Received: 9 June 2016 | Revised: 27 July 2017 | Accepted: 26 August 2017

DOI: 10.1002/sim.7500

RESEARCH ARTICLE

WILEY Statistics in Medicine

Validating effectiveness of subgroup identification for longitudinal data

Nichole Andrews[1] | Hyunkeun Cho[2]

[1]Department of Statistics, Western Michigan University, Kalamazoo, MI 49008, USA

[2]Department of Biostatistics, University of Iowa, Iowa City, IA 52242, USA

Correspondence
Hyunkeun Cho, Department of Biostatistics, University of Iowa, Iowa City, IA 52242, USA.
Email: hyunkeun-cho@uiowa.edu

In clinical trials and biomedical studies, treatments are compared to determine which one is effective against illness; however, individuals can react to the same treatment very differently. We propose a complete process for longitudinal data that identifies subgroups of the population that would benefit from a specific treatment. A random effects linear model is used to evaluate individual treatment effects longitudinally where the random effects identify a positive or negative reaction to the treatment over time. With the individual treatment effects and characteristics of the patients, various classification algorithms are applied to build prediction models for subgrouping. While many subgrouping approaches have been developed recently, most of them do not check its validity. In this paper, we further propose a simple validation approach which not only determines if the subgroups used are appropriate and beneficial but also compares methods to predict individual treatment effects. This entire procedure is readily implemented by existing packages in statistical software. The effectiveness of the proposed method is confirmed with simulation studies and analysis of data from the Women Entering Care study on depression.

KEYWORDS

classification algorithm, effectiveness of subgrouping, personalized treatment, random effects linear model

1 | INTRODUCTION

Depression is a serious illness that can significantly impact one's way of life. Treatments for depression include medication and cognitive behavioral therapy, among others. Longitudinal data have been studied to compare treatments and assess the response of a treatment on a patient over time with the goal being to find the most effective treatment. When analyzing data from the Women Entering Care study, Miranda et al[1] found that medication and cognitive behavioral therapy performed similarly in regards to lowering a depression score, leading to the recommendation of either treatment. The inability to show which treatment outperforms the other could be interpreted as a failed study; however, this is not the case. As research emerges on personalized treatment, the focus has shifted from finding one overall beneficial treatment to identifying a subgroup of the population that would have a positive effect from a given treatment.

For longitudinal data, let Y_{ij} be the response for the ith subject at time T_{ij}, where $i = 1, \dots, n, j = 1, \dots, n_i$, and n_i is the number of times measurements are taken on the ith patient. We suppose that n subjects are independent. To evaluate the treatment effect over time, the following marginal regression model could be considered:

$$Y_{ij} = \delta_0 + \delta_1 Z_i T_{ij} + \delta_2 T_{ij} + \xi_{ij}, \tag{1}$$

where $Z_i = 1$ or -1 represents the treatment assignment for patient i, $\delta = (\delta_0, \delta_1, \delta_2)'$ is the parameter vector, and ξ_{ij} are random errors. Generalized estimating equations were proposed to estimate δ using a working correlation structure with nuisance parameters.[2] This enables us to accommodate associations among measurements within the subject, yet it requires estimation of additional nuisance parameters involved in the working correlation structure. Qu et al[3] went a step further and developed quadratic inference functions, which avoids estimation of additional parameters by approximating the working correlation structure with several basis matrices. Both approaches can yield consistent and efficient estimators by accommodating the within-subject correlation commonly existing in longitudinal data.

Analysis with model (1), however, could indicate no difference in the outcome of 2 treatments, resulting in the recommendation of either treatment for use in the population. At the same time, individuals can react very differently to the same treatment. Outside factors, such as biological or environmental influences, can have a significant impact on the outcome of a given treatment. As such, a method for identifying an ideal treatment based on patient characteristics is desired rather than identifying a single beneficial treatment for the entire population.

