Journal of Statistical Research *https://doi.org/10.47302/jsr.2020540204*
2020, Vol. 54, No. 2, pp. 147-165 *ISSN 0256 - 422 X*

SUBGROUP IDENTIFICATION FOR DIFFERENTIAL CARDIO-RESPIRATORY FITNESS EFFECT ON CARDIOVASCULAR DISEASE RISK FACTORS: A MODEL-BASED RECURSIVE PARTITIONING APPROACH

MD YASIN ALI PARH

Department of Mathematical Sciences, Ball State University, Muncie, IN-47306, USA
Department of Statistics, Islamic University, Kushtia-7003, Bangladesh
Email: yasinstatiu@gmail.com

MUNNI BEGUM*

Department of Mathematical Sciences, Ball State University, Muncie, IN-47306, USA
Email: mbegum@bsu.edu

MATTHEW HARBER, BRADLEY S. FLEENOR, MITCHELL WHALEY

Clinical Exercise Physiology Program, Ball State University, Muncie, IN-47306, USA
Email: mharber@bsu.edu

W HOLMES FINCH

Department of Educational Psychology, Ball State University, Muncie, IN-47306, USA
Email: whfinch@bsu.edu

JAMES PETERMAN, LEONARD KAMINSKY

Fisher Institute of Health and Well-being, Ball State University, Muncie, IN-47306, USA

* Corresponding author

SUMMARY

The goal of this study is twofold: i) identification of features associated with three cardiovascular disease (CVD) risk factors, and (ii) identification of subgroups with differential treatment effects. Multivariate analysis is performed to identify the features associated with the CVD risk factors: hypertension, diabetes, and dyslipidemia. For subgroup identification, we applied model-based recursive partitioning approach. This method fits a local model in each subgroup of the population rather than fitting one global model for the whole population. The method starts with a model for the overall effect of treatment and checks whether this effect is equally applicable for all individuals under the study based on parameter instability of M fluctuation test over a set of partitioning variables. The procedure produces a segmented model with a differential effect of cardio-respiratory fitness (CRF) corresponding to each subgroup. The subgroups are linked to predictive factors learned by the recursive partitioning approach. This approach is applied to the data from the Ball State Adult Fitness Program Longitudinal Lifestyle Study (BALL ST), where we considered the level of CRF as a treatment variable. The overall results indicate that CRF is inversely associated with hypertension, diabetes and dyslipidemia. The partitioning factors that are selected are related to these risk factors. The subgroup-specific results indicate that for each subgroup, the chance of hypertension, diabetes and dyslipidemia increases with low CRF.

Keywords and phrases: Subgroup identification, multivariate analysis, model-based recursive partitioning approach, cardio-respiratory fitness

1 Introduction

Abundant evidence over the past three decades has established that cardio-respiratory fitness (CRF) is directly associated with proper function of human body, and its respiratory, cardiovascular, and musculoskeletal systems (Ross et al., 2016). Recent studies have observed that CRF is inversely associated with non-communicable diseases including cardiovascular diseases (CVD) (Harber et al., 2017; Arena et al., 2015; Blair, 2009; Kokkinos et al., 1995). Chase et al. (2009) found that physical activity and CRF are associated with a lower risk of developing hypertension. In a similar study of physical capacity, Agostinis-Sobrinho et al. (2018) showed that there is a significant inverse association between CRF and blood pressure. To investigate the relationship between CRF and diabetes, Sawada et al. (2003) found an inverse association between type 2 diabetes risk and CRF based on a prospective study of Japanese men. Breneman et al. (2016) found that high CRF at baseline and maintenance of CRF over time is protective against the development of atherogenic dyslipidemia.

Considerable effort has been devoted to the estimation of overall average effect of CRF on diseases. Despite this focus, in many cases, the efficacy of CRF might vary based on individual's lifestyle, physiological, and demographic characteristics. Thus the effect of CRF on a disease may be different than the estimated average effect of CRF. A differential treatment effect on groups of individuals can be studied as subgroup analysis and is important to identify such groups for gaining precise insight about their health condition. This study focuses on differential effects of CRF on hypertension, diabetes and dyslipidemia separately that leads to subgroup identification (Ciampi et al., 1995; Foster et al., 2011; Lipkovich et al., 2011) problem for these health conditions. In addition, we identify which characteristics, if any, lead to these differential effects (Seibold et al., 2016).

Subgroup analysis often plays an essential role in clinical trials to investigate consistency or heterogene-

ity of treatment effects across subgroups Alosh et al. (2015). A number of subgroup identification methods are available in the literature including Virtual Twins (VT) Foster et al. (2011), subgroup identification with enhanced treatment effects based on differential effect search (SIDES) Lipkovich et al. (2011), QUalitative INteraction Trees (QUINT) Dusseldorp and Mechelen (2014) and the Generalized Unbiased Interaction Detection and Estimation (GUIDE) (Loh (2002), Loh (2009),Loh et al. (2015)). These methods are either limited to specific outcome variables, feature selection bias, or computational constraints. Model based recursive partitioning Seibold et al. (2016) approach automatically detect patient subgroups that are identifiable by predictive factors. In this method, the partial score with respect to the intercept does not give any information about whether the partitioning variable is predictive or prognostic. In order to state whether a partitioning variable is predictive or prognostic, it is necessary to consider the model parameters in the segmented model. If the treatment parameter varies in the subgroups, then the chosen partitioning variables are predictive or both predictive and prognostic. If the treatment parameter is constant, the variables are only prognostic. Subgroup analysis has also been studied under Bayesian and decision theoretic point of view (Schnell et al. (2017), Nugent et al. (2019)).

