Received: 14 July 2018 | Accepted: 7 April 2020

DOI: 10.1111/biom.13284



Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

BIOMETRIC PRACTICE

# Repeated measures random forests (RMRF): Identifying factors associated with nocturnal hypoglycemia

**Peter Calhoun**[1] | **Richard A. Levine**[2] | **Juanjuan Fan**[3]

[1]Jaeb Center for Health Research, Tampa, Florida

[2]Department of Mathematics and Statistics, Analytic Studies and Institutional Research, San Diego State University, San Diego, California

[3]Department of Mathematics and Statistics, San Diego State University, San Diego, California

**Correspondence**
Juanjuan Fan, PhD, Department of Mathematics and Statistics, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182.
Email: jjfan@sdsu.edu

**Funding information**
NSF, Grant/Award Number: 1633130

**Abstract**

Nocturnal hypoglycemia is a common phenomenon among patients with diabetes and can lead to a broad range of adverse events and complications. Identifying factors associated with hypoglycemia can improve glucose control and patient care. We propose a repeated measures random forest (RMRF) algorithm that can handle nonlinear relationships and interactions and the correlated responses from patients evaluated over several nights. Simulation results show that our proposed algorithm captures the informative variable more often than naïvely assuming independence. RMRF also outperforms standard random forest and extremely randomized trees algorithms. We demonstrate scenarios where RMRF attains greater prediction accuracy than generalized linear models. We apply the RMRF algorithm to analyze a diabetes study with 2524 nights from 127 patients with type 1 diabetes. We find that nocturnal hypoglycemia is associated with HbA1c, bedtime blood glucose (BG), insulin on board, time system activated, exercise intensity, and daytime hypoglycemia. The RMRF can accurately classify nights at high risk of nocturnal hypoglycemia.

**KEYWORDS**

hypoglycemia, longitudinal data, random forest, repeated measures, type 1 diabetes, variable importance

## 1 | INTRODUCTION

Medical research heavily relies on longitudinal designs to identify factors associated with a disease or outcome. The standard methodology to test factors associated with an outcome often uses a repeated measures linear or logistic regression analysis and corresponding variable selection techniques. However, model misspecification or invalid assumptions can yield incorrect or misleading conclusions. Nonlinear relationships, outliers, or missing measurements require more sophisticated approaches while still handling the dependence resulting from repeated measures. To overcome these issues, we propose a repeated measures random forest (RMRF) algorithm.

Random forests (RFs) (Breiman, 2001) require few statistical assumptions and often outperform classical methods such as logistic regression or linear regression in terms of prediction accuracy. They are particularly effective at handling nonlinear relationships, interactions, outliers, and missing measurements. Additionally, they can handle data with a large number of predictors of mixed type (continuous, ordinal, or categorical with two or more levels) without the need for predictor variable transformation, dummy variable creation, or variable selection. We extend the RF algorithm to handle repeated measures and apply our proposed method in a diabetes study to identify factors associated with nocturnal hypoglycemia.

Segal (1992) first extended regression trees to handle longitudinal data by replacing the node impurity in classification and regression trees (CART) with a generalized least square split function (referred as Mahalanobis distance). Loh and Zheng (2013) follow a GUIDE approach by treating each longitudinal data series as a curve and using chi-squared tests of the residual curve patterns to select a variable to split.

  

Once a variable is selected, the best split is determined by minimizing the total sum of squared deviations of the two subnodes. Hajjem *et al.* (2011) propose splitting on transformed response data after removing the random-effect component from the original response. Sela and Simonoff (2012) independently propose a similar approach called random-effects/expectation-maximization (EM) trees where the splitting is based on maximizing the reduction in sum of squares in a node. Other authors have implemented linear mixed-effects tree algorithms; however, all of these approaches focus on regression trees for their simulations and analysis. For this paper, we are examining RFs for classification problems.

Zhang (1998) proposes a classification tree method with multiple responses using a generalized entropy index. Lee (2005) uses generalized estimating equations (GEE) to account for the correlation and demonstrates superior performance to Segal's (1992) Mahalanobis tree algorithm and Zhang's (1998) generalized entropy index method. Lee's approach involves fitting a marginal model using GEE techniques and using the residuals to determine the best split. However, Loh and Zheng (2013) find that partitioning the data by the signs of their average residuals is potentially ineffective. Additionally, Lee's method works only for multivariate outcomes where predictors have a fixed value within each cluster. Hajjem *et al.* (2017) propose using a penalized quasi-likelihood method and using an EM algorithm for computation. They show better performance over the standard classification tree when random effects are present, but this has not been extended to RFs and can be computationally intensive. Sharpsten *et al.* (2013) use a robust Wald statistic, derived from GEE, to determine the best split. The robust Wald statistic can handle all types of predictors. Furthermore, the authors show the robust Wald statistic outperforms Lee's (2005) approach at identifying the informative variable. These papers investigate a single tree when using either stopping rules or a CART-like approach. Only Segal and Xiao's (2011) and Hajjem *et al.*'s (2014) subsequent papers extend the research to multivariate RFs for regression problems. The goal of this paper is to develop an RF algorithm that can split on all types of predictors and handle repeated measurements. Our motivation to extend RF to handle longitudinal data is to identify factors associated with nocturnal hypoglycemia. We adopt the robust Wald statistic as the splitting statistic and use an acceptance-rejection criterion to reduce the computational intensity and variable selection bias. We show that the standard RF often misses informative variables with repeated measurements and demonstrate superior performance with the proposed RMRF algorithm. To our knowledge, this is the first paper to extend RFs to longitudinal data with a binary outcome using a marginal model approach.

