RF-DiD-SHAP: A Causal Inference Model in Healthcare Program Evaluation using Longitudinal Data

Peichang Shi,Qingkun Shang,Elizabeth Fuller,Jia Zhao,Stephen Tregear

Booz Allen Hamilton
Mclean, Virginia,22102 USA
shi_peichang@bah.com

Abstract

While mixed-effect models are commonly employed in healthcare program evaluation to manage repeated measures over time, their reliance on strict assumptions, such as normality and linearity, can be limiting. In contrast, the random forest model offers flexibility by not necessitating specific data distributions and accommodating nonlinearity. However, due to an inability to account for cluster and time effects, its effectiveness with time series data presents a challenge. In this study, we propose the RF-DiD-SHAP method, which is situated within the difference-in-differences (DiD) framework. This framework helps explain why the cluster effect and time effect for random forest may not lead to biased estimates of the intervention effect.Using both simulation data and real-world data, our algorithm demonstrates comparable performance to traditional mixed-effect models when assumptions are met. Notably, testing using simulation data revealed that the RF-DiD-SHAP methodology exhibits superior model performance in scenarios where traditional approaches fall short. Of particular significance is the capability of SHAP values to offer causal inference insights into the impact of healthcare programs.

Introduction

Conducting evaluations of healthcare programs is a crucial undertaking for the Centers for Medicare and Medicaid Services(CMS), as it offers a systematic approach to studying the effectiveness of a program, practice, intervention, or initiative in achieving its objectives. Evaluations play a vital role in assessing program performance, identifying successful aspects, and pinpointing areas for improvement. These assessments serve the dual purpose of showcasing impact and informing strategies for program enhancement. Given the nature of healthcare programs, the evaluation process often involves the analysis of longitudinal data, with ongoing assessments conducted over a specified time frame to gauge the efficacy of intervention programs(Adams and Neville 2020; Grembowski 2015).

High-quality randomized controlled trials (RCTs) are regarded as the most robust means of assessing the effectiveness of a program. The RCT approach stands as the gold standard for evaluating treatments or interventions(Hariton and Locascio 2018). In this method, the intervention group is randomly selected from the eligible population, and a control group is also randomly chosen from the same eligible population. However, certain interventions cannot be feasibly evaluated using the single simplicity of a randomized controlled trial. In such cases, quasi-experiments come into play. Quasi-experiments are studies designed to evaluate interventions without employing randomization. Despite lacking randomization, similar to randomized trials, quasi-experiments aim to establish a causal relationship between an intervention and its outcomes(Achen 2021).
The difference-in-differences (DiD) method stands out as a prevalent quasi-experimental technique, comparing the changes in outcomes over time between a population enrolled in a program (the treatment group) and a population that is not (the comparison group)(Wing, Simon, and Bello-Gomez 2018; Dimick and Ryan 2014; Lechner et al. 2011). Traditional methods for implementing the DiD model include Generalized Equation Modeling(GEE) and Generalized Linear Mixed-effect models(GLMM). However, these conventional statistical approaches are hindered by stringent assumptions, potentially limiting their efficacy in handling issues related to non-normality and nonlinearity(Koper and Manseau 2009; Gardiner, Luo, and Roman 2009).

In contrast, the random forest algorithm stands out as a widely adopted machine learning technique, demonstrating superior performance compared to many traditional statistical approaches(Breiman 2001). The strengths of the random forest lie in its ability to effectively handle nonlinearity while being robust against outliers and multicollinearity. These attributes render it a versatile and extensively utilized tool in the field of data science(Biau and Scornet 2016).

While the random forest is often viewed as a somewhat blackbox algorithm, there exist interpretation techniques tailored specifically for tree models. For instance,SHapley Additive exPlanations (SHAP) values represent a prominent method in the realm of explainable artificial intelligence and machine learning(Lundberg and Lee 2017). These values offer a means to comprehend and interpret predictions made by complicated machine learning models without transpancy(Hu and Szymczak 2023). By attributing contributions to each feature in a model's prediction, SHAP values facilitate an understanding of why a specific prediction was generated.Particularly well-suited for tree models, treeSHAP stands out as an existing and convenient interpretation algorithm with notable power(Lundberg, Erion, and Lee 2018).