In this study, we applied the model based recursive partitioning approach to identify subgroup of patients with hypertension, diabetes and dyslipidemia (HDD), for which CRF is effective. Data from the BALL ST Adult Fitness Program Longitudinal Lifestyle Study is considered and is discussed in section 2. Our objective is to detect subgroups of individuals suffering from HDD in which the subgroups differ in terms of the intercept and effect of CRF. Before considering to subgroup identification, we identified features that are associated with HDD. First, we considered a bi-variate analysis using chi-squared test to investigate whether CRF is associated HDD. Second, we considered multivariate analysis to identify the features of HDD with generalized linear models. Finally, we applied the model-based recursive partitioning approach (Seibold et al., 2016; Zeileis et al., 2008) to the generalized linear models for subgroup identification. The method allows us to focus attention on predictive factors/ partitioning variables, and fit a segmented model that includes CRF $\times$ covariate interactions that describe the relevant subgroups. Thus the objectives of this paper are: to identify features associated with HDD; to identify subgroups of individuals that differ in terms of the intercept and effect of CRF; and finally to compare the results of overall intercept and effect of CRF with subgroup-specific intercept and CRF effect for each scenario.

The rest of the sections of this paper is organized as follows. In section 2, we discuss our data and variables. In section 3, we describe the methodology that applied to our data to identify features and to identify subgroups. In section 4, we present the results of exploratory data analysis, multivariate, and subgroup analysis. Finally, in section 5, we present a discussion of our results and conclusions about our study.

2 Data and Variables

Data from the BALL ST Adult Fitness Program Longitudinal Lifestyle Study is used for the subgroup identification problem. All participants in this study performed an initial comprehensive health and physical fitness assessment between 1969 and 2017, including a maximal Cardiopulmonary Exercise (CPX) test. The baseline CPX test was performed using standardized treadmill protocols BRUCE et al. (1963), Ball State University Bruce Ramp Kaminsky and Whaley (1998), modified Balke-Ware Pollock et al. (1982), and individualized protocols to determine VO2peak. The VO2peak was determined by averaging the highest 2 to 3 consecutive measured VO2 values within 2 $ml\ kg^{-1}\ min^{-1}$, occurring in the last 2 min of the CPX test. In this pa-

per, we categorized VO2peak into low CRF group ($\leq 33rd$ percentile) (Imboden et al., 2018) and high CRF group ($> 33rd$ percentile) using percentiles from database of Fitness Registry and the Importance of Exercise (FRIEND) Kaminsky et al. (2015). For the US population, the FRIEND registry provides age-specific and sex-specific reference values for CPX-CRF.

In the original dataset, there were 3694 de-identified participants with 58 covariates. The covariates include demographic variables, physiological variables, and variables associated with lifestyle. Because of the large number of missing values, we excluded LDL, HDL, the proportion of body fat, waist circumference, hip waist ratio, peak systolic, and diastolic blood pressure from our dataset. The percent of missing values in these variables was $46\%, 17\%, 28\%, 11\%, 14\%, 34\%, 33\%$ respectively. The demographic variable, ethnicity, is also excluded from the dataset because 99.35% of participants belong to one race, white non-Hispanic. We also omitted a few participants because of their incomplete information. Some variables including record date, test number, and death date that are not relevant to our study are excluded from the dataset.

Three health conditions HDD with their binary status are considered as three independent response variables. These are major health conditions that affect public health substantially and also the major causes of death. The CRF with two levels (CRF low and CRF high) considered as treatment variables. Based on the available covariates in our dataset we considered age, BMI, glucose, triglyceride, cholesterol, sex, physical activity level, obesity, smoking status as explanatory variables.

3 Methodology

3.1 Identification of features associated with hypertension, diabetes, and dyslipidemia

To examine the association between CRF and the health conditions: hypertension, diabetes, and dyslipidemia, we applied the bi-variate Pearson chi-square test. The null hypothesis states that hypertension, diabetes, and dyslipidemia are not associated with CRF, while the research hypothesis states that they are associated with CRF. To investigate the relationship between a set of predictors and a response variable we considered the multivariate models. Since each of the response variables, hypertension, diabetes and dyslipidemia, has two levels: presence and absence, we consider a binary logistic regression model to investigate the relationship between predictors and the response variable.

3.2 Model-based recursive partitioning approach for subgroup identification for hypertension, diabetes, and dyslipidemia

The regular fit of a multivariate model gives the overall average effect of the covariate for all individuals. However, the overall average response of a treatment is not necessarily the same for all individuals. A treatment may be more effective for a particular group of people and less effective for another group of people. Especially in the presence of subgroups the assumption of universal effect of a treatment to all individuals does not hold. In such a situation, we can consider the existence of subgroups of individuals (Ciampi et al., 1995; Foster et al., 2011; Lipkovich et al., 2011). To identify subgroups for binary responses, we applied the logistic regression-based recursive partitioning approach (Zeileis et al., 2008; Seibold et al., 2016).