For simplicity, throughout this paper we let patients represent the clusters and nights as the repeated measurements in each cluster. In Section 2, we describe the proposed RMRF algorithm. In Section 3, we show our method correctly captures the true underlying data structure and outperforms standard RF implementations that ignore the dependence. We apply the RMRF algorithm to a diabetes study in Section 4. We conclude with the strengths of using the RMRF and discuss possible future research. The RMRF algorithm and all results in this paper can be found on GitHub with the link provided in the Web Appendix Table 1.

## 2 | METHOD

### 2.1 | Splitting criteria

We adopt the robust Wald statistic proposed by Sharpsten *et al.* (2013) as the splitting criteria. Suppose there are $N$ patients with the $i$th subject containing $n_i$ correlated nights. Let $Y_{ij}$ represent the binary response (eg, whether hypoglycemia occurred) for the $i$th subject on the $j$th night, and let $\boldsymbol{X}_{ij}$ represent a $p$-dimensional vector of covariates, $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$. To determine the splitting statistic for the $k$th variable ($k = 1, \ldots, p$), we fit the logit marginal model

$$\text{logit}\left(p_{ij}\right) = \beta_0 + \beta_1 I\left(x_{ijk} \leq c\right) \tag{1}$$

for $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$, where $p_{ij} = P(Y_{ij} = 1)$, $\text{logit}(p_{ij}) = \log\{p_i/(1 - p_{ij})\}$, $\beta_0$ and $\beta_1$ are unknown regression parameters, and $I(\cdot)$ is the indicator function.

The marginal model fits a single binary split using the $k$th continuous covariate at cutpoint $c$. If the $k$th covariate is categorical, the form of the binary split is determined by $I(X_{ijk} \in C)$, where $C$ can be any subset of categories. While linear combinations of covariates may be considered, we restrict to binary splits on a single covariate. The robust Wald statistic is calculated for testing $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ from the marginal model. Details of the robust Wald statistic formula are given in Web Appendix A. The splitting statistic is robust in the sense that the results are consistent even when the correlation structure is misspecified.

We use the *geeM* R package (McDaniel and Henderson, 2016) for fitting GEE models and calculating the robust Wald test statistic. The possible working correlation structures that can be implemented include independent, $m$-dependent, exchangeable, first-order autoregressive, unstructured, fixed, and user defined. We utilize the *partykit* (Hothorn and Zeileis, 2015) R package for growing the trees and RFs. While there are algorithms that can implement RFs for independent (non-clustered) data, we developed our own implementation to use the same growing parameters as the proposed RMRF algorithm. In Section 2.2, we describe how the splitting statistic can be used to grow RFs for longitudinal studies.

**TABLE 1** Repeated measures random forest algorithm

For $t = 1, 2, \ldots, ntree$, do

1. Generate subsample $L_s$ by sampling without replacement the patient ids and taking all data from the patients randomly selected.
2. Pick a random variable in $L_s$.
3. Pick a random cutpoint from the chosen variable.
4. Calculate robust Wald statistic and its respective $P$-value.
5. If $P$-value <`minpvalue` threshold, then take split. Otherwise, repeat steps 2 to 5.
6. Once a split is chosen, repeat steps 2 to 6 for data in each child node until a stopping criterion is reached.

## 2.2 | Repeated measures random forest algorithm

The first step in growing a tree involves subsampling the original data. The use of subsamples differs from the standard RF (Breiman, 2001) and Segal and Xiao's (2011) and Hajjem *et al.*'s (2014) multivariate RF algorithms that use bootstrap samples. Bootstrap samples would contain multiple responses at the same time point and the tree-growing procedure would need to account for both the subject and time dependence. Subsampling works under weaker assumptions for statistical inference compared with bootstrapping and subsample aggregating has been found to be computationally cheaper with approximately the same accuracy as bagging (Buhlmann and Yu, 2002). We sample 63.2% of patients to grow an unpruned tree; this is the approximate percentage of unique observations in bootstrap samples. We note that this step differs from the traditional subsampling technique in that patients (instead of nights) are sampled without replacement. Thus, the out-of-bag (OOB) observations include different subjects and are independent of the in-bag sample.

Table 1 summarizes the RMRF algorithm. The tree is grown by recursively partitioning the subsampled data using the robust Wald statistic until a stopping criterion is reached. Specifically, one variable is randomly selected and then one cutpoint is randomly chosen for that variable at each internal node. The $P$-value of the robust Wald statistic is calculated and used to assess the random split. If the random split does not attain a small enough $P$-value, the process is repeated where a variable and then a cutpoint are randomly selected. While the optimal $P$-value threshold (denoted `minpvalue`) can be selected for a particular dataset via cross-validation, our implementation requires a split must have a $P$-value <.16—the same threshold used in Hallett's (2015) dissertation with survival data. We find that variable importance and prediction accuracy results are not sensitive to choices of different $P$-value thresholds in the range of .01 to .25. The node stops partitioning when there are fewer than three observations or the node is pure. We also stopped partitioning the node when 50 candidate splits failed to reach the $P$-value threshold.

The proposed RMRF algorithm is an extension to the acceptance-rejection trees algorithm for independent data (Calhoun *et al.*, 2019). We also consider selecting `mtry` (eg, $\sqrt{p}$) variables at each node and using an exhaustive search, similar to the RF algorithm by Breiman (2001), or randomly selecting a single cutpoint from each variable selected, similar to the extremely randomized (ER) trees algorithm by Geurts *et al.* (2006). For these alternative approaches, the best split is determined by the maximum absolute value of the robust Wald statistic (or the smallest $P$-value).