Song and Pepe[4] proposed a method for subgrouping patients into a particular treatment according to a covariate determined by how this value compared to a prespecified threshold. The use of a single covariate was also used by Bonetti and Gelber,[5] in which patients were grouped by the value of this covariate and analyzed with a moving average procedure. Moskowitz and Pepe[6] used the concept of positive predictive values with a single covariate. The problem with these methods, however, is that more than one variable may be related to the outcome of the treatment. Cai et al[7] were able to use multiple baseline measurements with a two-stage method, where a parametric index score was calculated based on the estimated subject-specific mean response for the treatments. Zhao et al[8] also used a parametric scoring system with multiple baseline covariates. Foster et al[9] proposed the virtual twins method to identify a subgroup for which the treatment effect was better that than the average treatment effect.

Recently, random effects linear models have been studied for personalized treatments, as the model allows each patient to be considered an individual rather than only a member of the population.[10,11] Diaz et al[12] used a random intercept model to model the log of plasma concentrations given certain covariates. Diaz[13] proposed benefit functions for treatment comparison and provided a graphical method for investigating the severity of a disease. Here, the random effects incorporate variability of the response differences in personal characteristics of the patient. Cho et al[14] used a random forest approach in an unspecified random effects model. Zhu and Qu[15] personalized drug dosage over time with a log-linear mixed effect model. Diaz et al[10] also noted that an empirical Bayesian approach under the mixed model framework may have better results for individualizing drug doses.

While the above mentioned procedures can subgroup the data, the effectiveness of their classification has not been fully discussed. Shen and He[16] developed a procedure using a structured logistic-normal mixture model that not only classified the data but also tests for the existence of subgroups. This work was extended by Wu et al[17] for time-to-event data with the semiparametric logistic-cox mixture model. While these methods have advanced work in subgroup analysis, specifications for the data may not always be met.

In this paper, we offer a complete process from subgrouping to validation for personalized treatments in longitudinal studies. Our procedure starts by providing a random effects linear model. The random effects in the model evaluate individual treatment effects over time, yet the fixed effects still allow us to look at the population as a whole. Since the variation in the random effects acts as the variation between characteristics of the patients,[10,13,18] we use various classification approaches to build prediction models based on the individual effects and characteristics of the patients; both linear and nonlinear classification approaches are considered, since the association between the characteristics of the patient and the outcome are unknown in practice. While subgrouping can be performed based on the prediction models, the question of its appropriateness and which model is best remains unanswered. Therefore, a validation procedure has been developed to choose the best prediction model under the marginal regression framework.

While many methods have been developed for classifying data, the advantage of the proposed method is that it uses supervised learning algorithms already developed, making them easier to implement and interpret. In addition, the proposed procedure can be readily applied to a longitudinal medical study where all follow up appointments may not be attended, therefore resulting in missing measurements. Moreover, the validation approach allows us to not only analyze the treatment effect over time for those that received the treatment deemed beneficial with the prediction model but also takes into account a time effect. This is an important aspect; while we may desire that the outcome decreases over time, this may not happen. Including a time effect allows us to analyze whether or not the treatment slows the progression of the illness. Since we are able to assess the validity of our classification and determine the best prediction model, our steps outline the entire procedure for determining an appropriate subgroup.

The paper is organized as follows. Section 2 outlines the proposed methodology, including the random effects model for treatment effect over time, the prediction models for subgrouping, and the validation of the models. Results of the methodology on simulation studies are in Section 3. Section 4 analyzes data from the Women Entering Care study on depression among low-income and minority women. The discussion in the final section presents some additional thoughts.

2 | METHODOLOGY

2.1 | Evaluating individual treatment effects

Since a random effects linear model has been shown to be effective in the analysis of longitudinal data, we consider the model that evaluates the treatment effect, specifically its effect over time, on a response. Accordingly, the random slope intercept model is formulated as

$$Y_{ij} = \beta_0 + \alpha_{0i} + (\beta_1 + \alpha_{1i})Z_i T_{ij} + \beta_2 T_{ij} + e_{ij}, \quad i = 1, \ldots, n \quad j = 1, \ldots, n_i, \tag{2}$$

where α_{0i} and α_{1i} are the random intercept and slope for subject i, respectively, and e_{ij} are random errors. While β_1 represents the overall average of the treatment effect over time, α_{1i} enables us to take into account individual differences. By considering the interaction effect between the treatment and time, model (2) allows us to evaluate the individual treatment effect on the response over time.