Consider, the logistic regression model $M((Y, \mathbf{X}, Z), \boldsymbol{\Theta})$ that describes the conditional distribution of Y as a function of the treatment (Z) and all possible covariates $(\mathbf{X})$ through a vector of parameters $\boldsymbol{\Theta}$. The parameter vector $\boldsymbol{\Theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})^T$ typically contains one intercept parameter α, a set of parameters for covariates $\boldsymbol{\beta}$ and one treatment parameter γ. Given n observations Y_i $(i = 1, 2, \ldots, n)$, the estimate of the parameter vector $\boldsymbol{\Theta}$ can be obtained by applying the maximum likelihood (ML) estimation technique. The estimate of the parameter vector is obtained by taking partial derivatives with respect to the corresponding parameters of the objective function $\boldsymbol{\Psi}(\boldsymbol{Y_i}, \boldsymbol{\Theta})$, which is equivalent to solving the score equation

$$\sum_{i=1}^{n} \frac{\partial \boldsymbol{\Psi}((y, \mathbf{x}, \mathbf{z})_i, \boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}} = \sum_{i=1}^{n} \psi((y, \boldsymbol{x}, \boldsymbol{z})_i, \boldsymbol{\Theta}) = 0, \tag{3.1}$$

where ψ is the score function.

These estimates give the overall average response for all individuals $i = 1, 2, \ldots, n$. However, in the presence of patient subgroups that differ in their treatment effect γ, the mean treatment effect $\hat{\gamma}$ does not consider the positive or negative effect of a specific subgroup. Also, the treatment effect might depend on additional characteristics such as age, gender and other lifestyle factors of the patients. So, it might be possible to partition the observations with respect to some covariates such that it can be possible to fit a model at each subgroup of the patients (Seibold et al., 2016).

In such a situation, we use a recursive partitioning approach based on l partitioning variables X_j $(j = 1, 2, \ldots, l)$. The patient subgroups can be described as a partition $\{\mathcal{B}_b\}, (b = 1, 2, \ldots, B)$ of all patients $i = 1, 2, \ldots, n$ (Zeileis et al., 2008). The parameters of each subgroup should be different, which can be defined as varying coefficients (Hastie and Tibshirani, 1993). Let the subgroup-specific model parameters be denoted as $\boldsymbol{\Theta(b)}$. The coefficient varies based on the several patients characteristics or predictive factors and are always step functions with different levels for each subgroup.

In this paper, we are looking for the covariates that interact with the treatment variable (CRF) and the variables that have a direct effect on the outcome. So, we are interested in subgroups that differ in the intercept or the treatment effect or both. We assume that the effects of covariates are constant for all patient, so the subgroup specific parameter vector is $\boldsymbol{\Theta(b)} = (\boldsymbol{\alpha}(b), \boldsymbol{\beta}, \boldsymbol{\gamma}(b))^T$. Subgroup-specific intercept $\boldsymbol{\alpha}(b)$ and treatment parameter $\boldsymbol{\gamma}(b)$ are influenced by the additional patient characteristics that also treated as the partitioning variables to partition the sample space $\mathcal{X}$. Partitioning variables are the predictive variables that interact with treatment variables.

If the partition $\{\mathcal{B}_b\}$ is known, the partitioned model parameters $\boldsymbol{\Theta(b)}$ could be estimated by minimizing the segmented objective function:

$$\hat{\boldsymbol{\Theta}}_b = \underset{\boldsymbol{\Theta}_b}{argmin} \sum_{i=1}^{n} \sum_{b=1}^{B} \mathbf{1}(\boldsymbol{x_i} \in \mathcal{B}_b) \Psi(y, \boldsymbol{z})_i, \Theta_b, \tag{3.2}$$

where $\mathbf{1}$ denotes the indicator function and $(y, \boldsymbol{z})_i, \boldsymbol{x_i}$ are the realizations of $(Y, \boldsymbol{Z})$ and $\boldsymbol{X}$ for the ith patient. This allows us to write the subgroup-specific intercept and treatment parameters as a function of the partitioning variables.

$$\boldsymbol{\alpha}(\boldsymbol{z}) = \sum_{b=1}^{B} \mathbf{1}(\boldsymbol{z} \in \mathcal{B}_b) \boldsymbol{\alpha}(b) \; and \; \boldsymbol{\beta}(\boldsymbol{z}) = \sum_{b=1}^{B} \mathbf{1}(\boldsymbol{z} \in \mathcal{B}_b) \boldsymbol{\beta}(b).$$

However, without any prior knowledge about the partitioning $\mathcal{B}_b$, minimization of the objective function is more complicated, even if the number of the segments B is fixed. If there is more than one partitioning variable ($l > 1$), the number of potential partitions quickly becomes too large for an exhaustive search (Zeileis et al., 2008). To control the large number of partitions, we used the Bayesian information criteria (BIC) to prune the tree to avoid overfitting by increasing B (Su et al., 2004).