## 2.3 | Variable importance

We adopt an area under the curve (AUC) based permuted importance method proposed by Janitza *et al.* (2013). When growing a tree by subsampling from the original sample, the observations not selected are called the OOB sample. The area under the receiver-operating characteristic curve (AUC) for a tree is calculated by comparing each predicted observation in the OOB sample with its actual value. The RMRF algorithm uses the mean predicted value approach, which often outperforms the majority vote method commonly implemented (Malley *et al.*, 2012). The variable importance measures the difference between the OOB AUC before and after permuting the values for a particular variable.

# 3 | SIMULATION

The following simulations compare our proposed RMRF algorithm with the standard RF (Breiman, 2001) and ER trees (Geurts *et al.*, 2006) algorithms. For each simulation, we consider the following five additive models listed in Table 2, where $\rho$ is the number of predictors. Model A gives independent responses, Models B and C simulate responses with a random effect, and Models D and E simulate marginal responses. We use an exchangeable covariance structure when calculating the robust Wald statistic for all five models to assess performance even under model misspecification. Simulations will specify the number of input variables and their respective effect sizes, $\beta$. Most simulations considered only generate a single informative variable. Adding an intercept term was explored, but yielded similar conclusions. Throughout this paper, we refer to patient-level factors as variables having a fixed value within each subject (eg, $x_{ij} = x_{ij'}$ for $\forall j$), while night-level factors can have any value within each subject (eg, $x_{ij} \neq x_{ij'}$ for $j \neq j'$). Using the robust Wald statistic is listed as "Robust," while ignoring the dependence and using the Gini gain is called the "Naïve" approach.

**TABLE 2** Models used for simulations

| Model | Form | Description |
|---|---|---|
| A | $\text{logit}(Y_{ij}) = \sum_{k=1}^{p} \beta_k I(X_{ijk} \le c_k) + \varepsilon_{ij}$, where $\varepsilon_{ij} \sim N(0,1)$ | Independence |
| B | $\text{logit}(Y_{ij}) = \sum_{k=1}^{p} \beta_k I(X_{ijk} \le c_k) + w_i + \varepsilon_{ij}$, where $w \sim N(0,1)$, $\varepsilon_{ij} \sim N(0,1)$ | Random effects, $w \sim N(0,1)$ |
| C | $\text{logit}(Y_{ij}) = \sum_{k=1}^{p} \beta_k I(X_{ijk} \le c_k) + w_i + \varepsilon_{ij}$, where $w \sim N(0,1)$, $\varepsilon_{ij} \sim N(0,3)$ | Random effects, $w \sim N(0,3)$ |
| D | $\text{logit}(Y_{ij}) = \sum_{k=1}^{p} \beta_k I(X_{ijk} \le c_k) + \varepsilon_{ij}$, where $\varepsilon_{ij} \sim N(0,\Sigma), \Sigma_{ij} = \begin{cases} 1 \text{ for } i=j \\ 0.4 \text{ for } i \ne j \end{cases}$ | Marginal, exchangeable $\rho = 0.4$ |
| E | $\text{logit}(Y_{ij}) = \sum_{k=1}^{p} \beta_k I(X_{ijk} \le c_k) + \varepsilon_{ij}$, where $\varepsilon_{ij} \sim N(0,\Sigma), \Sigma_{ij} = \begin{cases} 1 \text{ for } i=j \\ 0.9 \text{ for } i \ne j \end{cases}$ | Marginal, exchangeable $\rho = 0.9$ |

In this paper, we consider simulation settings with a sample size of $N = 100$ and $n_i = 4$ for $i = 1, \ldots, N$ (eg, 100 patients each with 4 nights). We found that both the robust Wald statistic and Gini gain performed better for larger datasets, but yielded the same conclusions. This analysis consisted of 5000 simulations, except for the RF comparison (Section 3.2) where 500 simulations each with 1000 trees were analyzed. The focus of this paper is to compare the effectiveness of the proposed splitting method and the Gini gain for identifying informative variables. We also compared the performance of capturing the true cutpoint in Web Appendix B.

## 3.1 | Identifying informative variables

We consider two simulation scenarios with either one or two informative variables. For this analysis, we generate four covariates $X_1, \ldots, X_4$ from a discrete uniform distribution over 1/10, …, 10/10 where $X_1$ and $X_2$ are night-level factors and $X_3$ and $X_4$ are patient-level factors. We consider either a moderate effect size, $\beta_1 = \log(2)$, or a strong effect size, $\beta_1 = \log(4)$. We restrict attention to cutpoints at $c_1 = 0.3$. We note that only $X_1$ affects the response in the first simulation scenario.

For each simulated dataset, a single split is obtained by maximizing the robust Wald statistic or the Gini gain. The proportion of trees that split on the informative variable is recorded. Figure 1 gives a graph of the simulation results. For Model A under independence, the robust Wald statistic is slightly less likely to split on the informative variable over the Gini gain (83% vs 86%). However, for Models B-E, the robust Wald statistic performs better, particularly for highly correlated responses (55% vs 16% for Model C and 96% vs 47% for Model E). For stronger effects, $\beta_1 = \log(4)$, both splitting statistics capture the true informative variable for >95% of simulations under Models A, B, D, and E. However, for Model C, the robust Wald statistic has 91% of simulations split on $X_1$ compared to only 49% with Gini gain.