However, random forest models face criticism when applied to longitudinal data, primarily because the algorithm assumes independent observations. Violations of this independence assumption can result in underestimating standard errors and inaccuracies in confidence interval calculations(Raudenbush and Bryk 2002). Common violations include clustered data and time effects(Karpievitch et al. 2009). Given that longitudinal data is often collected repeatedly from the same subjects over time, time points may be clustered within a subject, or subjects may belong to the same units such as hospitals or nursing homes. Additionally, time effects can introduce autocorrelation.

To address these issues, several extensions of random forest have been proposed. These extensions, inspired by GLMM, aim to replace the linear model of the fixed effect component with a tree or random forest while retaining the modeling of the dependence structure with random effects. Various algorithms have been developed to incorporate trees or random forests into linear mixed-effect models. Many of these extensions build upon the concepts of Multi-step Error Reduction Trees (MERT) and Random Effects Expectation-Maximization (RE-EM) trees(Sela and Simonoff 2012; Liao 2005; Hajjem, Bellavance, and Larocque 2011). The MERT approach employs a decision tree or random forest to estimate fixed effects and bootstrapping to estimate random effects. In contrast, RE-EM trees consider the partition of samples formed by the regression tree, estimating local fixed effects within each partition while estimating global random effects.

While these extensions show promise in addressing cluster and time effects, challenges arise in terms of interpretation. The increased complexity of RE-EM and MERT raises transparency concerns, a critical consideration in the healthcare industry.

This paper introduces a hybrid approach, RF-DiD-SHAP, which integrates the conventional random forest with SHAP values to assess healthcare programs within a DiD framework, aiming to enhance model performance. The key contributions of this study include:

1. Demonstrating through simulation data that RF-DiD-SHAP yields comparable results to traditional linear mixed-effect models when data meets all assumptions;
2. Highlighting the capacity of the random forest model to outperform in situations where data deviates from normality and exhibits nonlinearity;
3. Emphasizing that RF-DiD-SHAP is interpretable and capable of providing causal inference for intervention programs at individual level.

Methods

1. Mixed effect model

   The general mixed-effects model formula is provided below. This formula incorporates the cluster effect into the model. All random effects and residual errors are assumed to have a mean of 0. Therefore, if we disregard the cluster effect, the fixed effect estimates remain unbiased, with the exception of the variance calculation,which is also supported by literature(Raudenbush and Bryk 2002; Liao 2005).

   $$Y_{ij} = \beta_0 + \beta_1 X_{ij} + b_i + e_{ij} \tag{1}$$

   where:

   $$\begin{aligned} Y_{ij} &: \text{the dependent variable,} \\ X_{ij} &: \text{the independent variable,} \\ \beta_0 &: \text{the fixed intercept,} \\ \beta_1 &: \text{the fixed slope,} \\ b_i &: \text{the vector of random effects,} \\ e_{ij} &: \text{the vector of residual errors} \end{aligned}$$

   The random effects b_i are assumed to follow a multivariate normal distribution with mean zero and covariance matrix D:

   $$b_i \sim \text{MVN}(0, D) \tag{2}$$

   The residual errors e_{ij} are assumed to follow a multivariate normal distribution with mean zero and covariance matrix Σ:

   $$e_{ij} \sim \text{MVN}(0, \Sigma) \tag{3}$$

   In our experiments, GEE and GLMM are both utilized for analyzing repeated measures or clustered data, serving as our benchmark models. GEE, being a population-averaged model, is focused on estimating the population mean response. It exhibits robustness to misspecification of the correlation structure and makes fewer assumptions about the distribution of random effects, rendering it less sensitive to model assumptions. On the other hand, GLMM is a subject-specific or individual-level model that incorporates both fixed and random effects, allowing for the modeling of both population-level and subject-specific responses. However, GLMM assumes specific distributions for the random effects and may be sensitive to the correct specification of the random effects distribution. In our model, we adopt an autoregressive model of order 1 to address both cluster and time effects.

2. Time series clustering matching

   The parallel trend assumption is a critical assumption in the DiD model, particularly when analyzing observational data. This assumption is, in the absence of treatment, treatment and control groups would have followed parallel trends over time. Here we employ time series clustering matching techniques based on Euclid distance. This involves grouping units based on their temporal patterns before the treatment, ensuring that treatment and control groups exhibit similar trends. By forming matched clusters, the time series clustering approach helps control for potential confounding factors.