We estimate the parameters in model (2) using maximum likelihood estimation. Without loss of generality, we suppose that the number of measurements taken on each subject are the same (ie, $n_i = k$ for all i) and rewrite model (2) as

$$Y = G\beta + D\alpha + e, \tag{3}$$

where Y and e are nk-dimensional vectors of the responses and errors, and G and D are $nk\times3$ and $nk\times2n$ matrices of covariates corresponding to the fixed effect $\beta = (\beta_0, \beta_1, \beta_2)'$ and random effect $\alpha = (\alpha_{01}, \ldots, \alpha_{0n}, \alpha_{11}, \ldots, \alpha_{1n})'$, respectively. Assuming a multivariate normal distribution

$$\begin{pmatrix} \alpha \\ e \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Omega & 0 \\ 0 & \Sigma \end{pmatrix}\right),$$

model (3) can also be expressed as $Y = G\beta + e^*$, where $e^* = D\alpha + e$, resulting in $e^* \sim N(0, M)$ with $M = D\Omega D' + \Sigma$. Accordingly, the estimators for β and α are obtained as $\hat{\beta} = (G'\hat{M}^{-1}G)^{-1}G'\hat{M}^{-1}Y$ and $\hat{\alpha} = \hat{\Omega}D'\hat{M}^{-1}(Y - G\hat{\beta})$, respectively. Here, $\hat{\Omega}$ and $\hat{\Sigma}$, and ultimately $\hat{M}$, are obtained by maximizing the following log-likelihood function:

$$\begin{aligned} l(\Omega, \Sigma) = &-\frac{1}{2}(Y - G(G'M^{-1}G)^{-1}G'M^{-1}Y)'M^{-1}(Y - G(G'M^{-1}G)^{-1}G'M^{-1}Y) \\ &- \frac{1}{2}log|M| - \frac{nk}{2}log(2\pi), \end{aligned}$$

where $|M|$ is the determinant of the covariance matrix M. While this is computationally intensive, advances with technology and software make this a nonissue. These estimates are asymptotically consistent and efficient.[19] When the estimate of M is biased, the restricted maximum likelihood is a viable alternative approach.[20]

Since model (2) provides the individual treatment effect on the response over time, we can split all subjects into 2 groups according to whether or not they had a positive effect. For this, an indicator of C_i is assigned to each subject based on the sum of the fixed slope estimate and random slope estimate for the interaction between treatment and time, where $C_i = 1$ if $\hat{\beta}_1 + \hat{\alpha}_{1i} > 0$ and -1 otherwise.

2.2 | Building prediction models

After model (2) is fitted to longitudinal data, we build prediction models to subgroup the data by treating the binary outcome of C_i as the response variable. The corresponding independent variables, denoted by X_i, contains characteristics of patient i that are deemed influential to the assignment of the treatment. This could include variables such as, but not limited to, age, gender, and race. The use of C_i as the response is key, as it is determined by the parameter estimate for the interaction between treatment and time for each patient and is not an observed value from the data.

Since we classify observations into one of 2 subgroups, the desired prediction model is specified as $f(X_i) = P(C_i = 1|X_i)$, where $f(\cdot)$ is a function representing the association between C_i and X_i. In reality this relationship is unknown. It could be either linear or nonlinear, however this lack of information makes the function $f(\cdot)$ unidentifiable. As such, various

prediction models are constructed through both types of supervised learning algorithms; linear (logistic regression, linear discriminant analysis (LDA), and support vector machine (SVM) with a linear kernel) and nonlinear (quadratic discriminant analysis (QDA), decision tree, random forest, and SVM with a radial kernel). We denote the estimated prediction model by $\hat{f}(X_i)$ and classify patient i as $\hat{C}_i = 1$ if $\hat{f}(X_i) > 0.5$ and -1 otherwise.

Among these supervised learning algorithms, we expect logistic regression, LDA, and SVM with a linear kernel to provide an accurate prediction model if the predictors are linearly associated with the response. Likewise, we expect these methods to perform poorly and QDA, decision tree, random forest, and SVM with a radial kernel to perform well if the relationship between C_i and X_i is not linear. After we use the supervised learning algorithms to build the prediction models, we assess these results with the validation approach outlined below.