3.2.1 Recursive partitioning algorithm

The key idea underlying this method is the ability to detect parameter instability by looking at the score function of the model. To check the parameter instability, the basic idea is that each node is associated with a single model. To assess whether the splitting of the node is necessary, a fluctuation test for parameter instability is performed (Zeileis and Hornik, 2004). If there is significant instability with respect to any of the partitioning variables X_j, split the node into B locally optimal segments and repeat the procedure. If no more significant instabilities is found, the recursion stops and returns a tree where each terminal node is associated with a model $M(Y, \boldsymbol{\Theta}_b)$ (Zeileis et al., 2008), The steps of the algorithm are,

1. Fit the multivariate model once to all observations in the current node of the tree by estimating parameters of the objective function via Maximum likelihood estimation technique.
2. Assess whether the parameter estimates are stable with respect to every ordering covariates, $X_1, X_2, \ldots, X_l$. If there is some overall instability, select the variable X_j associated with the highest parameter instability, otherwise stop.
3. Compute the split point(s) that locally optimize the objective function, either for a fixed or adaptively chosen number of splits.
4. Split the node into daughter nodes and repeat the procedure.

3.2.2 Testing the parameter instability

The algorithm to check the parameter instability is to find out whether the parameters of the fitted model are stable over each particular ordering implied by the partitioning variables X_j, or whether splitting the sample with respect to one of the X_j might capture instabilities in the parameters and thus improve the fit. Since we are only interested in detecting non-constant intercepts $\boldsymbol{\alpha}(\boldsymbol{X})$ and treatment effects $\gamma(\boldsymbol{X})$ we focus on the partial score functions $\psi_\alpha((Y, \boldsymbol{X}), \boldsymbol{\Theta}) = \partial \boldsymbol{\Psi}((Y, \boldsymbol{X}), \boldsymbol{\Theta})/\partial \boldsymbol{\alpha}$ and $\psi_\gamma((Y, \boldsymbol{X}), \boldsymbol{\Theta}) = \partial \boldsymbol{\Psi}((Y, \boldsymbol{X}), \boldsymbol{\Theta})/\partial \gamma$ (Seibold et al., 2016). If the model parameters are in fact constant and do not depend on any of the partitioning variables $\boldsymbol{X}$, the partial score function $\psi_\alpha((Y, \boldsymbol{X}), \boldsymbol{\Theta})$ and $\psi_\gamma((Y, \boldsymbol{X}), \boldsymbol{\Theta})$ are independent of $\boldsymbol{X}$. Consequently, parameter instability corresponds to a correlation between either of the partial score functions and at least one of the partitioning variables $X_1, X_2, \ldots, X_l$. Formally, to detect deviations from independence between the partial score functions and the partitioning variables, model-based recursive partitioning utilizes independent tests. The null hypotheses are

$$H_0^{\alpha,j} : \psi_\alpha((Y, \boldsymbol{X}), \boldsymbol{\Theta}) \perp X_j, j = 1, 2, \ldots, J$$
$$H_0^{\gamma,j} : \psi_\gamma((Y, \boldsymbol{X}), \boldsymbol{\Theta}) \perp X_j, j = 1, 2, \ldots, J.$$

For the model $M((Y, \boldsymbol{X}), \boldsymbol{\Theta})$, the partial score functions with respect to $\boldsymbol{\alpha}$ and γ, are independent of the partitioning variable X_j ($j = 1, 2, \ldots, J$). Hence, these null hypotheses correspond to an appropriate

model fit regarding the intercept and treatment parameter. Because the partial score functions under the null-hypotheses are at least asymptotically normal in many model families, asymptotic M-fluctuation tests with appropriate correction for multiplicity were introduced by Zeileis et al. (2008) and Zeileis and Hornik (2004)). Alternatively, permutation tests can be applied in situations where asymptotic normality of the partial score is not guaranteed Zeileis and Hothorn (2013).

If at least one of the $2 \times J$ null hypotheses for the global model $M((Y, \mathbf{X}, Z), \mathbf{\Theta})$ is rejected at a pre-specified nominal level, model-based recursive partitioning selects the partitioning variable X_j^* associated with the highest correlation to any of the partial score functions. This is usually done by means of the smallest p value. The dependency structure between the partitioning variableX_j^* and either one of the partial score functions is described by a simple cut-point model. Once an optimal cut point $X_j^* < \mu$ using a suitable criterion is found (Zeileis et al., 2008; Hothorn et al., 2006b), the subjects are split into two subgroups according to $X_j^* < \mu$. For both subgroups, two separate models are estimated with parameters $\hat{\mathbf{\Theta}}_1$ and $\hat{\mathbf{\Theta}}_2$, respectively, followed by testing independence of hypotheses with the corresponding partial score functions. If deviations from independence is found, a cut-point is selected according to the most highly associated partitioning variable and split again. The procedure of testing independence of partial score functions and partitioning variables is repeated recursively until deviations from independence can no longer be detected. The R codes for the subgroup analysis can be found in https://github.com/mbegum/GitHubManuscript/R-code-subgroup analysis.txt.