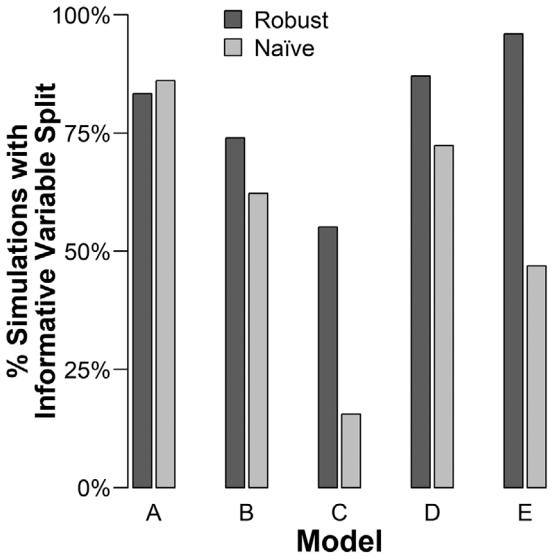


**FIGURE 1** Percentage of simulations where best split uses informative variable

*Note*. Each simulation consists of two night-level variables and two patient-level variables. One of the night-level variables is informative with $\beta_1 = \log(2)$. A single binary split is formed and the percentage of simulations with the informative variable is reported.

We also compare the robust Wald statistic and Gini gain when simulating two informative variables: one night-level and one patient-level factor. We let $\beta_1 = \log(2)$ and $\beta_3 = \log(2)$ with cutpoints 0.3 and grow a tree with a depth of 2 (ie, four terminal nodes). The "Inclusive" signal indicates both $X_1$ and $X_3$ (and possibly variables $X_2$ and $X_4$) are in the tree,

**TABLE 3** Performance of splitting statistics identifying informative variables

| | | Variable selection | | Prediction | |
|---|---|---|---|---|---|
| Model | Method | Inclusive | Exclusive | Average accuracy | Average AUC |
| A | | | | | |
| | Robust | 79 | 33 | 58.5 | 0.574 |
| | Naïve | 82 | 36 | 58.4 | 0.576 |
| B | | | | | |
| | Robust | 68 | 24 | 56.5 | 0.552 |
| | Naïve | 65 | 22 | 56.3 | 0.551 |
| C | | | | | |
| | Robust | 44 | 13 | 52.6 | 0.517 |
| | Naïve | 29 | 7 | 52.1 | 0.513 |
| D | | | | | |
| | Robust | 75 | 33 | 58.2 | 0.570 |
| | Naïve | 70 | 27 | 57.9 | 0.567 |
| E | | | | | |
| | Robust | 64 | 35 | 58.1 | 0.567 |
| | Naïve | 53 | 16 | 57.0 | 0.554 |

*Note*. Each simulation consists of two night-level variables and two patient-level variables. One of the night-level variables and one of the patient-level variables are informative with $\beta_1 = \log(2)$ and $\beta_3 = \log(2)$. A tree with a depth of 2 is grown. The "Inclusive" signal indicates that both informative variables are in the tree, while the "Exclusive" signal indicates that only informative variables are in the tree. The average accuracy and area under the receiver-operating characteristic (AUC) is calculated on a test dataset.

while the "Exclusive" signal indicates only $X_1$ and $X_3$ are in the tree. We assess prediction accuracy by simulating a large test dataset. The average accuracy and AUC are calculated for each simulation.

Table 3 gives the percentage of simulations capturing the two informative variables. For Model A, the robust Wald statistic performs slightly worse than the Gini gain with a smaller proportion of "Inclusive" simulations (79% vs 82%) and "Exclusive" simulations (33% vs 36%). For Models B–E, the robust Wald statistic outperforms the Gini gain, especially for Model C (44% vs 29% Inclusive and 13% vs 7% Exclusive, respectively) and Model E (64% vs 53% Inclusive and 35% vs 16% Exclusive).

The results with two informative variables were highly dependent on the effect sizes. A larger patient-level effect gives much better relative performance for the robust Wald statistic. For example, if we let $\beta_1 = \log(2)$ and $\beta_3 = \log(4)$ under Model E, the exclusive signal increases to 53% for the robust Wald statistic, while remains at 16% with the Gini gain. However, letting $\beta_1 = \log(4)$ and $\beta_3 = \log(2)$, the exclusive signal is both 33% for both splitting statistics under Model E. Since the Gini gain favors patient-level factors with a dependent response, it only needs a small effect size to split on $X_3$. However, when the night-level effect is small, the Gini gain will often miss the informative variable. The robust Wald statistic has also slightly higher prediction accuracy under all five models and slightly higher AUC under Models B–E. Overall, the differences in prediction performance are not as extreme as differences in informative variable detection.

## 3.2 | Comparing random forest algorithms

The previous simulations focused on a single tree and showed that the standard algorithm that naïvely assumes independence prefers patient-level factors under dependent responses. We now investigate differences between our proposed RMRF algorithm with ER trees, standard RF, and generalized linear models (GLM).

We generate five covariates $X_1, \ldots, X_5$ from a discrete uniform distribution over 1/10, …, 10/10 where $X_1$, $X_2$, and $X_3$ are night-level factors and $X_4$ and $X_5$ are patient-level factors. We let $\beta_1 = \log(2)$ and $c_1 = 0.3$ as before; thus, only $X_1$ affects the response. For each simulation, an RF is grown with each tree having a maximum depth of 1 (ie, a forest of decision stumps). This analysis consists of 500 simulations with each simulation growing an RF of 1000 trees. Since each tree consisted of a single split, we let variable importance be the proportion of trees containing a given predictor for this simulation analysis. A GLM was also fit using standardized variables to determine the importance. A robust GLM uses a marginal method with an exchangeable covariance structure, while the naïve is a logistic regression model. GLM treated the covariates as continuous as the true binary split would not be known a priori.