2.3 | Validating prediction models

While classification can be performed on our data, the question still remains of whether the subgrouping was effective or not. Therefore, a validation approach has been developed to tackle this problem. Suppose that a higher response is desired over time. Then treatment $Z = 1$ is deemed more beneficial than treatment $Z = -1$ for patient i if $\hat{C}_i = 1$, as $\hat{\beta}_1 + \hat{\alpha}_{1i}$ is the parameter estimate for the interaction term of $Z_i T_{ij}$ in model (2) and $C_i = 1$ means this estimate is positive. Likewise, treatment $Z = -1$ is deemed more beneficial for patient i if $\hat{C}_i = -1$.

In this section, we assume that the desired outcome is for the response to decrease over time, which corresponds to our application of the depression study (ie, treatments $Z = 1$ and -1 are deemed more beneficial for patients whose $\hat{C}_i$ are -1 and 1, respectively). For each subgrouping method described in Section 2.2, let U_i be the indicator that the patient received the treatment deemed to be more beneficial through the prediction model: $U_i = 1$ if the patient received the treatment predicted to be more beneficial and $U_i = -1$ otherwise. We then formulate the following marginal regression model:

$$Y_{ij} = \gamma_0 + \gamma_1 U_i T_{ij} + \gamma_2 T_{ij} + \varepsilon_{ij} \tag{4}$$

and estimate parameters γ_k, $k = 0, 1, 2$, using the generalized estimating equation approach[2] that can yield unbiased and more efficient estimators than the one ignoring the within-subject correlation. We remark that while this may appear to be similar to model (1), the key difference is the use of U_i rather than Z_i. We are no longer concerned with which treatment the patient received, as we were in model (1), but rather with whether or not the subject received the treatment that was deemed beneficial.

For patients who receive the more beneficial treatment, we should notice a larger decrease in their response over time, thus $\hat{\gamma}_1$ should be significantly negative. Here we let the proposed subgrouping analysis be appropriate and beneficial if $H_0 : \gamma_1 = 0$ is rejected against $H_a : \gamma_1 < 0$ using the Wald test. It may be the case that multiple subgrouping approaches prove to be beneficial but one must be chosen. The best subgrouping approach is the one that distinguishes the two groups (did and did not receive the treatment predicted to be beneficial) the most. This is determined by the one with the largest Wald test statistic among effective prediction models.

3 | SIMULATION STUDIES

In this section, we assess the proposed method through 3 different types of simulation studies. First, we assume that subgrouping is appropriate and use both a linear and nonlinear form of the random slopes. Finally, we assume that subgrouping is not appropriate and generate random slopes that are not dependent on the data. For these, a sample size of 200 for the training dataset and 100 for the testing dataset were modeled as

$$Y_{ij} = \beta_0 + \alpha_{0i} + (\beta_1 + \alpha_{1i}) Z_i T_{ij} + \beta_2 T_{ij} + e_{ij}, \quad j = 1, \ldots, 6, \tag{5}$$

where $(\beta_0, \beta_1, \beta_2)' = (0, 0, -0.2)'$, Z_i was randomly chosen as either -1 or 1 for the treatment assignment with a probability of .5, T_{ij} was the index of time j, α_{0i} was randomly generated from a uniform distribution between -1 and 1, and $e_i = (e_{i1}, \ldots, e_{i6})'$ was randomly selected from a multivariate normal distribution with mean 0 and variance-covariance matrix R, where all elements on the diagonal of R are 1 and 0.7 otherwise, which corresponds to a compound symmetry structure with a correlation coefficient of 0.7. Six independent variables were used, which act as characteristics of the patient; X_{1i}, X_{2i}, and X_{3i} were generated randomly from a standard normal distribution, while X_{4i}, X_{5i}, and X_{6i} were binary variables assigned a value randomly chosen as either -1 or 1 for subject i with a probability of .5.

The training dataset was used to fit model (5) and the 7 supervised learning algorithms described in Section 2.2 were used based on the predictor vector $X_i = (X_{1i}, X_{2i}, X_{3i}, X_{4i}, X_{5i}, X_{6i})'$. Once 7 prediction models were developed on the training dataset, subjects in the testing dataset were classified based on these models; misclassification error rates were computed, defined as the proportion of times $C_i \neq \hat{C}_i$ for $i = 1, \ldots, 100$. Finally, the proposed validation approach was used to determine the appropriateness of our subgrouping and the best subgrouping method. A total of 1000 simulations were run for each type of random slope.