4 Findings

4.1 Results of Exploratory Data Analysis

We summarize the characteristics of the study participants with descriptive statistics and graphical representation. Since CRF is expressed in percentile, we present the median and interquartile range (IQR) of fitness rank according to the level of categorical variables in Table 1. The summary statistics of the quantitative variables according to the levels of CRF are shown in Table 2.

Table 1 shows that the median CRF for a female is higher than a male. Obese people, smokers, and physically inactive individuals have CRF rank compared to their counterparts. In addition, the CRF of the people having hypertension, diabetes, dyslipidemia, is lower.

Table 2 and Figure 1 show that the CRF is lower among participants having a higher amount of triglyceride, glucose, and cholesterol. The average age is similar for participants with both high and low CRF.

The results of bi-variate analysis in Table 3 indicate that there is a strong association between CRF and hypertension, diabetes, and dyslipidemia.

4.2 Identification of features associated and identification of subgroups with different CRF effect

Identification of features using multivariate analysis and the subgroup analysis are conducted for hypertension, diabetes and dyslipidemia separately.

In order to determine the predictive ability of the covariates considered in this study, we constructed receiving operating characteristics (ROC) curves and calculated the area under the curves (AUC) from a validation data for each of the three health conditions.

Table 1: Median and IQR of the fitness rank according to the levels of categorical variables

Categorical Variable	Levels	n	Median	IQR
Sex	Female	1502	46	42
	Male	1701	39	45
Obesity	Absence	1100	55	43
	Presence	2103	22	28
Smoking Status	Smoker	349	30	37
	Non-smoker	2854	44	43
Physical activity level	Inactive	2319	34	39
	Active	884	66	43
Diabetes	Absence	3014	44	44
	Presence	189	19	31
Hypertension	Absence	2219	48	44
	Presence	984	30	41
Dyslipidemia	Absence	1980	50	36
	Presence	1523	34	41

Figure 2 shows that the covariates have high predictive ability for hypertension and diabetes with AUC 0.82 and 0.92 respectively, and acceptable predictive ability for dyslipidemia with AUC 0.76. It is to be noted that these covariates are standard in the CRF literature.

Before applying the model based methods to the entire dataset, we divided the dataset into training with approximately two-third of the observations and test with the rest. For hypertension, subgroups in the training data were identified based on BMI and Age whereas in the test data, only age was informative to subgroup identification. For diabetes, subgroups in both training and test data are identified based on a single predictor, Glucose that defines the health condition. Finally for dyslipidemia, subgroups in the training data are identified based on total cholesterol, sex, and triglyceride whereas subgroups in the test data are identified based on total cholesterol and sex. These results indicate that other than Glucose, none of the covariates are informative to subgroup identification for this data. The subgroup analysis results from training and test data are attached in the supplementary information.

Next we applied the model-based recursive partitioning approach to the entire dataset. For each of the three health condition, hypertension, diabetes, and dyslipidemia the results from the feature identification and subgroup analysis are presented below in that order.

Hypertension: There are 984 (30.72%) individuals with hypertension in the dataset, and the rest 2219 (69.28%) do not have hypertension. Considering the status of hypertension as a response variable with two categories, presence and absence, we identified features associated with hypertension and the impact of CRF

Table 2: Mean and standard deviation of the quantitative variable by CRF

		CRF High		CRF Low	
Quantitative variables	Unit	Mean	SD	Mean	SD
Age	years	45.18	12.37	44.37	12.00
Weight	kg	74.64	14.09	92.54	21.22
BMI	kg/m^2	25.27	3.85	31.05	6.49
Glucose	mg/dl	94.45	16.10	101.02	27.29
Cholesterol	mg/dl	200.65	39.28	200.65	39.28
Triglyceride	mg/dl	115.92	69.14	167.08	128.47

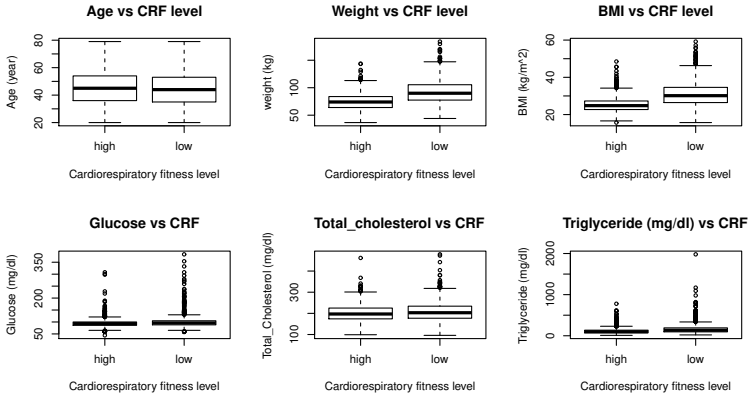

Figure 1: Box plot of continuous explanatory variables by CRF

on hypertension by applying the binary logistic regression model.

Table 4 shows the estimates of the odds ratios with corresponding 95% confidence intervals. The 95% confidence intervals for the odds ratios for CRF, age, sex, BMI, and dyslipidemia do not include one, indicating that each of these covariates are associated with the risk of hypertension. The estimates of regression coefficients (not shown here) of the continuous covariates, age, BMI, triglyceride, glucose and cholesterol, are positive, indicating that an increase in these covariates is associated with an increase in the risk of hypertension. The estimated coefficient for the categorical variable, sex, is positive, indicating that men are more prone to hypertension than women. Similarly, the estimate of the CRF is positive suggesting that the risk of hypertension of low CRF people is higher than that of high CRF people.