Table 4 gives the percentage of simulations where $X_1$ had the highest variable importance. The RMRF algorithm performed better than ER and RF for all models when using the robust Wald statistic or Gini gain. The most pronounced difference was under a strong random-effect dependence (Model C) using the robust Wald statistic where the RMRF and ER algorithm correctly identified the informative variable on 57.8% of simulations compared to 51.2% with RF. Under independence (Model A), the robust Wald statistic was slightly less likely to identify the informative variable compared with the Gini gain (89.6% vs 91.4% with RMRF). However, for dependent responses the robust Wald statistic performed much better, particularly for strongly correlated response (57.8% vs 21.4% Model C with RMRF, 99.2% vs 68.4% Model E with RMRF). The RF algorithms were also more successful at identifying the informative variable compared with GLM.

Better detection for the RMRF algorithm was also seen when simulating only night-level factors: the RMRF outperformed ER and RF for the majority of models considered and was still more accurate than using Gini gain for the dependent response models (59.6% vs 48.2% Model C, 98.4% vs 91.8% Model E; Web Appendix Table 3). However, the difference in performance between the robust Wald statistic and

**TABLE 4** Performance of random forests and a generalized linear model at identifying informative variables

| Model | Method | RMRF | ER | RF | GLM[a] |
|---|---|---|---|---|---|
| A | | | | | |
| | Robust | 89.6 | 88.8 | 87.4 | 89.4 |
| | Naïve | 91.4 | 87.8 | 89.6 | 89.6 |
| B | | | | | |
| | Robust | 79.8 | 79.6 | 75.8 | 77.8 |
| | Naïve | 73.0 | 71.2 | 68.4 | 76.0 |
| C | | | | | |
| | Robust | 57.8 | 57.8 | 51.2 | 36.6 |
| | Naïve | 21.4 | 20.8 | 14.0 | 30.4 |
| D | | | | | |
| | Robust | 90.4 | 87.4 | 89.8 | 85.0 |
| | Naïve | 80.6 | 78.6 | 78.2 | 83.0 |
| E | | | | | |
| | Robust | 99.2 | 98.6 | 99.2 | 81.6 |
| | Naïve | 68.4 | 68.2 | 55.2 | 74.0 |

*Note*. Each simulation consists of three night-level variables and two patient-level variables. One of the night-level variables is informative with $\beta_1 = \log(2)$. A random forest is grown with 1000 trees each having a single binary split. The variable importance is calculated for each variable and the percentage of simulations with the informative variable having the greatest importance is reported.

[a]Robust GLM uses a marginal model with a logit link function, naïve GLM uses a logistic regression model with a logit link function. Only linear model fits without any interactions.

Gini gain was smaller when there are no patient-level factors. The robust GLM performed better with only night-level factors. However, we explored simulating linear and quadratic relationships, with and without interactions, in Web Appendix Table 4. As the true relationship is unknown and there are often too many possible nonlinear relationships or interactions to consider, the robust GLM treated all covariates as linearly related to the response without any interactions; this is a common assumption in practice. When the relationship was truly linear without any interactions, the GLM model correctly identified the informative variables 98% of the time compared to around 85% for the RF algorithms. However, the RF algorithms were greatly superior when the response had a quadratic relationship or interaction effect.

## 4 | FACTORS ASSOCIATED WITH NOCTURNAL HYPOGLYCEMIA

For patients with diabetes, hypoglycemia is a common phenomenon and is associated with a broad range of adverse events including seizures, unconsciousness, and, on rare occasions, death. Many hypoglycemic events occur at night when patients may not immediately react to some of the symptoms. In a large continuous glucose monitoring (CGM) JDRF study (2010), hypoglycemia (CGM glucose concentration of ≤60 mg/dL) occurred on 8.5% of nights, with 23% of these events lasting more than 2 hours. Additionally, fear of hypoglycemia may result in increased anxiety about diabetes management, deliberately keeping blood glucose levels too high, feelings of guilt and frustration, and other various complications. Identifying factors associated with nocturnal hypoglycemia can improve glucose control and patient care.

The Pump Shutoff (PSO) studies are assessed to determine factors associated with nocturnal hypoglycemia. The datasets are publicly available and detailed information of the two studies is located in the Web Appendix. The two studies (PSO3 and PSO4) were nearly identical but enrolled different cohorts: PSO3 enrolled patients 15 to 45 years old with an HbA1c level ≤8.0% and PSO4 enrolled patients 4 to 14 years old with an HbA1c level ≤8.5%. Additional details of the study and eligibility criteria are described elsewhere (Maahs *et al.*, 2014; Buckingham *et al.*, 2015) and summarized herein.

Enrolled patients used a predictive low-glucose suspend system for 42 nights with half the nights randomly assigned intervention nights and half control nights. Each night the system prompted the subject to enter whether a bedtime snack was consumed and the level of exercise intensity for that day. The continuous glucose, blood glucose, and insulin data were obtained from CGM, metered, and pump downloads, respectively.

The data assessed include only control nights where a predictive low-glucose suspend algorithm was turned off to assess factors associated with nocturnal hypoglycemia. This analysis restricts to nights with at least 6 hours of CGM data in the clinical trial phase. Hypoglycemia is defined as having at least six measurements (30 minutes) below 60 mg/dL.

