed the subject-level bootstrapping strategy to replace the original one, and the resulting algorithm is named as RF++. Specifically, when drawing bootstrap samples to construct decision trees in an RF, instead of resampling at the observation level, as shown in Figure 2, bootstrap resampling at the subject level is performed and all observations from the selected subjects are included as in-bag observations.

Adler *et al.* (2011a) and Adler *et al.* (2011b) [26, 27] further extended this idea to a two-stage bootstrapping strategy. Firstly, one subject i is chosen randomly and all associated observations are in bag. Afterwards for each chosen subject *i* the training samples are chosen by randomly selecting one observation from all n_i measurements. Adler et al. (2011b) [27] showed in their simulation studies that subject-level resampling based on one observation per subject yields the best prediction results compared to the standard RF, averaging approach and RF++, although one should also notice that different settings may lead to different results and there could be cases where the other methods are more preferable.

The adoption of subject-level bootstrapping avoids the problem of potentially exposing individual trees to all subjects.

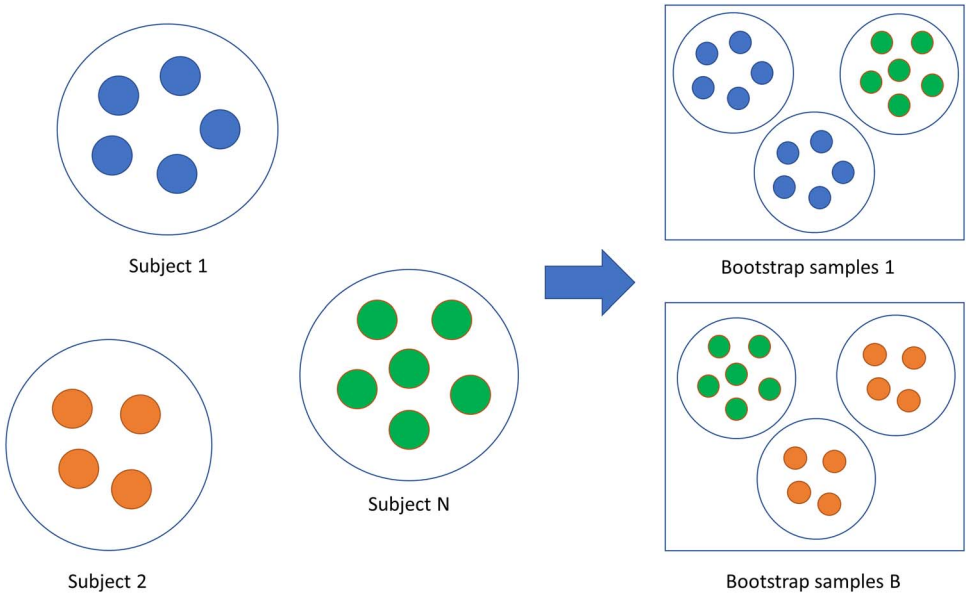

Figure 2. Illustration of subject-level bootstrap samples used to construct decision trees in RF++.

The two-stage bootstrapping strategy could further mitigate the negative effect the intrasubject correlation casts on the prediction performance; when only one observation per subject is selected, even though the same subject might be used in construction of different trees, likely different observations are selected for the training of different trees, which further reduces the similarity between trees.

Besides the usual observation-level classification, Karpievitch *et al.* (2009) [24] showed that classification at subject level is also possible. A majority vote can be performed across the observations belonging to the same subject to result in the subject classification. With such results, a subject-level misclassification rate estimate based on OOB samples is also made possible. This information may be more beneficial and easier to interpret in clinical trials.

However, as pointed out in Hajjem *et al.* (2014) [28], the subject-level bootstrapping only adjusts the sampling method for clustering; thus, no random effects are incorporated in the modeling as well as prediction. Furthermore, for longitudinal studies where time plays a role, this strategy cannot fully utilize the information contained in the data set.

Time effect considered

For many research questions, not only values of predictors at the current time point, but also from the past are helpful, sometimes even crucial, for a good prediction performance. A large value of a particular biomarker might be relevant if it had rather small values in the past, pointing to an early change on the molecular level. Therefore, in this section, we would like to review several RF extensions that take time effect into consideration.

Historical RF

The historical RF, proposed by Sexton and Laake (2018) [29], is an approach that explicitly considers the history of predictors. Assume that we have training data $\{y_{ij}, t_{ij}, X_{ij}\}$, $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$. Here y_{ij} denotes the response, X_{ij} the vector of predictors and t_{ij} the time of the j-th observation on the i-th subject. The method estimates a model for the response y_{ij} using both $(t_{ij}; X_{ij})$ (the observations concurrent with y_{ij}) and all preceding observations of the i-th subject up to (but not including) time t_{ij}. Thus, for a time-varying predictor, its historical information along with its current value are both used for modeling. For a time-invariant predictor, of course, only its current value is used as in the standard RF.

For time-invariant predictors the standard splitting procedure is adopted when constructing each decision tree. In case of a time-varying predictor, its historical information, i.e. values within a specific time interval before the time concurrent with y_{ij}, is first represented by a summary function. One exemplary such function for subject i at time point j counts the number of past observation values, including both response and predictor variables, that are measured at a maximum of η_1 units of time before the current time point j and smaller than η_2, i.e.

$$s_c(\eta; \bar{z}_{ijk}) = \sum_{t_{il} \in [t_{ij} - \eta_1, t_{ij})} I(z_{ilk} < \eta_2), \tag{2}$$

where $\bar{z}_{ij} = \{z_{il} = (y_{il}, x_{il}) : t_{il} < t_{ij}\}$ denotes the past observations of subject i prior time j, and $\bar{z}_{ijk}$ is its k-th component. This aggregation results in a single number per observation and variable for a fixed value of $\eta = (\eta_1, \eta_2)$. For each summary function, there is also a windowed version where the time interval considered is further limited by an upper bound, i.e. $t_{il} \in [t_{ij} - \eta_1, t_{ij} - \eta_3)$ in equation (2). The different functions usually lead to similar prediction performance (personal communication). However, it should be noted that only the frequency based functions are scale invariant, which is one of the properties that make the standard RF algorithm robust. Finally, the partitioning at a particular node is performed using the predictor with the smallest Gini impurity or sum-of-squares error for categorical or quantitative response, respectively. However, determination of an optimal cut-off point for time-varying predictors includes optimization of the parameters in the summary function such as η, which largely increases the computing expenses especially when the number of time-varying predictors is large such as in some omics datasets.