Subgroup identification: Table 5 reports the test statistics values and the p values of parameter instability based on M-fluctuation tests, and Figure 3 presents the logistic regression-based tree with different intercept and differential impact of CRF in each node, where the partitioning variables, age, triglyceride and BMI, interact

Table 3: Chi-square test between the levels of CRF and different diseases

Variables	Test Statistic	Degree of Freedom	P-value
Hypertension	96.603	1	$< .0001$
Dyslipidemia	108.05	1	$< .0001$
Diabetes	66.285	1	$< .0001$

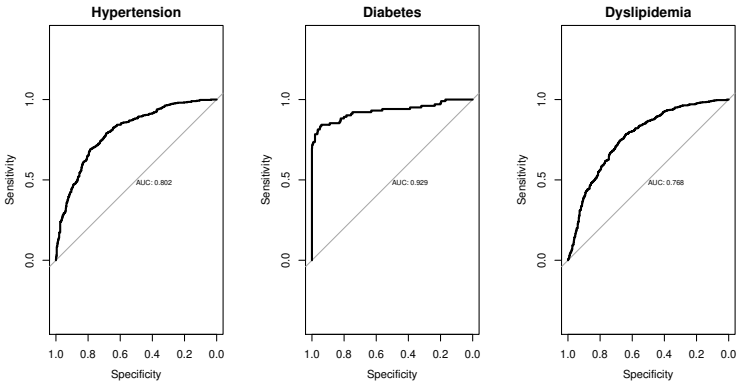

Figure 2: ROC curves for three health conditions: hypertension, diabetes, and dyslipidemia

with CRF.

The results show that age has the smallest p value and is highly significant and thus used for splitting the data in the first node. In the second node, triglyceride has the lowest p value and is highly significant and thus used for splitting the node again. Similarly, fifth node is split based on BMI. No further considerable parameter instabilities are detected in the third, fourth, sixth, and seventh nodes and hence partitioning stops in those nodes.

It is observed that the most significant predictors in multivariate model is age (highest Z-value) that is also statistically significant predictive factor in the M-fluctuation test and is used to split the data at the very beginning node.

According to the logistic regression-based recursive partitioning tree in Figure 3, the intercept and estimated effect of CRF with 95% confidence interval for each subgroup is reported in Table 6. It is to be noted that subgroups I and II are obtained with age and triglyceride as interacting covariates whereas subgroups III and IV are obtained with age and BMI as interacting covariates. The intercept of subgroup I indicates that individuals with triglyceride less than $108.0\ mg/dl$ and younger than 49 years have on average a low risk (-2.23) of hypertension compared to overall average risk (-1.20) of hypertension, however, this increases with low CRF since the effect of CRF (0.69) for subgroup I is positive. Thus, within this subgroup the risk of hypertension is inversely associated with CRF. In contrast, individuals with triglyceride greater than $108.0\ mg/dl$ and younger than 49 years have on average a higher risk (-1.29) of hypertension compared to subgroup I and this also increases with low CRF. In addition, individuals with body mass index higher than $27.1\ kgm^{-2}$ and older than

Table 4: Results of multivariate analysis for hypertension

		95% CI	
	Odds Ratio	lower	upper
CRF(low)	1.46	1.19	1.78
Age	1.06	1.06	1.07
Sex(male)	1.57	1.31	1.88
BMI	1.09	1.07	1.12
Obesity	1.03	0.8	1.34
Dyslipidemia	1.22	1.01	1.46
Glucose	1.00	0.99	1.01
Triglyceride	1.00	1.00	1.00
Cholesterol	0.99	0.99	1.00
Smoking status	0.85	0.64	1.12
Diabetes	1.36	0.86	2.14

49 years, have a high risk of hypertension that also increases with low CRF.

Diabetes: Out of a total of 3203 participants one hundred and eighty nine (5.9%) have diabetes. The status of diabetes is considered as a response variable with two categories, presence and absence of diabetes. The results of multivariate analysis in Table 7 show that Glucose associated to the risk of diabetes which is a general knowledge for diabetes. Age, sex, BMI, cholesterol, and smoking status are also associated with diabetes and are borderline significant. CRF not statistically significant for diabetes which may be due to the small number of diabetic patients in our data.

Subgroup identification: Similar to hypertension, we applied the logistic regression-based recursive partitioning approach to identify subgroups of individuals. Figure 4 presents the logistic regression-based tree with different intercept and differential impact of CRF in each node, where a single predictive factor Glucose interact with CRF. Similar to hypertension, the predictive factors are selected based on the smallest p value of the parameter instability of M-fluctuation test (we did not show the result of parameter instability here).