One hundred twenty-seven patients are analyzed displaying hypoglycemia in 515 (20%) of the 2524 nights. Six patient-level factors and seven night-level factors are assessed to determine any influence on nocturnal hypoglycemia. Web Appendix Tables 5 and 6 give the percentage of nights with hypoglycemia for the patient-level and night-level factors, respectively.

The RMRF algorithm is utilized to predict nocturnal hypoglycemia and assess variable importance. An exchangeable covariance structure is used for this analysis, although a first-order autoregressive covariance structure yielded similar results. We used the AUC-based approach to determine the variable importance.

The variable importance represents the relative importance of the predictor variables, but does not indicate if the influence on nocturnal hypoglycemia is statistically significant. We adopt a permuted importance method proposed by Altmann *et al.* (2010) to test significance. Our AUC-based permuted importance method permutes the responses, constructs a repeated measures an RMRF, and calculates the null variable importance for each predictor. Repeating this process

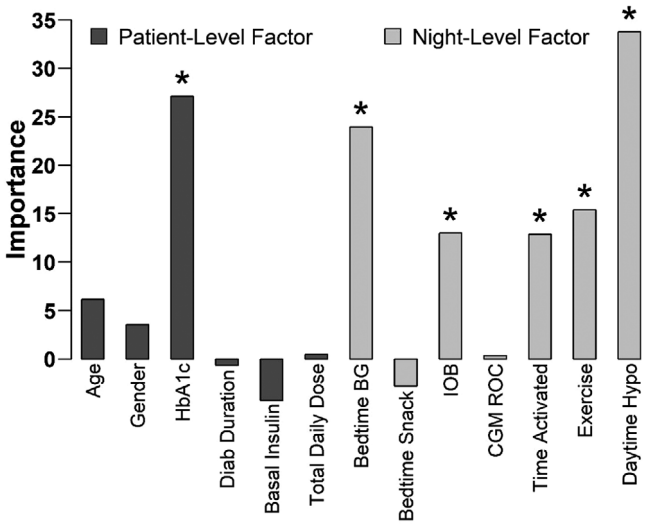


**FIGURE 2** AUC-based variable importance rankings using RMRF algorithm

*Note*. Predictors with an asterisk (*) indicate a statistically significant variable importance score.

several times yields the empirical null distribution of variable importance for uninformative variables. The *P*-value is calculated by comparing the observed variable importance with the null distribution. We use a nonparametric method where the *P*-value is simply the proportion of null importance as or more extreme than the observed importance. This method has been shown to accurately test significance of variables and reduce the variable selection bias found in RFs, since the bias would also occur in the null distribution.

Web Appendix Figure 1 gives the histogram of the null distribution and the observed variable importance for the six significant variables using 100 permutations. HbA1c ($P = .01$), bedtime blood glucose (BG) ($P = .01$), insulin on board (IOB, $P = .03$), time system activated ($P = .02$), exercise ($P = .01$), and daytime hypoglycemia ($P = .01$) were associated with nocturnal hypoglycemia.

The variable importance for all of the factors considered is shown in Figure 2. Predictors with an asterisk (*) indicate significance. Using the standard misclassification (instead of AUC) to calculate variable importance missed many informative factors: total daily dose ($P = .02$) and daytime hypoglycemia ($P = .01$) were the only significant variables when using misclassification. Additionally, we found that naively assuming independence and using the Gini gain will mistakenly find all patient-level factors as significant—reaffirming our simulation studies that Gini gain favors patient-level factors.

We also compared the variable importance of RMRF with a generalized linear mixed-effects model (GLMM) that ranked variables using the estimated standardized regression coefficients. Daytime hypoglycemia had the greatest variable importance, followed by time system activated, bedtime blood glucose, and CGM rate of change. The six most important variables in RMRF were in the top seven most important variables using GLMM although the order of the most important variables differed. Using step-wise or backward selection with GLMM found bedtime BG, exercise, and daytime hypoglycemia were associated with nocturnal hypoglycemia, but did not indicate significance for HbA1c, insulin on board, and time system activated. The AUC from the GLMM after variable selection was 0.589 using fivefold cross-validation compared to 0.622 from RMRF. The GLMM could not achieve the same AUC as RMRF, even when using the same variables determined from RMRF or using variable selection based on the maximum AUC (0.581 and 0.589, respectively). Thus, the RMRF achieved better predictive performance compared with GLMM. The superior predictive performance of RMRF algorithm can be used to classify nights as being high or low risk for nocturnal hypoglycemia, which can help patients better control their blood sugar and manage their diabetes.

## 5 | DISCUSSION

This paper proposes a repeated measures random forest (RMRF) algorithm for analyzing longitudinal data. Our algorithm subsamples the data by subjects and uses a robust Wald statistic to account for the dependence. We also adopt the mean prediction (over majority vote) and an AUC-based permuted importance method. We show that under various simulations the proposed RF algorithm correctly captures the true cutpoint and informative variables.

Ignoring the dependence and using the standard RF algorithm can yield biased variable importance metrics. For strongly correlated responses, RF algorithms that ignore the dependence favor patient-level variables. Under a marginal response with correlation 0.9 (Model E), the robust Wald statistic detects the one informative night-level on 96% of simulations, whereas Gini gain detects only 47%. In this sense, the Gini gain tries to handle the dependence by picking a noninformative subject effect. The robust Wald statistic can capture both night- and patient-level informative factors with equal effect sizes, while the Gini gain only accurately detects the informative patient-level factor.