To mitigate the effects of additional optimization of the parameters in the summary function, Sexton and Laake (2018) [29, 30] incorporate an additional level of randomization where instead

Table 2. Overview of different RF extensions from (G)LMM

Outcome	Tree	Forest
Quantitative (Gaussian)	MERT	MERF
	RE-EM tree	REEMforest
	SMERT	SMERF
	SREEM tree	SREEMforest
Exponential family	GMERT	
	GMET	GMERF
Binary	BiMM tree	BiMM forest

of using all observations within the specified time interval, only a sub-sample is randomly selected and used for optimizing the cut-off point. In addition, subject-level resampling strategy is also adopted in RF construction. This not only enjoys the advantages mentioned in Section 2.2.2, but also keeps the complete observation history of a subject.

This method also does not include random effects in the modeling, so the prediction is solely based on the estimated fixed effects.

Extensions from (generalized) linear mixed effects model

A different approach to adjust for the longitudinal structure is to combine (generalized) linear mixed models ((G)LMMs) with the decision tree or RF algorithm. The (G)LMM is a classic statistical methodology for the analysis of longitudinal or more general clustered data. As with (G)LMMs the predictors can be constant or varying over time and different time points are possible for each subject. One advantage of (G)LMM is its explicit modeling of intrasubject correlation structure as well as subject-level random effects besides the main fixed effects of interest. After properly adjusting for these random effects and correlation structure, the longitudinal data become conditionally independent; thus, the estimation of the fixed effect component of the model follows exactly the same way as if independent observations were observed. Moreover, prediction can now be generalized to a wider population. However, the drawbacks of this approach include its computational complexity to fit mixed effects models as well as the possibility to misspecify the intrasubject correlation structure. More detailed descriptions, discussions and applications on classical methods for longitudinal data analysis can be found in several textbooks [4, 5].

The general idea of the RF extension from (G)LMM is to replace the linear model of the fixed effect component by a tree or RF while keeping the modeling of the dependence structure with random effects. Multiple algorithms have been developed to incorporate the tree or RF into the (G)LMM and are summarized in Table 2. As can be seen, most extensions are based on two approaches, namely, MERT and RE-EM trees. For binary response, a Bayesian approach called BiMM has also been proposed.

We first describe approaches for a regression setting based on LMMs, followed by more general methods using GLMMs that can be employed in the context of classification but also for other types of outcomes such as count variables.

Quantitative (Gaussian) response variable

For a normally distributed quantitative outcome, the classic LMM model can be written as

$$y_i = X_i\beta + Z_i b_i + \epsilon_i,$$
$$b_i \sim N(0, D), \epsilon_i \sim N(0, R_i),$$

where $y_i = (y_{i1}, \ldots, y_{in_i})'$ is the $n_i \times 1$ vector of the outcome for the n_i observations of subject i, $X_i = [x_{i1}, \ldots, x_{in_i}]'$ is the $n_i \times p$ matrix of predictors considered as fixed effects, $Z_i = [z_{i1}, \ldots, z_{in_i}]'$ is the $n_i \times q$ matrix of predictors modeled as random effects, $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{in_i})'$ is the $n_i \times 1$ vector of random errors, β is the $p \times 1$ unknown vector of parameters of the fixed effects, and b_i is the $q \times 1$ unknown vector of random effects of subject i. Both b_i and ϵ_i are assumed to follow a normal distribution with mean zero and covariance matrix D and R_i, respectively. It is further assumed that they are independent and that the observations between subjects are also independent. The parameters can be estimated by maximum likelihood (ML) or restricted maximum likelihood (REML) methods.

Two different strategies have been proposed in the literature to replace the linear dependency between the predictors and the outcome. The first approach is the mixed effects regression tree/forest (MERT by Hajjem *et al.* (2011) [31] and MERF by Hajjem *et al.* (2014) [28]) where the fixed effects are estimated using a regression tree or RF. Specifically, the modified model can be written as

$$y_i = f(X_i) + Z_i b_i + \epsilon_i, \quad (3)$$
$$b_i \sim N(0, D), \epsilon_i \sim N(0, R_i),$$

where it is further assumed that $R_i = \sigma^2 I_{n_i}$ where I_n denotes the identity matrix with size n, and the function $f(X_i)$ is estimated by the regression tree or RF. For model fitting an expectation-maximization (EM) algorithm [32] is used, which iterates between estimation of the fixed and random effects components. The general approach can be described as in Algorithm 1 (slightly modified from [31] and [28]).

Algorithm 1 Modified EM algorithm

1: Initialize with $\hat{b}_i = 0$, $\hat{\sigma}^2 = 1$, and $\hat{D} = I_q$
2: **while** not convergent **do**
3: Estimate a regression tree or RF with $y_i^* = y_i - Z_i\hat{b}_i$, and X_i, and denote the predictions with $\hat{f}(X_i)$
4: Fit the linear random effect model $y_i = \hat{f}(X_i) + Z_i b_i + \epsilon_i$
5: **end while**

The convergence is based on a generalized log-likelihood criterion

$$GLL(f, b_i|y) = \sum_{i=1}^{N}\{[y_i - f(X_i) - Z_i b_i]^T R_i^{-1}[y_i - f(X_i) - Z_i b_i] + b_i^T D^{-1} b_i + \log|D| + \log|R_i|\}.$$

This method assumes that the correlation is only due to between subject variation, i.e. the covariance matrix R_i of the errors ϵ_i is assumed to be diagonal. The MERT approach uses a decision tree to estimate $f(X_i)$, whereas the MERF method improves prediction performance by considering a standard RF. It can be noticed that in MERF the bootstrap sample for each tree is drawn on the observation level and predictions are based on the OOB sample to reduce the risk of overfitting. Note that resampling of individual observations is possible in this setting since it is assumed that the correlation between observations can be completely modeled by the random effects. Thus,

using the modified outcome variable y_i^* results in independent observations.