According to the logistic regression-based recursive partitioning plot in Figure 4, the intercept and estimated effect of CRF with 95% confidence intervals for each subgroup are reported in Table 8. If we compare results of each subgroup with all participants, we see that individuals with Glucose less than or equal 111 mg/dl have on average, a low risk of diabetes (-5.38) compared to overall average risk (-3.52) of diabetes; however, this increases with low CRF. In contrast, individuals with Glucose higher than 111 mg/dl have a higher risk of diabetes that also increases with low CRF.

Table 5: Results of parameter instability based on M-fluctuation test

	Node 1		Node 2		Node 5	
Variables	Test value	P value	Test value	P value	Test value	P value
Age	258.3	$< .0001$	61.2	$< .0001$	34.4	$< .0001$
Sex	37.3	$< .0001$	44.3	$< .0001$	5.0	0.61
BMI	130.8	$< .0001$	71.5	$< .0001$	38.7	$< .0001$
Triglyceride	114.9	$< .0001$	71.7	$< .0001$	24.9	$< .0001$
Obesity	109.5	$< .0001$	49.4	$< .0001$	38.1	$< .0001$
Glucose	108.7	$< .0001$	46.3	$< .0001$	38.4	$< .0001$
Cholesterol	30.7	$< .0001$	34.4	$< .0001$	4.9	0.94
Dyslipidemia	76.9	$< .0001$	38.2	$< .0001$	15.9	0.02
Inactivity	29.7	0.0001	2.6	1.0	11.1	0.16
Smoking	4.9	0.99	5.2	.997	38.9	0.99
Diabetes	58.1	$< .0001$	3.3	.999	18.1	0.03

Dyslipidemia: About forty eight percent of individuals have dyslipidemia in the dataset. Table 9 shows that CRF, age, sex, BMI, obesity, triglyceride, and cholesterol are statistically significant predictors for dyslipidemia and are associated with the risk of dyslipidemia.

Subgroup identification: Three partitioning variables, cholesterol, triglyceride, and sex, are selected to identify the subgroup based on the parameter instability of M-fluctuation tests (not shown here). These variables are also highly significant in multivariate analysis for dyslipidemia.

The results in Table 10 show that women with cholesterol less or equal $198.7\ mg/dl$ have, on average (-1.82), a low risk of dyslipidemia compared to the overall mean response (0.41) of all participants , however, this increases with low CRF. In contrast, men with the same amount of cholesterol have a relatively high risk of dyslipidemia than women. The intercept of subgroup IV indicates that individuals with more than $198.7\ mg/dl$ cholesterol and $103.2\ mg/dl$ triglyceride have an extremely high risk of dyslipidemia, and that increases with low CRF.

We compared the results from the model based recursive partitioning approach to that from virtual twin(VT) which also generates subgroups by selecting covariates that interacts with CRF. But unlike the model based approach, VT does not take into account of intercepts of each node. We compared the subgroups for hypertension obtained by VT and the subgroups obtained by the model based approach. After running a number of iterations we found that VT generates different subgroups selecting different covariates at each iteration, whereas, the subgroups from the model based approach stays the same. We attached subgroups using VT from two iterations to the supplementary information.

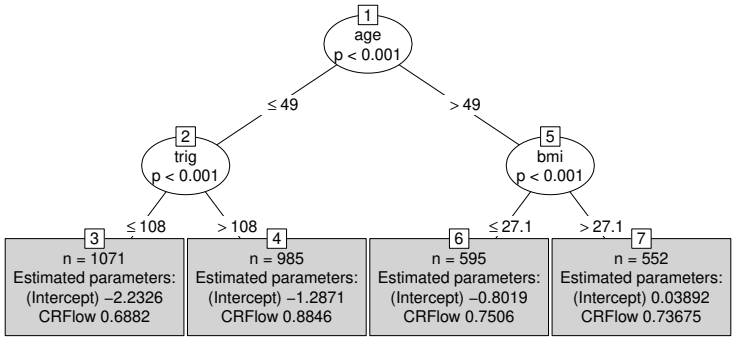

Figure 3: Logistic regression based tree for hypertension. The terminal nodes of the tree plot report the intercept and estimate of CRF.

Table 6: Intercept and estimate of the CRF with 95% confidence interval (CI) for each subgroup based on GLM tree

All participants/ Subgroup	Intercept[CI]	CRF(Low)[CI]
All participants	-1.20 [-1.30, -1.09]	0.88 [0.73, 1.03]
Subgroup I (age$<=49$& triglyceride $<=108$)	-2.23 [-2.48, -2.00]	0.69 [0.31, 1.06]
Subgroup II ((age$<=49$& triglyceride>108)	-1.29 [-1.52, -1.06]	0.88 [0.60, 1.17]
Subgroup III (age>49& BMI$<=27.1$)	-0.80 [-0.99, -0.61]	0.75 [0.34, 1.16]
Subgroup IV(age>49& BMI>27.1)	0.04 [-0.21, 0.28]	0.74 [0.39, 1.08]

5 Discussion and Conclusion

In this paper, we accomplished two broad objectives. First, we identified features associated hypertension, diabetes and dyslipidemia using the data from the BALL ST Adult Fitness Program Longitudinal Lifestyle Study. Second and most importantly, we identified subgroups of individuals with differential effects of CRF by applying logistic regression based recursive partitioning approach. This method detects the predictive factors that interact with the treatment (CRF). In the model-based recursive partitioning approach, the partitioning variable for each split was selected based on the smallest p value in the M-fluctuation test. After identifying subgroups based on the partitioning variables for each subgroup of hypertension, diabetes and dyslipidemia, we obtained the confidence intervals of the parameters (intercept and CRF).