Even for datasets with only night-level factors, the RMRF algorithm is more likely to identify the influential variable. With strong random-effects dependence (Model C), we find that the informative variable had the highest rank more often for the proposed algorithm over the naïve method (59.6% vs 48.2%). For the scenarios considered, the RMRF algorithm is typically better than ER and RF at detecting the informative variable. The RF method using the robust Wald statistic requires fitting a marginal model for each possible cutpoint, which can be computationally intensive. As there were only

a few possible cutpoints for each predictor in the scenarios considered, the RF algorithm was manageable with parallel computing, but may not be computationally feasible for most other datasets.

The proposed marginal model performed well even under model misspecification. We typically found similar results when a different correlation structure was specified. Furthermore, the robust Wald statistic often captures the true cutpoint and informative variable, even under a random-effects model.

The RF algorithms were better than the GLM at detecting the informative variable when splitting on both night- and patient-level factors. RFs were also able to handle quadratic relationships and detect interactions. Couronne *et al.* (2018) compared RF and logistic regression for 243 real datasets and found that RF performed better in terms of accuracy on 69% of the datasets. We believe that the advantages with RF would naturally extend to repeated measurements, but more research is needed to confirm this conjecture.

Our simulation results did not optimize input parameters. Breiman (2001) recommends choosing the number of randomly selected variables by minimizing the prediction error from the OOB sample. The *P*-value threshold (`minpvalue`) for the RMRF algorithm could also be tuned using the OOB sample. While optimizing parameters could potentially improve performance, the relative differences between the two splitting statistics should be similar.

Our algorithm subsampled the patients (not nights), so the OOB sample was independent of the in-bag sample and allowed for unbiased accuracy estimates. Subsampling also allows the splitting statistic to handle the time dependence since each measurement from the same subject occurs at different times. However, subsampling may not be optimal with highly unbalanced clusters. Additionally, patients with relatively more nights in the OOB sample would have more influence on the variable importance. Our algorithm also used a splitting statistic from a marginal model, but an alternative approach is to use a random-effects model. Typically, marginal models perform best when the true underlying model is marginal, but may not be optimal with random effects (Su and Fan, 2004). However, we ensure that the OOB samples are evaluated using different subjects, so any random effects would be unknown. Even if the predictions were more accurate with a random-effects model, we would still expect the relative difference between the observed and permuted predictions to be similar. Further research to confirm this conjecture is warranted.

This paper considered various simulations: additive models, linear and quadratic relationships, and interactions. We focused on smaller datasets with each predictor having the same number of distinct cutpoints. The RF algorithm under independence encounters a variable selection bias where variables with many distinct cutpoints are given higher importance (Strobl *et al.*, 2007). In a separate paper (Calhoun *et al.*, 2019), we found that the proposed acceptance-rejection algorithm reduced the variable selection bias under independence; however, this bias should also be investigated in a longitudinal setting. The ability of the proposed RMRF algorithm to handle missing observations should also be explored.

We applied our proposed RMRF algorithm to identify factors associated with nocturnal hypoglycemia. The study included 2524 nights from 127 patients. We found that HbA1c, bedtime BG, insulin on board, time system activated, exercise intensity, and daytime hypoglycemia were associated with nocturnal hypoglycemia. Our results agree with the literature. Higher frequency of hypoglycemia has been associated with lower HbA1c level, lower bedtime BG, an exercise session, occurrence of prior hypoglycemia, no bedtime snack, and bolus insulin deliveries unassisted by the bolus estimation calculator (Cryer, 1993; Kalergis *et al.*, 2003; Diabetes Research in Children Network (DirecNet) Study Group, 2007; Weiss *et al.*, 2015). To our knowledge, time system activated was the only factor with no physiological reason for affecting hypoglycemia. A post hoc analysis found that nights with a late system activation time were of shorter duration and had fewer CGM measurements (data not shown); thus, we hypothesize that a late system activation time did not prevent hypoglycemia, but instead reduced the chance of detecting hypoglycemia by having less data.

The RMRF algorithm had greater predictive performance than a conventional GLMM. Greater precision could help identify ways of preventing hypoglycemia and improve patient care. Additionally, controlling for factors associated with nocturnal hypoglycemia (or risk profiles) could reduce variability and improve power for clinical trials comparing two treatments. The RMRF algorithm could also be applied to identify factors that may influence the effectiveness of a device or drug controlling for nonlinear relationships and interactions.

There are several directions for future research. The focus of this paper was on variable importance, but computation time and prediction accuracy should also be investigated. We found that using the robust Wald statistic gave slightly better predictions than the Gini gain for a single tree, so would expect better prediction accuracy with RF. RFs could also be grown with a random subject effects model; further research is warranted comparing trees grown using marginal and random subject effects models.

## ACKNOWLEDGMENTS

We thank the Jaeb Center for Health Research for making the diabetes studies publicly available. This research was supported in part by NSF grant 1633130.

## ORCID

*Peter Calhoun* https://orcid.org/0000-0002-5325-7200

## REFERENCES


Altmann, L., Tolosi, L., Sander, O. and Lengauer, T. (2010) Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26, 1340–1347.

Breiman, L. (2001) Random Forest. *Machine Learning*, 45, 5–32.

Buckingham, B., Raghinaru, D., Cameron, F., Bequette, B., Chase, H., Maahs, D. *et al.* (2015) Predictive low-glucose insulin suspension reduces duration of nocturnal hypoglycemia in children without increasing children ketosis. *Diabetes Care*, 38, 1197–1204.

Buhlmann, P. and Yu, B. (2002) Analyzing bagging. *The Annals of Statistics*, 30, 927–961.