3. RF-DiD-SHAP

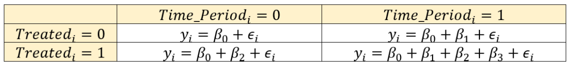
The basic formula for DiD model is below:

$$Y_{ij} = \beta_0+\beta_1\text{post}_i+\beta_2\text{treatment}_j+\beta_3(\text{treatment}_i\times\text{Post}_j)+\epsilon_{ij} \tag{4}$$

where:

Y_{ij} is the outcome for unit i at time j,
treatment$_i$ is 1 if treated, 0 otherwise,
post$_j$ is 0 before,1 after treatment,
ϵ_{ij} is the error term.

In this equation, our focus is on the coefficient $beta_3$ representing the interaction term treatment*post. Utilizing random forest regression for the specified model, we employ treeSHAP to derive SHAP values for $beta_3$ across individual IDs and time points. Subsequently, we apply the following set of four equations to assess the program's impact. ,see figure 1. In order to get intervention effect, we need the following calculations:



	$Time\_Period_i = 0$	$Time\_Period_i = 1$
$Treated_i = 0$	$y_i = \beta_0 + \epsilon_i$	$y_i = \beta_0 + \beta_1 + \epsilon_i$
$Treated_i = 1$	$y_i = \beta_0 + \beta_2 + \epsilon_i$	$y_i = \beta_0 + \beta_1 + \beta_2 + \beta_3 + \epsilon_i$

Figure 1: Equations in DiD model

$$\begin{aligned}
&Y_{00} = E(y|treatment = 0, time = 0) = \beta_0;\\
&Y_{01} = E(y|treatment = 0, time = 1) = \beta_0 + \beta_1;\\
&Y_{10} = E(y|treatment = 1, time = 0) = \beta_0 + \beta_2;\\
&Y_{11} = E(y|treatment = 1, time = 1) = \beta_0 + \beta_1 + \beta_2 + \beta_3;\\
&DiD = Y_{11} - Y_{10} - Y_{01} + Y_{00} = Treatment * Post;
\end{aligned} \tag{5}$$

Assume that the random forest model struggles to effectively account for both cluster and time effects, treating them as unknown bias functions. Υ_c and Υ_t.

For the cluster effect, our assumption posits that entities within the same cluster share identical cluster effects. Therefore, a specific ID exhibits the same cluster effect both before and after treatment. As for the time effect, we assume that the function governing time effect maintains a consistent value for a given time point,so ID in control group and treatment group has the same time effect for a given time point. This assumption leads to the following formulas:

$$\begin{aligned}
&Y_{00} = \beta_0 + \upsilon_{c0} + \upsilon_{t0};\\
&Y_{01} = \beta_0 + \beta_1 + \upsilon_{c0} + \upsilon_{t1};\\
&Y_{10} = \beta_0 + \beta_2 + \upsilon_{c1} + \upsilon_{t0};\\
&Y_{11} = \beta_0 + \beta_1 + \beta_2 + \beta_3 + \upsilon_{c1} + +\upsilon_{t1};\\
&DiD = Y_{11} - Y_{10} - Y_{01} + Y_{00} = Treatment * Post;
\end{aligned} \tag{6}$$

It is evident that while the random forest model might produce biased estimates owing to an unknown error function related to cluster and time effects, such biases are mitigated within the DiD framework, which could cancel out these effects, resulting in minimal impact on the program impact estimate.