3.1 | Linear association

We model our random slopes as

$$\alpha_{1i} = -0.5X_{1i} + X_{4i} + \zeta_i,$$

where ζ_i is the error term randomly generated from a standard normal distribution. The average of the misclassification error rates reported in Table 1 show that all methods except the decision tree perform similarly with the linear approaches producing slightly lower error rates.

In addition, the validity of our classification was assessed on the testing dataset. Table 1 also displays the proportion of times each method produced a significantly negative $\hat{\gamma}_1$ at a nominal level of 0.05, as well as the average and standard deviation of $\hat{\gamma}_1$ among the 1000 simulations. Since we were assuming that a lower response is desired over time, $\hat{\gamma}_1$ is significantly negative if the proposed prediction model is effective. This was achieved, as shown in Table 1; each subgrouping approach produced a significant parameter estimate all 1000 times, indicating that the proposed method performs well in the case of linear random slopes. Moreover, the average and standard deviation of $\hat{\gamma}_1$ were approximately the same for all methods except the decision tree, suggesting that both linear and nonlinear classification approaches performed relatively equally. Due to ease of interpretation and simplicity, however, we would recommend the use of a linear classification approach here.

We note that we intentionally set X_{4i} to be a strong variable and X_{1i} to be weaker to find if logistic regression and the random forest algorithm would detect these variables as significant. For logistic regression, variables were considered significant if their corresponding P value was less than .05. X_{4i} was always shown to be a significant variable in the model while X_{1i} was significant 99.3% of the time. Moreover, the remaining 4 variables were significant 5% to 6% of the time. In addition, we were able to identify important factors when the random forest algorithm was implemented; X_{4i} was always considered the most important factor and X_{1i} was the second most important variable over 91% of the time.

3.2 | Nonlinear association

We used the same information as above, while adding 2 nonlinear components to the random slopes in Section 3.1. The nonlinear random slopes were then generated as

$$\alpha_{1i} = -0.5X_{1i} + X_{4i} + X_{1i}X_{2i} - 0.7X_{3i}^2X_{4i} + \zeta_i.$$

The results in Table 2 confirm that all nonlinear approaches outperformed the linear ones in terms of a lower misclassification error rate; SVM with a radial kernel had the lowest error rate with the random forest algorithm less than 1% behind. We also remark that X_{5i} and X_{6i} were never considered among the top 3 most influential variables among all 1000

TABLE 1 Misclassification error rates and validation results on testing data for linear random slopes

Type	Method	Error Rate	Proportion	Mean ($\hat{\gamma}_1$)	SD ($\hat{\gamma}_1$)
Linear	Logistic	19.17%	1.000	−1.976	0.233
	LDA	18.87%	1.000	−1.994	0.227
	SVM (linear)	19.05%	1.000	−1.989	0.228
Nonlinear	QDA	19.45%	1.000	−1.970	0.234
	Decision tree	24.45%	1.000	−1.670	0.301
	Random forest	20.57%	1.000	−1.899	0.240
	SVM (radial)	19.12%	1.000	−1.982	0.229

Abbreviations: LDA, linear discriminant analysis; QDA, quadratic discriminant analysis; SVM, support vector machine

TABLE 2 Misclassification error rates and validation results on testing data for nonlinear random slopes

Type	Method	Error Rate	Proportion	Mean ($\hat{\gamma}_1$)	SD ($\hat{\gamma}_1$)
Linear	Logistic	39.06%	0.718	−0.815	0.383
	LDA	39.02%	0.718	−0.814	0.385
	SVM (linear)	39.27%	0.666	−0.737	0.400
Nonlinear	QDA	33.24%	0.913	−1.088	0.369
	Decision tree	34.73%	0.952	−1.280	0.418
	Random forest	30.39%	0.997	−1.587	0.360
	SVM (radial)	29.63%	0.999	−1.615	0.350