The second approach is independently proposed in Sela and Simonoff (2012) [19], called random effects expectation-maximization (RE-EM) trees. It still considers the model (3), but it does not directly use tree or RF algorithms to estimate the fixed effects. Instead it considers the partition of samples formed by the regression tree and estimates local fixed effects within each partition while estimating the random effects globally. The algorithm is similar to MERT in using the generalized log-likelihood as convergence criterion.

More specifically, for model fitting, the initialization is the same as for MERT, and the rest is modified as shown in Algorithm 2.

Algorithm 2 RE-EM algorithm

1: Initialize with $\hat{b}_i = 0$, $\hat{\sigma}^2 = 1$, and $\hat{D} = I_q$
2: **while** not convergent **do**
3: Estimate a regression tree or RF with $y_i^* = y_i - Z_i\hat{b}_i$, and X_i. Construct indicator matrix Φ^i with size $n_i \times T$ where $\Phi^i_{jt} = I(y_{ij} \in g_t)$, $I(\cdot)$ is the indicator function and g_t is the t-th terminal node of the tree, and T is the total number of terminal nodes
4: Fit the linear mixed effects model

$$y_i = \Phi^i \mu + Z_i b_i + \epsilon_i,$$

where $\mu = (\mu_1, \ldots, \mu_T)$ denotes the local fixed effects within each terminal node
5: **end while**

The tree is thus only used to define the partition of the sample space and a system of LMM models is fitted with global random effects and each partition having its own local fixed effects. The *lme* function in the R package *nlme* is employed for LMM model fitting which allows a general within-subject correlation structure; for instance, an autocorrelation structure within the errors is possible so that R_i can be a nondiagonal matrix.

An extension of the RE-EM tree has recently been proposed by Capitaine *et al.* (2021) [33] and is called REEMforest where an ensemble of RE-EM trees is generated for the fixed effects estimation. The function $\hat{f}(X_i)$ is estimated by the mean of the K fitted RE-EM trees:

$$\hat{f}(X_i) = \frac{1}{K}\sum_{k=1}^{K} \Phi^{i,k}\hat{\mu}_k,$$

where $\Phi^{i,k}$ is the $n_i \times T$ indicator matrix based on the tree k and $\hat{\mu}_k$ is the $T \times 1$ vector of fitted local fixed effects from tree k.

To further consider serial correlation within the observations of the same subject, the MERT and RE-EM tree and their corresponding forest variants have also been extended to include an additional stochastic component in Capitaine *et al.* (2021) [33]. The resulting approaches are correspondingly called SMERT, SREEMtree etc. The model with the additional stochastic component can be written as follows:

$$y_i = f(X_i) + Z_i b_i + \omega_i + \epsilon_i,$$

where $\omega_i = (\omega_i(t_1), \ldots, \omega_i(t_{n_i}))'$ is a centered Gaussian process with $\mathrm{Cov}(\omega_i(s), \omega_i(t)) = \gamma^2\Gamma(s,t)$. The $\omega_i(t)$ are independent for different subjects $i = 1, \ldots, N$ and b_i, ϵ_i and $\omega_i(t)$ are mutually independent. Again, a variant of the EM algorithm is used to estimate the parameters where the definition of the new variable y_i^* now includes the additional stochastic component: $y_i^* = y_i - Z_i\hat{b}_i - \hat{\omega}_i$. In their simulation studies, Capitaine *et al.* (2021) [33] showed that both MERT and RE-EM based tree and RF algorithms are applicable to high-dimensional datasets and provide more accurate prediction than LMM and standard RF.

Generalized response variable

The approaches described so far in this section assume a quantitative outcome that is normally distributed. Further extensions have been proposed for other types of outcomes by using generalized linear mixed models (GLMMs) instead of LMMs.

The GLMM assumes that, conditional on the random effects, the outcome y_i follows a distribution from the exponential family. The GLMM model can be further specified as:

$$g(\mu_i) = \eta_i = X_i\beta + Z_i b_i,$$
$$b_i \sim N(0, D),$$

where $\mu_i = E(y_i|b_i)$, $g(\cdot)$ is a known link function, and η_i is a $n_i \times 1$ vector. Commonly used link functions include identity link, logit link and log link functions for quantitative, binary and count outcomes, respectively. Parameters of GLMMs are estimated by ML or REML methods using numerical optimization algorithms such as penalized quasi-likelihood (PQL) [34], iteratively reweighted least squares or a Newton–Raphson method [35].

Similar to the quantitative outcome case, the RF extensions from GLMM replace the linear relationship between outcome and fixed effects predictor variables by a nonparametric alternative such as a decision tree or RF. The essential estimation procedure is again using an iterative algorithm inspired by the EM algorithm [32] to estimate the fixed and random effects separately and iteratively. Here, we only provide a brief summary of the approaches and mention their quantitative counterparts. For more details, we refer the reader to the original publications.

The MERT approach has been extended to the generalized mixed effects regression tree (GMERT) in Hajjem *et al.* (2017) [36]. The PQL algorithm of the GLMM is modified so that a weighted MERT pseudo-model is used instead of the weighted linear mixed-effects pseudo-model. The fixed part is again estimated with CART. In this implementation it is necessary to specify initial estimates of the mean values $\hat{\mu}_i$. In the simulation study with a binary outcome the authors used predetermined values $\hat{\mu}_{ij} = 0.25$ if $y_{ij} = 0$ and $\hat{\mu}_{ij} = 0.75$ if $y_{ij} = 1$. Unfortunately, no further discussions on this initialization were presented.