Algorithm 1: RF-DiD-SHAP algorithm

Input: Time series data
Parameter:treatment, post, treatment*post, time and other covariates
Output: program impact: SHAP value of interaction term trt*post in the model

```
1: Using time series clustering to get matched treatment group and control group.
2: Run random forest regression model
3: Calculate time effect using SHAP value
4: Deduct time effect from raw outcome
5: Rerun random forest model using updated outcome data
6: Recalculate SHAP values for interaction term treatment*post using SHAP value
7: return solution
```

Results

Data

1. Simulation data
   We employed three simulation datasets to demonstrate the consistent results of RF-DiD-SHAP when compared to generalized linear mixed-effect models. Each dataset spans 12 weeks, with an intervention starting at week 7.
   - The first dataset reflects observations without time effects, such as weight remaining relatively constant over time, with a potential increase post-nutrition program.
   - The second dataset utilizes a control group with consistent weight and cluster effects. The first two datasets follow a normal distribution, serving to assess consistency between traditional statistical models and RF-DiD-SHAP.
   - The third dataset tests the advantages of RF-DiD-SHAP in handling non-linear and distribution-free data. It is characterized by a non-normal Cauchy distribution, with random intervention impact with a mean of 15, and incorporates both time and cluster effects, as seen in the weight rise over time for young kids even without a nutrition program.

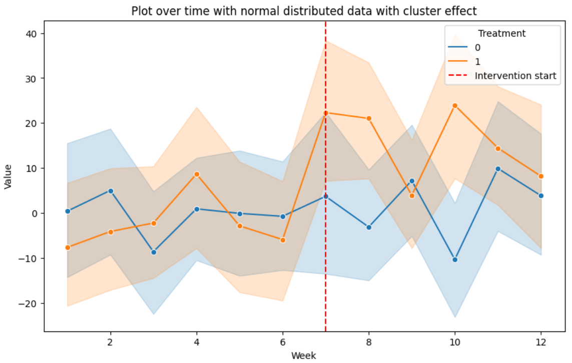


Figure 2: Simulation data with cluster effect

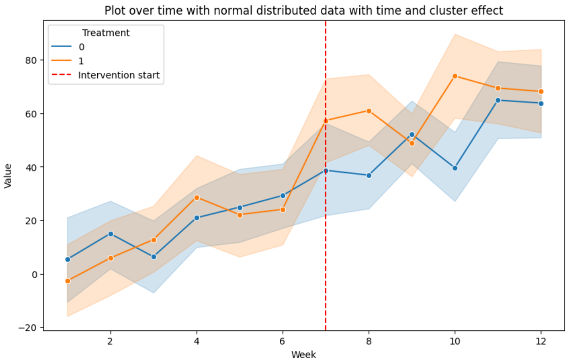


Figure 3: Simulation data with time and cluster effect

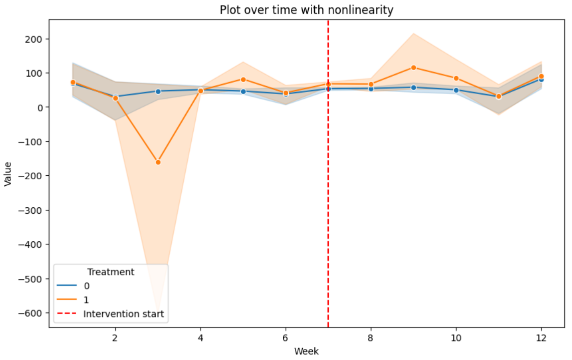


Figure 4: Simulation data with nonlinear and nonnormal distribution

2. Real-world data

   Our population of interest was comprised of 15,447 Medicare-certified nursing homes, of which 3,382 participated in a CMS-sponsored COVID-19 intervention program. We applied time series clustering to COVID incidence rates, considering six time points prior to the intervention. We chose facilities in cluster 0 (see figure 6) as our target population since they meet the parallel trend assumption for DiD analysis. We matched 200 participating facilities with 200 non-participating facilities.In the final matched dataset, the intervention starts at week 6, with 12 additional follow up time points (see figure7).