Calhoun, P., Hallett, M., Su, X., Levine, R., Cafri, G. and Fan, J. (2019) Random forest with acceptance-rejection trees. *Computational Statistics*. Available at: https://doi.org/10.1007/s00180-019-00929-4.

Couronne, R., Probst, P. and Boulesteix, A.-L. (2018) Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics*, 19, 270.

Cryer, P. (1993) Hypoglycemia begets hypoglycemia in IDDM. *Diabetes*, 42, 1691–1693.

Diabetes Research in Children Network (DirecNet) Study Group. (2007) Impaired overnight counterregulatory hormone responses to spontaneous hypoglycemia in children with type 1 diabetes. *Pediatric Diabetes*, 8, 199–205.

Geurts, P., Ernst, D. and Wehenkel, L. (2006) Extremely randomized trees. *Machine Learning*, 63, 3–42.

Hajjem, A., Bellavance, F. and Larocque, D. (2011) Mixed effects regression trees for clustered data. *Statistics and Probability Letters*, 81, 451–459.

Hajjem, A., Bellavance, F. and Larocque, D. (2014) Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, 84, 1313–1328.

Hajjem, A., Larocque, D. and Bellavance, F. (2017) Generalized mixed effects regression trees. *Statistics and Probability Letters*, 126, 114–118.

Hallett, M. (2015) *Novel Random Forest and Variable Importance Methods for Correlated Survival Data, with Applications to Tooth Prognosis*. Claremont, CA: Claremont Graduate University.

Hothorn, T. and Zeileis, A. (2015) partykit: A modular toolkit for recursive partytioning in R. *Journal of Machine Learning Research*, 16, 3905–3909.

Janitza, S., Strobl, C. and Boulesteix, A. (2013) An AUC-based permutation variable importance measure for random forests. *BMC Bioinformatics*, 14, 119.

Juvenile Diabetes Research Foundation Continuous Glucose Monitoring Study Group. (2010) Prolonged nocturnal hypoglycemia is common during 12 months of continuous glucose monitoring in children and adults with type 1 diabetes. *Diabetes Care*, 33, 1004–1008.

Kalergis, M., Schiffrin, A., Gougeon, R., Jones, P. and Yale, J. (2003) Impact of bedtime snack composition on prevention of nocturnal hypoglycemia in adults with type 1 diabetes undergoing intensive insulin management using Lispro insulin before meals. *Diabetes Care*, 26, 9–15.

Lee, S. (2005) On generalized multivariate decision tree by using GEE. *Computational Statistics and Data Analysis*, 49, 1105–1119.

Loh, W. and Zheng, W. (2013) Regression trees for longitudinal and multiresponse data. *The Annals of Applied Statistics*, 7, 495–522.

Maahs, D., Calhoun, P., Buckingham, B., Chase, H., Hramiak, I., Lum, J. *et al.* (2014) A randomized trial of a home system to reduce nocturnal hypoglycemia in type 1 diabetes. *Diabetes Care*, 37, 1885–1891.

Malley, J., Kruppa, J., Dasgupta, A., Malley, K. and Ziegler, A. (2012) Probability machines: Consistent probability estimation using nonparametric learning machines. *Methods of Information in Medicine*, 51, 74–81.

McDaniel, L. and Henderson, N. (2016) geeM: Solve Generalized Estimating Equations. R package version 0.10.0. Available at: https://mran.microsoft.com/snapshot/2016-08-05/web/packages/geeM/geeM.pdf.

Segal, M. (1992) Tree-structured methods of longitudinal data. *Journal of the American Statistical Association*, 87, 407–418.

Segal, M. and Xiao, Y. (2011) Multivariate random forests. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1, 80–87.

Sela, R. and Simonoff, J. (2012) RE-EM trees: a data mining approach for longitudinal and clustered data. *Machine Learning*, 86, 169–207.

Sharpsten, L., Fan, J., Barr, J., Su, X., Demirel, S. and Levine, R. (2013) Predicting glaucoma progression using decision trees for clustered data by goodness of split. *International Journal of Semantic Computing*, 7, 157–172.

Strobl, C., Boulesteix, A., Zeileis, A. and Hothorn, T. (2007) Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8, 25–46.

Su, X. and Fan, J. (2004) Multivariate survival trees: a maximum likelihood approach based on frailty models. *Biometrics*, 60, 93–99.

Weiss, R., Garg, S., Bergenstal, R., Klonoff, D., Bode, B., Bailey, T. *et al.* (2015) Predictors of hypoglycemia in the ASPIRE in-home study and effects of automatic suspension of insulin delivery. *Journal of Diabetes Science and Technology*, 9, 1016–1020.

Zhang, H. (1998) Classification tree for multiple binary responses. *Journal of the American Statistical Association*, 93, 180–193.


## SUPPORTING INFORMATION

Supplementary Material

Web Appendices, Tables, and Figures referenced in Sections 2 through 4 and the RMRF algorithm and code are available with this paper at the Biometrics website on the Wiley Online Library. In addition, the code and results can also be found on GitHub: https://github.com/pcalhoun1/RMRF-Code. The diabetes data are publicly available, located at http://jdrfconsortium.jaeb.org. The two protocols can also be found on the ClinicalTrials.gov website (registration numbers NCT01591681 and NCT01823341).

**How to cite this article:** Calhoun P, Levine RA, Fan J. Repeated measures random forests (RMRF): Identifying factors associated with nocturnal hypoglycemia. *Biometrics*. 2020;1–9. https://doi.org/10.1111/biom.13284