Similarly, Fontana *et al.* (2018) [37] extends the RE-EM tree to the generalized mixed effects tree (GMET). Again, a regression tree is used and the indicator variables for the terminal nodes are modeled as fixed effects in the mixed effects model. The modified outcome variable for the regression tree is $y_i^* = \eta_i - Z_i b_i$. However, η_i needs to be estimated that is usually achieved with a standard generalized linear model (GLM) using the predictors as fixed effects covariates (in [37, 38]). Note that this approach is not possible for high-dimensional data due to this need to estimate η_i with GLM since the number of variables then cannot exceed the number of observations. To the best of our knowledge, no solutions for high-dimensional data have so far been proposed in the literature along this direction. GMET has further been extended to generalized mixed effects random forest (GMERF) in Pellagatti *et al.* (2021) [38] where instead of growing only a single decision tree, an RF is trained.

A Bayesian approach

For binary outcomes, Binary Mixed Model (BiMM) tree proposed in Speiser *et al.* (2020) [39] considers a Bayesian implementation of GLMM. The GLMM portion of the BiMM method has the form

$$\text{logit}(\mu_{it}) = \text{CART}(X_{it})\beta + Z_{it}b_{it},$$

where $\text{CART}(X_{it})$ are indicator variables reflecting membership of each longitudinal observation t for subject i in terminal nodes within the decision tree. Therefore, the use of the tree in this approach is again not to model the fixed effects directly, but rather to determine similar groups of observations after random effects have been properly adjusted.

For estimation the BiMM tree method again adopts the EM-like algorithm and iterates between developing CART models using all predictors and then using information from the CART model within a Bayesian GLMM to adjust for the clustered structure of the outcome.

This tree method is further extended to a forest-based method by Speiser *et al.* (2019) [40] where all $\text{CART}(X_{it})$ are replaced by $\text{RF}(X_{it})$. More details of the algorithms can be found in the corresponding paper.

Lin and Luo (2019) [41] considers the same model for binary outcomes and proposed a multilevel CART (M-CART) for estimation. The only difference between their method and BiMM is that they fit a multilevel logistic model instead of a Bayesian GLMM in the EM-like iterations to adjust for the clustered structure of the outcome.

Compared with the M-CART and previously reviewed frequentist methods, the Bayesian approach, as pointed out by the authors, can avoid issues with model convergence, especially when data are high dimensional. In addition, when uninformative priors are used, frequentist GLMM results can be obtained. That is to say, the Bayesian approach provides a more general framework with frequentist approaches such as RE-EM tree/forest as special cases.

Prediction with RF extensions from (G)LMM model

When making predictions with aforementioned RF extensions from a (G)LMM model, two different settings have to be distinguished. The first case is prediction for a new subject i for which no random effects $\hat{b}_i$ are available. This happens when the subject does not belong to the training dataset. Thus, prediction is solely based on the fixed effect component, which is either given by the prediction of the tree or RF or the predicted effect associated with the terminal node in which the new observation lands ($\hat{\Phi}^{i}_{jt}\hat{\mu}_t$). The other case is to predict a new observation for a subject i used in the training process. In this case, the estimated random effects are available, so the sum of the fixed component and the corresponding random effect of subject i can be used together.

Multivariate response longitudinal data

So far, the reviewed methods are designed for univariate response longitudinal data; however, we can also treat measurements at different time points together as multivariate responses or a discretized response curve. Therefore, in this section, we would like to shift our attention to extensions of the RF algorithm that can accommodate multivariate response variables.

Before we start, we would like to note that the algorithms in this section are directly applicable with time-invariant predictors such as genetic data. If predictors are also observed at multiple time points, techniques from the previous sections need to be used along with the modifications reviewed in this section for an adequate analysis.

Repeatedly measured univariate longitudinal responses as multivariate response

Even though the within-subject correlation cannot be disregarded, at the subject level, the usual independency assumption is still reasonable. Therefore, one strategy for longitudinal data analysis is to consider observations at different time points jointly so that each subject has only one multi-dimensional response.

To accommodate multivariate responses, the common strategy to extend the RF algorithm largely focuses on modifying the split criterion in the construction of each decision tree, where impurity measures are modified so that multivariate responses can be handled properly. In addition the covariance structure needs to be considered when defining the impurity measure to account for the inter-dimensional correlation. The modifications can be roughly categorized into two classes, using either distance or likelihood based split criteria.

Distance-based impurity functions

Segal (1992) [42] was among the first to extend CART to longitudinal data by using a distance based measure for node impurity. Specifically, the author considered a univariate quantitative response but treated measurements at different time points jointly as a multivariate response. It is further assumed that the observation times for all subjects are the same, so that the dimension of the multivariate response is fixed and not changing across subjects. For a given node t, Segal (1992) [42] considers the following generalized sum of square function:

$$SS(t) = \sum_{i\in t}(y_i - \bar{y}_t)'\mathbf{V}(\theta, t)^{-1}(y_i - \bar{y}_t), \tag{4}$$

where $y_i,\ i = 1\ldots, N$ is a $n \times 1$ vector, $\bar{y}_t$ is the sample average of y_i's within node t, $\mathbf{V}(\theta, t)$ denotes the $n \times n$ covariance matrix of the responses within node t and depends on unknown parameters θ, which can be estimated within the node. In principle, the estimated parameters $\hat{\theta}$ can differ for node t and its daughters t_L and t_R, which as the author noticed may lead to impurity increase. Hence, the author further imposes the restriction that for each candidate split the covariance parameters are determined from the parent node t so that

$$\mathbf{V}(\theta, t) = \mathbf{V}(\theta_L, t_L) = \mathbf{V}(\theta_R, t_R).$$

Furthermore, the author provides several candidates for the covariance structure, namely, independence (i.e. diagonal matrix), 1st-order autoregression (AR1), compound symmetry (CS) and sample covariance matrix.