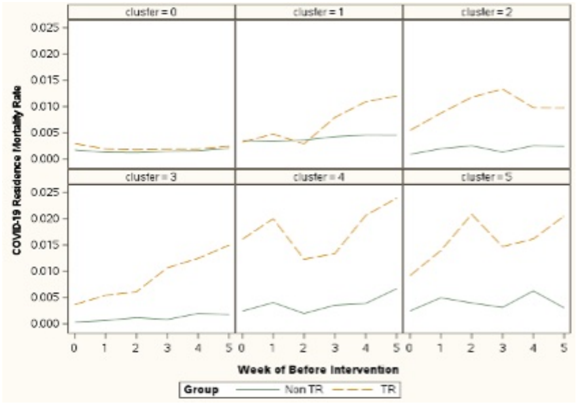


Figure 5: COVID incidence data cluster

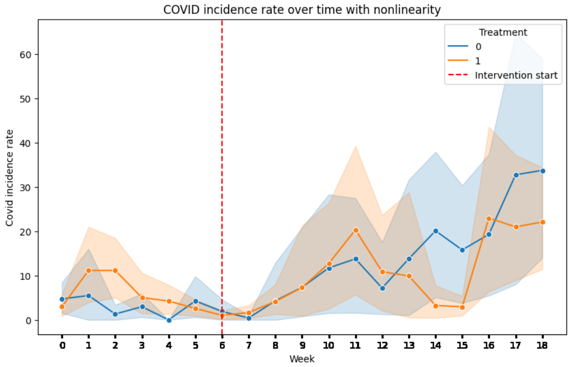


Figure 6: COVID incidence data over time

Findings

1. Data meets all assumptions

   When the data adhered to the necessary model assumptions, the RF-DiD-SHAP model yielded a predicted program impact estimate similar to other models. All three models closely approximated the true fixed program effect of 15 (see table 1). Although the RF-DiD-SHAP model had a slightly wider confidence interval, it provided a more accurate estimate. This suggests that neglecting the cluster effect does not introduce bias into the estimates for both coefficients and errors, as this bias is nullified within the DiD framework. Additionally, both GEE and GLMM produced nearly identical results.

2. Data violates assumptions

   When data come to violation of model assumptions, such as nonlinear and non normality, our RF-DiD-SHAP provide much accurate result compared to traditional statistical approach due to advantages of random forest model based on simulation data, which could greatly enhance our model performance gave different types of data if they could not meet model assumption(see table **??**.

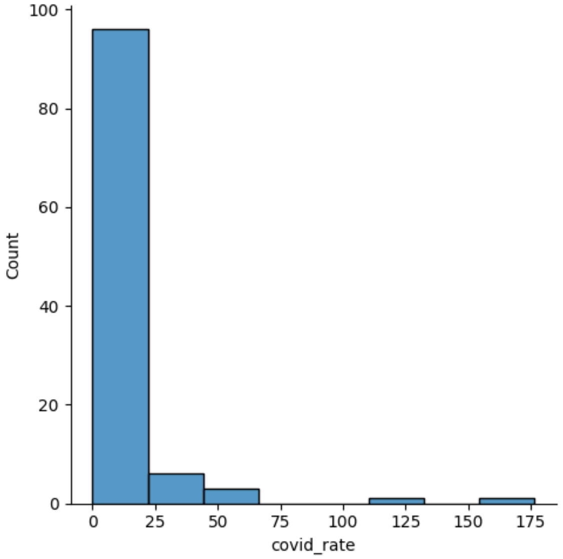


Figure 7: COVID incidence data distribution

In our real-world data,figure 7 illustrates that the data distribution is evidently not normally distributed. In light of this violation of the normality assumption,table 2 displays significantly divergent results for GEE and GLMM (2.73 vs -8.03). This discrepancy raises concerns about the reliability of both models. Given the outcomes of our previous simulations, we are inclined to place greater trust in the RF-DiD-SHAP model, which yielded a value of -1.36.

3. RF-DiD-SHAP could also provide extra value at individual level and time effect As indicated in figure 8, it is evident that the treatment effect is not constant. In the initial 7 weeks, the treatment program exhibits minimal impact on reducing the COVID rate. This observation may be attributed to the time required for facilities to implement the treatment plan, and some outcomes may take time to manifest. Subsequently, from week 8 to week 10, the treatment effect reaches its maximum. Following this period, the treatment effect diminishes, suggesting the possibility of redirecting focus to other areas or modifying the existing intervention programs.