Abbreviations: LDA, linear discriminant analysis; QDA, quadratic discriminant analysis; SVM, support vector machine

TABLE 3 Misclassification error rates and validation results on testing data for randomly generated random slopes

Type	Method	Error Rate	Proportion	Mean ($\hat{\gamma}_1$)	SD ($\hat{\gamma}_1$)
Linear	Logistic	50.10%	0.057	−0.002	0.200
	LDA	50.10%	0.057	−0.002	0.200
	SVM (linear)	50.10%	0.053	0.002	0.209
Nonlinear	QDA	50.10%	0.054	−0.001	0.200
	Decision tree	49.90%	0.064	−0.002	0.202
	Random forest	50.10%	0.048	0.008	0.200
	SVM (radial)	50.20%	0.060	0.004	0.208

Abbreviations: LDA, linear discriminant analysis; QDA, quadratic discriminant analysis; SVM, support vector machine

simulations with the random forest approach. Considering these 2 variables were the only ones not used in the calculation of the above random slopes, this result was not surprising.

Table 2 also shows that the prediction models based on the nonlinear classification approaches were better than those of the linear type in terms of a higher proportion of times that $\hat{\gamma}_1$ was significant; this estimate also has a smaller value, indicating that the predicted beneficial treatment lowers the response more over time. When comparing the nonlinear algorithms, SVM with a radial kernel and random forest were the best classification methods studied. For these methods, the average $\hat{\gamma}_1$ was −1.615 and −1.587, respectively. On the other hand, the linear approaches all had the highest error rates and accordingly had the fewest significant parameter estimates with the validation approach. In fact, the proportion of significant parameter estimates decreased by about 30% with the linear approaches from when we had linear random slopes in Section 3.1. In addition, the average $\hat{\gamma}_1$ for these methods ranged from −0.737 to −0.815, indicating the subgrouping is beneficial but not as beneficial as that of the nonlinear methods.

3.3 | No association

In the 2 previous simulation studies, the random slopes were modeled based on a subset of the independent variables. We now investigate when the random slopes are not dependent on the data at all. This represents the null hypothesis of subgrouping not being appropriate. Accordingly, we generated α_{1i} randomly from a standard normal distribution.

Table 3 displays the results from the validation approach. Regardless of the classification approach, the proportion of times that the prediction model is deemed significant through validation is close to a nominal level of 0.05. This proportion also represents the type I error, where we recommend subgrouping when it is not appropriate. These results indicate that when subgrouping is not appropriate, all the classification methods do not recommend any subgrouping. The averages of the misclassification error rates are also reported in Table 3. These are all near 50%, indicating that we are just as likely to correctly classify an individual as we are to misclassify them. This is because the random slopes are not associated with the data at all, yet the prediction models are built with the independent variables of X_i. As such, the classification approaches cannot perform well.

4 | DATA ANALYSIS FOR DEPRESSION STUDY

In this section, the proposed subgrouping method was applied to the Women Entering Care study on depression that involved low-income and minority women. Information on this data can be found in Miranda et al.[1] Here, we present a brief summary. The response was the Hamilton Depression Scores, assessed every month for the first 6 months, including a baseline observation, then every other month for the remainder of the year. Each patient was randomly assigned to 1 of 3 treatment groups: medication ($n = 88$), cognitive behavioral therapy ($n = 90$), and referral to community care (treatment as usual) ($n = 89$). Since the goal is to analyze the treatment effect over time, we only considered patients who had a baseline depression score and at least one follow-up. While all patients had an initial score, 11 did not have any other depression scores and were therefore excluded from our analysis (medication $n = 86$, cognitive behavioral therapy $n = 85$, referral to community care $n = 85$).

Table 4 gives us an initial look at the data at various time points. Once the assigned treatment had been implemented, we noticed better (lower) scores among those in the medication and cognitive behavioral therapy groups. Miranda et al[1] also found these 2 treatments to be more effective at treating depression than being referred to community care when evaluating the data from just the first 6 months.