The independence structure leads to the sum of square about the mean:

$$SS(t) = \sum_{i\in t}\sum_{j=1}^{n}(y_{ij} - \bar{y}_j)^2,$$

where y_{ij} is the outcome for subject i and component j, and all subjects are assumed to have same number of n components. This is a direct generalization from the univariate regression tree, and has been used by De'Ath (2002) [43] for applications in ecology

and by Segal and Xiao (2011) [44] in the construction of the multivariate RF.

When the sample covariance matrix is adopted in Segal's approach, the generalized sum of square function in equation (4) is closely related to the Mahalanobis distance where the Mahalanobis distance of an observation y_i from a set of observations with mean μ_i and (nonsingular) covariance matrix $\mathbf{S}$ is defined as

$$D(y_i) = \sqrt{(y_i - \mu_i)'\mathbf{S}^{-1}(y_i - \mu_i)}.$$

Larsen and Speckman (2004) [45] directly considered the Mahalanobis distance as node impurity measurement and split criterion. Instead of updating the covariance structure during the tree construction, they estimate the covariance matrix from the whole data set at the very beginning and use the estimate throughout the whole process. They still consider the simple average of observations in each node for μ_i, but different estimators such as trimmed mean could also be adopted.

Besides the Mahalanobis distance based split criterion, De'Ath (2002) [43] proposed the distance-based multivariate regression tree (db-MRT), where the impurity of a given node is measured based on the pairwise dissimilarities between observations within the node. Sim *et al.* (2013) [46] put this approach into a more formal construction where the dissimilarities between observations are captured by a distance matrix $D = \{D_{ij}\}_{1\le i,j\le N}$ with D_{ij} measuring the distances between y_i and y_j. Then, the impurity of each node t is defined as

$$Imp(t) = \sum_{i,j\in t} D_{ij}^2.$$

This approach is more general than the aforementioned extensions in that the distance matrix does not necessarily depend on the dimension of the original responses. In fact, it is possible to analyze longitudinal responses at irregular time points with this approach as long as an appropriate distance measure can be defined. However, how to make prediction with the resulting RF needs further consideration because now within a leaf, it is possible to have responses with different dimensions thus, usual sample average would not make sense in such cases.

Lastly, when the distance is measured by l_1-norm, given the well-known relationship that

$$\sum_{i=1}^{N} |y_i - \bar{y}|^2 = \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N} |y_i - y_j|^2,$$

this distance matrix based approach is connected to Segal's approach with an assumed independence covariance structure.

Likelihood based impurity function

Zhang (1998) [47] extended CART to multiple binary response variables. For responses from an exponential family distribution, the author considered the log-likelihood as the node impurity that depends only on the linear terms and the sum of the 2nd-order products of the responses. Specifically, for subject i, y_i is assumed to follow the joint probability distribution:

$$f(y_i; \Psi, \theta) = \exp(\Psi' y_i + \theta w_i - A(\Psi, \theta)),$$

where Ψ and θ are arrays of parameters, $A(\Psi, \theta)$ is the normalization function depending only on Ψ and θ, and $w_i = \sum_{j<k} y_{ij}y_{ik}$. The node impurity is defined as the maximum of the log-likelihood derived from this distribution; that is, for node t,

$$Imp(t) = \sum_{i\in t}(\hat{\Psi}' y_i + \hat{\theta} w_i - A(\hat{\Psi}, \hat{\theta})),$$

where $\hat{\Psi}$ and $\hat{\theta}$ are the maximum likelihood estimates of Ψ and θ within the node. Zhang and Ye (2008) [48] applied the same technique to ordinal responses by first transforming them to binary-valued indicator functions.

When multivariate normally distributed responses are considered, Abdolell *et al.* (2002) [49] proposed a likelihood-ratio test statistic as impurity function. Specifically, suppose that $y_i \sim N_n(\mu_i, \Sigma)$, the authors define the deviance function for a single observation as

$$D(\mu_i; y_i) = 2[\ell(y_i; y_i) - \ell(\mu_i; y_i)] = (y_i - \mu_i)'\Sigma^{-1}(y_i - \mu_i),$$

where $\ell(\mu_i; y_i)$ is the log-likelihood function. Assuming that Σ is constant and given for all i, they further define the deviance within a node t as

$$D(\hat{\mu}; y, t) = \sum_{i\in t} D(\hat{\mu}; y_i),$$

where $\hat{\mu}$ is the restricted maximum likelihood estimate within the node. The impurity of a node is then measured by the negation of the deviance. As pointed out by the authors, this deviance function in the context of the multivariate normal distribution is the Mahalanobis distance between μ_i and y_i. In addition, they also noticed that deviance assessed via the multivariate analysis of variance approach such as Hotelling's T^2 is again in a form of the Mahalanobis distance. These observations connects the likelihood-ratio test statistics-based impurity function with the aforementioned Mahalanobis distance based one.

The likelihood-ratio test statistics based impurity function is also considered by Segal (1992) [42] for multivariate normally distributed responses. However, their splitting rule focuses on the intrasubject variation structure other than the mean structure of responses, which we think may be difficult to interpret and less of interest in terms of precision medicine.

Prediction with multivariate RF extensions

While the RF extensions for univariate response longitudinal data predict the outcome at a single time point, the multivariate models predict the outcome at each of the predefined time points simultaneously.

Implementation

In Table 3, we provide a summary of the software implementations of the RF extensions reviewed in previous sections. Briefly, almost all extensions are presented in an R package on CRAN or as an R program. The majority of RF extensions provides variable importance measures that rely on standard importance measures of the underlying basic RF implementations. For instance, R package `LongitudiRF` uses `randomForest` [50] for RF construction, which provides both permutation and Gini importance measures. RF++ and Historical RF only give permutation importance measure and in the R package `MultivariateRandomForest`, the variable importance is based on the frequency of the variable being used as a splitting variable.