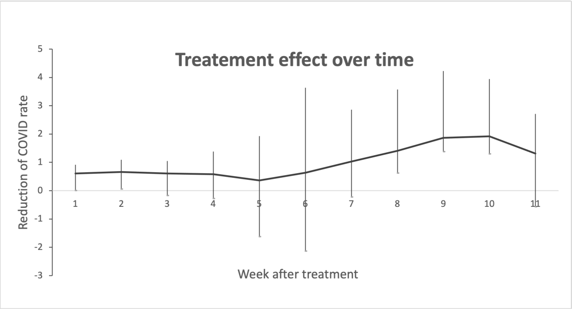


Figure 8: Treatment effect over time for COVID data

Conclusion

In summary, this study introduces the RF-DiD-SHAP method within the difference-in-differences framework. Although the random forest approach alone has limitations in handling time series data, when integrated into the DiD framework and tested through simulation data, the algorithm exhibits comparable performance to traditional mixed-effect models under appropriate assumptions. Notably, RF-DiD-SHAP surpasses traditional approaches in scenarios where they may fall short, as supported by simulation data. The noteworthy capability of SHAP values to provide causal inference insights into the impact of healthcare programs adds to the method's significance. The proposed approach stands as a promising avenue for fortifying the robustness and interpretability of healthcare program evaluations, offering valuable insights even in challenging scenarios.

References

Achen, C. H. 2021. *The statistical analysis of quasi-experiments*. University of California Press.

Adams, J.; and Neville, S. 2020. Program evaluation for health professionals: What it is, what it isn't and how to do it. *International Journal of Qualitative Methods*, 19: 1609406920964345.

Biau, G.; and Scornet, E. 2016. A random forest guided tour. *Test*, 25: 197–227.

Breiman, L. 2001. Random forests. *Machine learning*, 45: 5–32.

Dimick, J. B.; and Ryan, A. M. 2014. Methods for evaluating changes in health care policy: the difference-in-differences approach. *Jama*, 312(22): 2401–2402.

Gardiner, J. C.; Luo, Z.; and Roman, L. A. 2009. Fixed effects, random effects and GEE: What are the differences? *Statistics in medicine*, 28(2): 221–239.

Grembowski, D. 2015. *The practice of health program evaluation*. Sage Publications.

Hajjem, A.; Bellavance, F.; and Larocque, D. 2011. Mixed effects regression trees for clustered data. *Statistics & probability letters*, 81(4): 451–459.

Hariton, E.; and Locascio, J. J. 2018. Randomised controlled trials—the gold standard for effectiveness research. *BJOG: an international journal of obstetrics and gynaecology*, 125(13): 1716.

Hu, J.; and Szymczak, S. 2023. A review on longitudinal data analysis with random forest. *Briefings in Bioinformatics*, 24(2): bbad002.

Karpievitch, Y. V.; Hill, E. G.; Leclerc, A. P.; Dabney, A. R.; and Almeida, J. S. 2009. An introspective comparison of random forest-based classifiers for the analysis of cluster-correlated data by way of RF++. *PloS one*, 4(9): e7087.

Koper, N.; and Manseau, M. 2009. Generalized estimating equations and generalized linear mixed-effects models for modelling resource selection. *Journal of Applied Ecology*, 590–599.

	True fixed effect	Simulation data with cluster effect	Simulation data with time and cluster effect	Simulation data with nonlinear, nonnormal
GEE	15	15.67[14.60,16.73]	14.09[11.89,16.20]	49.77[-21.09,120.63]
GLMM	15	15.67[14.50,16.83]	14.09[11.74,16.44]	49.72[-21.42,120.21]
RF-DiD-SHAP	15	14.91[10.58,17.08]	15.62[11.70,18.13]	12.36[1.89,21.53]

Table 1: Model performance comparison using simulation data

	COVID data
GEE model	2.73[-7.1,12.56]
GLMM model	-8.02[-18.86,2.82]
RF-DiD-SHAP	-1.36[-3.34,0.16]

Table 2: Model performance comparison using COVID data

Lechner, M.; et al. 2011. The estimation of causal effects by difference-in-difference methods. *Foundations and Trends® in Econometrics*, 4(3): 165–224.

Liao, T. W. 2005. Clustering of time series data—a survey. *Pattern recognition*, 38(11): 1857–1874.

Lundberg, S. M.; Erion, G. G.; and Lee, S.-I. 2018. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Raudenbush, S. W.; and Bryk, A. S. 2002. *Hierarchical linear models: Applications and data analysis methods*, volume 1. sage.

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

Wing, C.; Simon, K.; and Bello-Gomez, R. A. 2018. Designing difference in difference studies: best practices for public health policy research. *Annual review of public health*, 39: 453–469.