To assess our procedure, we split the data into 2 smaller datasets with two-thirds of the data in the training dataset and a third in the testing. Our proposed method had similar findings as that of Miranda et al,[1] yet we used all observations rather than just the first 6 months. When using the training dataset to compare medication and cognitive behavioral therapy to the referral group, our random effects linear model estimated that the parameters for the interactions between treatment and time were $\hat{\beta}_1 = -0.1383$ and -0.1344 (t-value $= -2.60$, $df = 95$, P value $= .009$, and t-value $= -2.46$, $df = 93.5$, P value $= .013$), respectively. Therefore, our method was also able to show that medication and cognitive behavioral therapy are better at treating depression than being referred to community care.

Our attention shifted to comparing the medication and cognitive behavioral therapy groups. We start by fitting mean regression model (1) to the training dataset and using the generalized estimating equations with an AR(1) correlation structure, as this is an established method. With this, the parameter estimate for the interaction between treatment and time for the training dataset was $\hat{\delta}_1 = 0.0238$ (Wald $= 0.19$, P value $= .660$), meaning we could not determine a difference in outcomes of the treatment and would recommend either to a patient. Using proposed random effects model (2) resulted in an estimate of $\hat{\beta}_1 = -0.0017$ (t-value $= -0.0168$, $df = 94.6$, P value $= .980$), leading to the inability to conclude a significant difference in average outcomes between treatments. Therefore, analysis with the random effects of model (2) was performed, taking into account individual treatment effects over time. When building the prediction models, 8 independent variables were used for each patient: 6 binary variables (marital status, schooling, housing, ethnicity, where the patient was born, and whether or not the patient works) and 2 continuous variables (baseline depression score

TABLE 4 Means and confidence intervals for depression scores

Time	Medication Mean (95% CI)	Cognitive Behavioral Therapy Mean (95% CI)	Control Mean (95% CI)
Baseline	18.08 (16.99-19.17)	16.35 (15.21-17.49)	16.54 (15.42-17.66)
Month 3	9.60 (8.02-11.19)	10.24 (8.56-11.92)	13.05 (11.22-14.88)
Month 6	9.17 (7.41-10.94)	10.73 (8.95-12.52)	11.92 (10.14-13.70)
Month 12	9.71 (7.70-11.72)	8.38 (6.72-10.05)	10.22 (8.70-11.75)

TABLE 5 Validation results on testing dataset for depression data

Type	Method	$\hat{\gamma}_1$	SE	Wald	*P* value
Linear	Logistic	−0.029	0.203	0.02	.444
	LDA	−0.029	0.203	0.02	.444
	SVM (linear)	−0.046	0.205	0.05	.411
Nonlinear	QDA	−0.187	0.199	0.89	.173
	Decision tree	−0.243	0.192	1.61	.103
	Random forest	−0.583	0.183	10.10	.001
	SVM (radial)	−0.182	0.199	0.84	.180

Abbreviations: LDA, linear discriminant analysis; QDA, quadratic discriminant analysis; SVM, support vector machine

and age). Table 5 displays the results when using model (4) to check the validity of our approach on the testing dataset with prediction models built on the training dataset. All parameter estimates for the interaction between treatment and time were negative, indicating that the depression score decreases over time for those individuals that received the treatment deemed to be beneficial. Not all methods, however, produced significant results; the only method that found subgrouping to be appropriate and beneficial was the random forest algorithm.

None of the linear classification approaches were considered significant. In fact, logistic regression did not detect any of the predictors as significant. Siddique et al[21] found similar results in their study with growth mixture modeling. Of the nonlinear subgrouping approaches, the random forest algorithm produced the best results. This algorithm detected that whether or not the patient worked and where they were born were the 2 most important variables in classifying the data.

5 | DISCUSSION

Unlike most of the existing methods referred to in Section 1, our proposed procedure offers a complete process for subgrouping and validation; it uses a random effects linear model to assess the treatment effects over time for each subject, builds prediction models based on classification algorithms, and determines whether or not the subgroups are appropriate and beneficial. This whole process can be easily implemented using existing packages in statistical software such as R and SAS. To secure good performance for subgroup identification, repeated measures within the subject are required to separate variance components and identify individual treatment effects successfully.[22]

With the numerical studies, all classification methods performed about the same with the linear random slopes; however, for the nonlinear random slopes, the linear classification approaches performed poorly. This could lead to the recommendation of always using nonlinear classification approaches as the preferred method for subgrouping. While the results with such a nonlinear approach would still be good, the interpretation would not be as easy. As such, it is recommended that various classification approaches are considered to find the best subgrouping strategy. Real data analysis also confirms the importance of performing multiple subgrouping approaches and comparing the results. Our analysis showed that the nonlinear approaches were best but also showed that not all these nonlinear approaches perform well. In fact, only the random forest algorithm produced significant results here. Moreover, the simulation results indicate that our validation approach is not only simple but also powerful. The validation approach produced significant results more often when the proper type of classification approach was used, while also having a type I error close to a nominal level under the null hypothesis.