Table 3. Overview of the implementations of the reviewed RF extensions

Name	Implementation	Type	Response	VImp	References
RF++	Stand-alone software (binary) (https://sourceforge.net/projects/rfpp)	F	C, R	Yes	[24]
Historical RF	R package htree (CRAN)	F	C, R	Yes	[30]
MERT	R package LongituRF (CRAN)	T	R	No	[31]
MERF	R package LongituRF (CRAN)	F	R	Yes	[28]
RE-EM	R package REEMtree (CRAN)	T	R	No	[19, 51]
SMERT, SMERF (S)REEMforest	R package LongituRF (CRAN)	T, F	R	Yes	[33, 52]
GMERT	R code (supplement to original paper)	T	C	No	[36]
BiMM forest	R code (supplement to original paper)	F	C	Yes	[39, 40]
Multivariate RF	R package MultivariateRandomForest (CRAN)	F	R	Yes	[44, 53, 54]
	R package mvpart (CRAN)	T	R	No	[43, 55]
	R package randomForestSRC (CRAN)	F	C, R	Yes	[44, 56]

VImp = Variable importance, T = tree, F = forest, R = regression, C = classification

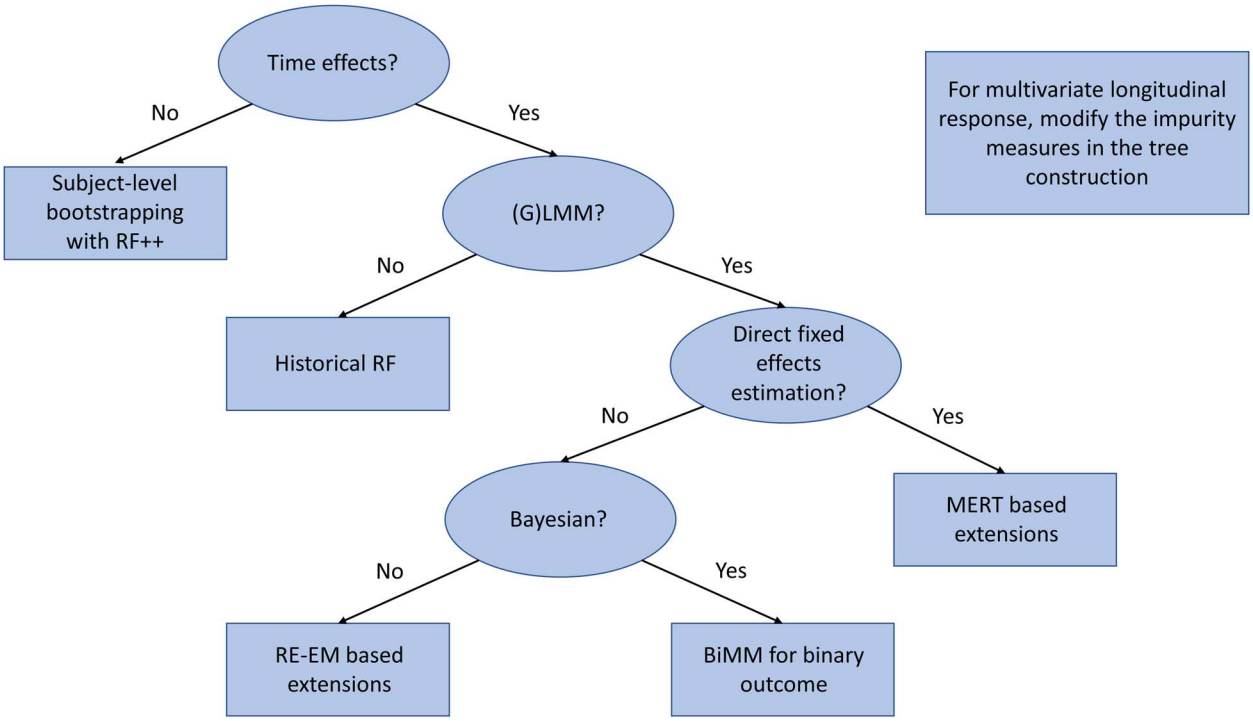

Figure 3. Summary of the concepts and methods reviewed in the paper.

Discussion

In this paper, we review extensions of the CART-based RF algorithm for the analysis of longitudinal data. An overview of the concepts and methods reviewed is shown in Figure 3.

There are certainly other RF construction approaches that can be considered. Examples include GUIDE [57] and conditional inference [58] approaches. For longitudinal data, extensions of non-CART-based tree and RF have also been proposed such as repeated measures random forest [59], mixed effect machine learning framework [60] and partially additive linear model trees [61]. See [62] for a review on different tree construction approaches and related discussions on extensions for longitudinal data analysis. Some extensions reviewed here such as MERT and MERF can easily switch to other approaches for tree/forest construction. In fact, the EM-based mixed effects approaches can also be adapted to be used with not only tree-based methods, but also other ML algorithms such as support vector machines or neural networks. However, one advantage of RF algorithm is that the prediction error will not increase when the number of trees in the forest increases, reflecting that it is not very prone to overfitting [10]. But it is unknown whether such property holds as well for the reviewed extensions.

As we pointed out in Section 1, missing values represent a major challenge. Segal (1992) [42] considered the surrogate splitting variable approach, which is one of the standard solutions for missing values in the literature of tree and forest methods. However, different missing mechanisms may require different approaches to handle missing values. In general, this is still an important research area in the context of developing and applying statistical methods and ML approaches in general.

Variable importance measure is a unique feature that RF can offer to support variable selection. Some reviewed extensions consider permutation-based variable importance, which is easy to implement but computationally expensive. Other approaches

such as the variable-delete approach may also considered, but correlation between predictors may negatively affect its performance. How to measure variable importance in a tailored fashion for longitudinal data warrants further study, because this could be beneficial in both understanding disease progression and searching for target biomarkers for drug design. In addition, as pointed out by Speiser *et al.* (2019) [40], it is also interesting to investigate within-cluster variable importance measures.

Another direction for future methodology development is on the effective handling of high dimensional longitudinal data. Except for the BiMM and REEMforest methods, the other reviewed methods have so far not been evaluated on high-dimensional data sets. For instance, for the extensions from GLMM with nonquantitative outcomes, there is a need for an initial GLM fit, which would be very difficult, if not impossible, with the high-dimensional data sets. Having RF extensions being able to handle such data would provide fruitful insights on their effects on complex diseases.

Lastly, while many of the mentioned studies compare the proposed modifications with standard RF or statistical approaches, only a few of them include comparisons between different RF extensions. One example is the simulation study performed by Capitaine *et al.* (2021) [33], which demonstrates that MERT and RE-EM-based RFs have similar prediction performance. However, comprehensive neutral comparison studies [63] for specific research questions and different types of longitudinal data with realistic simulation studies are needed to systematically investigate prediction performance, variable selection and computational efficiency. We plan to conduct such a benchmarking study to further understand these RF extensions and to give recommendations to practitioners analyzing real data sets.

Key Points

- Extensions of standard random forest algorithm for longitudinal data have been comprehensively reviewed.
- For clustered data where there is no time effect, subject-level bootstrap sampling technique is the key to take care of the clustering effect, and when the time effect is relevant, extensions from (generalized) linear mixed effects model and historical RF can be considered.
- Extensions to handle multivariate response are also available.
- Only a limited number of methods can analyze high-dimensional data such as omics data.
- Variable importance measure, a unique and essential feature of standard random forest for variable selection, warrants further development in the context of longitudinal data analysis.

Author contributions statement

J.H. and S.S. wrote and reviewed the manuscript.

Funding

This work was supported by the German Federal Ministry of Education and Research (BMBF) funded e:Med Programme on Systems Medicine [grant 01ZX1510 (ComorbSysMed) to S.S.].

References

1. Ashley EA. Towards precision medicine. *Nat Rev Genet* 2016; **17**(9): 507–22.
2. Larry Jameson J, Longo DL. Precision medicine-personalized, problematic, and promising. *Obstet Gynecol Surv* 2015; **70**(10): 612–4.
3. Matchett KB, Niamh Lynam-Lennon R, Watson W, *et al.*Advances in precision medicine: tailoring individualized therapies. *Cancer* 2017; **9**(11): 146.
4. Fitzmaurice GM, Laird NM, Ware JH. *Applied Longitudinal Analysis*. John Wiley & Sons, Hoboken, New Jersey, 2012.
5. Hedeker D, Gibbons RD. *Longitudinal Data Analysis*. Wiley-Interscience, Hoboken, New Jersey, 2006.
6. Krasniqi E, Schramm W, Reichenbach A. Data-driven stratification of parkinson's disease patients based on the progression of motor and cognitive disease markers datengetriebene stratifizierung von patienten mit parkinson-krankheit anhand von verlaufsdaten motorischer und kognitiver kennzahlen der erkrankung. *GMS Medizinische Informatik, Biometrie und Epidemiologie* 2021; **17**(1) ISSN:1860–9171.
7. Latourelle JC, Beste MT, Hadzi TC, *et al.*Large-scale identification of clinical and genetic predictors of motor progression in patients with newly diagnosed parkinson's disease: a longitudinal cohort study and validation. *Lancet Neurol* 2017; **16**(11): 908–16.
8. Zhang X, Chou J, Liang J, *et al.*Data-driven subtyping of parkinson's disease using longitudinal clinical records: a cohort study. *Sci Rep* 2019; **9**(1): 1–12.
9. König IR, Fuchs O, Hansen G, *et al.*What is precision medicine? *Eur Respir J* 2017; **50**(4).
10. Breiman L. Random forests. *Mach Learn* 2001; **45**(1): 5–32.
11. Ishwaran H, Kogalur UB, Blackstone EH, *et al.*Random survival forests. *Ann Appl Stat* 2008; **2**(3): 841–60.
12. Chen X, Ishwaran H. Random forests for genomic data analysis. *Genomics* 2012; **99**(6): 323–9.
13. Richard Cutler D, Edwards TC, Beard KH, *et al.*Random forests for classification in ecology. *Ecology* 2007. Preprint; **88**(11): 2783–92. https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1890/07-0539.1.
14. Mooney SD. Progress towards the integration of pharmacogenomics in practice. *Hum Genet* 2015; **134**(5): 459–65.
15. Ritchie MD. The success of pharmacogenomics in moving genetic association studies from bench to bedside: study design and implementation of precision medicine in the post-gwas era. *Hum Genet* 2012; **131**(10): 1615–26.
16. Svetnik V, Liaw A, Tong C, *et al.*Application of breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules. In: *International Workshop on Multiple Classifier Systems*. Springer-Verlag, Berlin Heidelberg, 2004, 334–43.
17. Raudenbush SW, Bryk AS. *Hierarchical Linear Models: Applications and Data Analysis Methods*, Vol. **1**. Sage, Thousand Oaks, California, 2002.
18. Fokkema M, Smits N, Zeileis A, *et al.*Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behav Res Methods* 2018; **50**(5): 2016–34.
19. Sela RJ, Simonoff JS. RE-EM trees: a data mining approach for longitudinal and clustered data. *Mach Learn* 2012; **86**(2): 169–207.
20. Mangino AA, Holmes Finch W. Prediction with mixed effects models: a Monte Carlo simulation study. *Educ Psychol Meas* 2021; **81**(6): 1118–42.