ORCID

Hyunkeun Cho http://orcid.org/0000-0002-6735-2487

REFERENCES

1. Miranda J, Chung JY, Green BL, et al. Treating depression in predominantly low-income young minority women: a randomized controlled trial. *J Am Med Assoc*. 2003;290:57-65.
2. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986;73:12-22.
3. Qu A, Lindsay BG, Li B. Improving generalised estimating equations using quadratic inference functions. *Biometrika*. 2000;87:823-836.
4. Song X, Pepe MS. Evaluating markers for selecting a patient's treatment. *Biometrics*. 2004;60:874-883.
5. Bonetti M, Gelber RD. Patterns of treatment effects in subsets of patients in clinical trials. *Biostatistics*. 2004;5:465-481.
6. Moskowitz CS, Pepe MS. Quantifying and comparing the predictive accuracy of continuous prognostic factors for binary outcomes. *Biostatistics*. 2004;5:113-127.
7. Cai T, Tian L, Wong PH, Wei LJ. Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics*. 2011;12:270-282.
8. Zhao L, Tian L, Cai T, Claggett B, Wei LJ. Effectively selecting a target population for a future comparative study. *J Am Stat Assoc*. 2011;108:527-539.
9. Foster JC, Taylor JMC, Ruberg SJ. Subgroup identification from randomized clinical trial data. *Stat Med*. 2011;30:2867-2880.
10. Diaz FJ, Yeh HW, Leon J. Role of statistical random-effects linear models in personalized medicine. *Curr Pharmacogenomics Person Med*. 2012;10:22-32.
11. Diaz FJ, de Leon J. The mathematics of drug dose individualization should be built with random effects linear models. *Ther Drug Monit*. 2013;35:276-277.

12. Diaz FJ, Rivera TE, Josiassen RC, Leon J. Individualizing drug dosage by using a random intercept model. *Stat Med*. 2007;26:2052-2073.
13. Diaz FJ. Measuring the individual benefit of a medical or behavioral treatment using generalized linear mixed-effects models. *Stat Med*. 2016;35:4077-4092.
14. Cho H, Wang P, Qu A. Personalized treatment for longitudinal data using unspecified random-effects model. *Stat Sin*. 2017;27:187-205.
15. Zhu X, Qu A. Individualizing drug dosage with longitudinal data. *Stat Med*. 2016;35:4474-4488.
16. Shen J, He X. Inference for subgroup analysis with a structured logistic-normal mixture model. *J Am Stat Assoc*. 2015;110:303-312.
17. Wu RF, Zheng M, Yu W. Subgroup analysis with time-to-event data under a logistic-cox mixture model. *Scand J Stat*. 2016;43:863-878.
18. Senn S. Individual therapy: New dawn or false dawn. *Drug Inf J*. 2001;35:1479-1494.
19. Hartley HO, Rao JNK. Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika*. 1967;54:93-108.
20. Harville DA. Maximum likelihood approaches to variance component estimation and to related problems. *J Am Stat Assoc*. 1977;72:48-53.
21. Siddique J, Chung JY, Brown CH, Miranda J. Comparative effectiveness of medication versus cognitive-behavioral therapy in a randomized controlled trial of low-income young minority women with depression. *J Consult Clin Psych*. 2012;80:995-1006.
22. Senn S. Mastering variation: Variance components and personalised medicine. *Stat Med*. 2016;35:966-977.

How to cite this article: Andrews N, Cho H. Validating effectiveness of subgroup identification for longitudinal data. *Statistics in Medicine*. 2018;37:98–106. https://doi.org/10.1002/sim.7500