21. Breiman L, Friedman J, Stone CJ, *et al. Classification and Regression Trees*. CRC Press, Boca Raton, FL, 1984.
22. Zhang H, Singer BH. *Recursive Partitioning and Applications*. Springer Science & Business Media, New York, 2010.
23. Degenhardt F, Seifert S, Szymczak S. Evaluation of variable selection methods for random forests and omics data sets. *Brief Bioinform* 2019; **20**(2): 492–503.
24. Karpievitch YV, Hill EG, Leclerc AP, *et al.*An introspective comparison of random forest-based classifiers for the analysis of cluster-correlated data by way of RF++. *PLoS One* 2009; **4**(9):e7087.
25. Vlahou A, Giannopoulos A, Gregory BW, *et al.*Protein profiling in urine for the diagnosis of bladder cancer. *Clin Chem* 2004; **50**(8): 1438–41.
26. Adler W, Brenning A, Potapov S, *et al.*Ensemble classification of paired data. *Comput Stat Data Analysis* 2011; **55**(5): 1933–41.
27. Adler W, Potapov S, Lausen B. Classification of repeated measurements data using tree-based ensemble methods. *Comput Stat* 2011; **26**(2): 355.
28. Hajjem A, Bellavance F, Larocque D. Mixed-effects random forest for clustered data. *J Stat Comput Simulation* 2014; **84**(6): 1313–28.
29. Sexton J, Laake P. *Historical random forests* Working paper, 2018.
30. Sexton J. *htree: historical tree ensembles for longitudinal data*. R package version 2.0.0, 2018.
31. Hajjem A, Bellavance F, Larocque D. Mixed effects regression trees for clustered data. *Stat Probability Lett* 2011; **81**(4): 451–9.
32. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982;963–74.
33. Capitaine L, Genuer R, Thiébaut R. Random forests for high-dimensional longitudinal data. *Stat Methods Med Res* 2021; **30**(1): 166–84 PMID: 32772626.
34. Rodríguez G. Multilevel generalized linear models. In: *Handbook of Multilevel Analysis*. Springer, New York, 2008, 335–76.
35. McCullagh P, Nelder JA. *Generalized Linear Models*. CRC Press, Boca Raton, FL, 2019.
36. Hajjem A, Larocque D, Bellavance F. Generalized mixed effects regression trees. *Stat Probability Lett* 2017; **126**:114–8.
37. Fontana L, Masci C, Ieva F, *et al.*Performing learning analytics via generalized mixed-effects trees. *MOX-Modelling and Scientific Computing, Department of Mathematics, Politecnico di Milano, via Bonardi* 2018; **9**:1–17.
38. Pellagatti M, Masci C, Ieva F, *et al.*Generalized mixed-effects random forest: a flexible approach to predict university student dropout. *Statistical analysis and data mining: the ASA. Data Sci J* 2021. _eprint:; **14**(3): 241–57. https://onlinelibrary.wiley.com/doi/pdf/10.1002/sam.11505.
39. Speiser JL, Wolf BJ, Chung D, *et al.*BiMM tree: a decision tree method for modeling clustered and longitudinal binary outcomes. *Commun Stat Simul Comput* 2020; **49**(4): 1004–23.
40. Speiser JL, Wolf BJ, Chung D, *et al.*BiMM forest: a random forest method for modeling clustered and longitudinal binary outcomes. *Chemom Intel Lab Syst* 2019; **185**:122–34.
41. Lin S, Luo W. A new multilevel cart algorithm for multilevel data with binary outcomes. *Multivar Behav Res* 2019; **54**(4): 578–92.
42. Segal MR. Tree-structured methods for longitudinal data. *J Am Stat Assoc* 1992; **87**(418): 407–18.
43. Glenn De'ath.Multivariate regression trees: a new technique for modeling species-environment relationships. *Ecology* 2002; **83**(4): 1105–17.
44. Segal M, Xiao Y. Multivariate random forests. *Wiley Interdisciplinary Rev* 2011; **1**(1): 80–7.
45. Larsen DR, Speckman PL. Multivariate regression trees for analysis of abundance data. *Biometrics* 2004; **60**(2): 543–9.
46. Sim A, Tsagkrasoulis D, Montana G. Random forests on distance matrices for imaging genetics studies. *Stat Appl Genet Mol Biol* 2013; **12**(6): 757–86.
47. Zhang H. Classification trees for multiple binary responses. *J Am Stat Assoc* 1998; **93**(441): 180–93.
48. Zhang H, Ye Y. A tree-based method for modeling a multivariate ordinal response. *Statistics Interface* 2008; **1**(1): 169.
49. Abdolell M, LeBlanc M, Stephens D, *et al.*Binary partitioning for continuous longitudinal data: categorizing a prognostic variable. *Stat Med* 2002; **21**(22): 3395–409.
50. Liaw A, Wiener M. Classification and regression by randomforest. *R News* 2002; **2**(3): 18–22.
51. Sela JSR, Jing W. *REEMtree: regression trees with random effects for longitudinal (panel) data*. R package version 0.90.4, 2021.
52. Capitaine L. *LongituRF: random forests for longitudinal data*. R package version 0.9, 2020.
53. Rahman R. *MultivariateRandomForest: models multivariate cases using random forests*. R package version 1.1.5, 2017.
54. Rahman R, Otridge J, Pal R. IntegratedMRF: random forest-based framework for integrating prediction from different data types. *Bioinformatics* 2017; **33**(9): 1407–10.
55. De'ath G. *mvpart: multivariate partitioning*. R package version 1.6–2, 2014.
56. Kogalur Hemant Ishwaran UB. *randomForestSRC: fast unified random forests for survival, regression, and classification (RF-SRC)*. R package version 3.1.0, 2022.
57. Loh W-Y. Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica* 2002; **12**(2): 361–86. Institute of Statistical Science, Academia Sinica.
58. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: a conditional inference framework. *J Comput Graph Stat* 2006; **15**(3): 651–74.
59. Calhoun P, Levine RA, Fan J. Repeated measures random forests (rmrf): identifying factors associated with nocturnal hypoglycemia. *Biometrics* 2021; **77**(1): 343–51.
60. Ngufor C, Van Houten H, Caffo BS, *et al.*Mixed effect machine learning: a framework for predicting longitudinal change in hemoglobin a1c. *J Biomed Inform* 2019; **89**:56–67.
61. Seibold H, Hothorn T, Zeileis A. Generalised linear model trees with global additive effects. *Adv Data Anal Classification* 2019; **13**(3): 703–25.
62. Loh W-Y. Fifty years of classification and regression trees. *Int Stat Rev* 2014; **82**(3): 329–48.
63. Boulesteix A-L, Lauer S, Eugster MJA. A plea for neutral comparison studies in computational sciences. *PLoS One* 2013; **8**(4):e61562.