The overall results from the multivariate analyses suggest that CRF is inversely associated with hypertension, diabetes, and dyslipidemia. That is, individuals with low CRF have a higher chance of getting hypertension, diabetes, and dyslipidemia than individuals with high CRF which has an important implication for the policymakers in public health. The results of this study are consistent with the previous literature (Chase et al.

Table 7: Results of multivariate analysis for diabetes

		95% Conf. Int	
	Odds Ratio	lower	upper
CRF(low)	1.51	0.8	2.86
Age	1.02	1	1.05
Sex(male)	0.53	0.31	0.93
BMI	1.05	1	1.11
Obesity	1.2	0.57	2.54
Dyslipidemia	1.01	0.57	1.8
Glucose	1.18	1.15	1.20
Triglyceride	1.00	0.99	1.00
Cholesterol	0.99	0.98	1.00
Hypertension	1.3	0.74	2.30
Smoking status	1.89	0.90	3.96

Table 8: Intercept and estimate of the CRF with 95% confidence interval(CI) for each subgroup based on GLM tree

All participants/ Subgroup	Intercept[CI]	CRF(Low)[CI]
All participants	-3.52 [-3.80, -3.26]	1.37 [1.05, 1.69]
Subgroup I (glucose$<=111$)	-5.38 [-6.22, -4.71]	1.34 [0.48, 2.29]
Subgroup II (glucose>111)	-0.57 [-0.95, -0.19]	0.64 [0.17, 1.12]

(2009),Agostinis-Sobrinho et al. (2018), Sawada et al. (2003), Breneman et al. (2016)). Our results also suggest that BMI and age are positively associated with hypertension, diabetes, and dyslipidemia. The chance of having hypertension and dyslipidemia for men is higher than women, but the chance of diabetes for women is higher than men. Individuals with a higher amount of triglyceride have a higher chance of having hypertension and dyslipidemia. Other than Glucose that defines diabetes, no other covariates appeared as significant factors for the health condition. A relatively small number of diabetic patients in the data may attribute to this result.

The results of subgroup analysis for hypertension based on the entire dataset indicate that, four subgroups are identified with different intercepts and differential effects of CRF based on partitioning variables, age, triglyceride and BMI. However, subgroups in the training data are based on age and BMI and subgroups in the test data are based on age only. A test data with larger number of observations will be helpful to properly validate the results.

Regarding diabetes, two subgroups are identified from the entire dataset with different intercepts and differ-

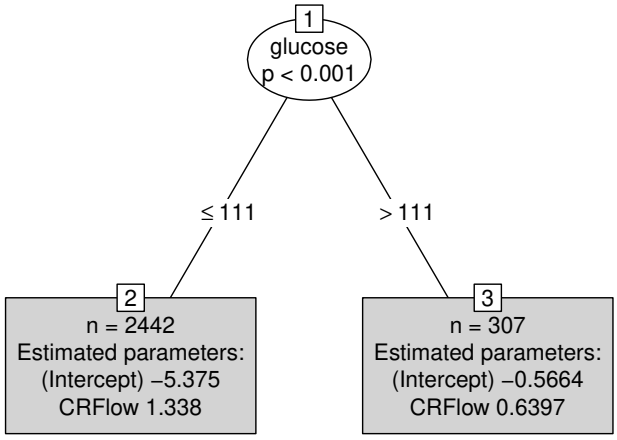

Figure 4: Logistic regression based tree for diabetes. The terminal nodes of the tree plot report the intercept and estimate of CRF.

ential effect of CRF based on Glucose only. Subgroups based on both training and test data are consistent with this result. Finally for dyslipidemia patients, four subgroups are identified with different intercepts and differential effects of CRF based on cholesterol, triglyceride, and sex. Subgroups in the training data are obtained based on the same covariates as in the entire dataset. However, triglyceride did not appear to be informative for subgroup identification for dyslipidemia in the test data. Again, a test data with larger number of observations will be helpful to properly validate the results.

In our analysis, we observed a connection between the most significant features in multivariate analysis and predictive factors in the first node of subgroup analysis. For example, in case of hypertension, age is the most significant feature which is also selected as the partitioning variable in the first node of subgroup analysis. Similarly, for diabetes and dyslipidemia, the most significant factors are glucose and cholesterol, respectively, and those are also selected as the partitioning variables in the first node of subgroup analysis. To the best of our knowledge, there is no literature on this connection, and thus we did not discuss any theory against this argument.

Limitations: One of the limitations of our study is a relatively small number of diabetic patients. Another limitation is that in our dataset, 99.35% of individuals belong to one race, white non-Hispanic. Therefore, we were not able to find how race or ethnicity would influence the subgroup analysis.

Table 9: Results of multivariate analysis for dyslipidemia

		95% Conf.Int	
	Odds Ratio	lower	upper
CRF(low)	1.33	1.11	1.6
Age	1.02	1.01	1.02
Sex(male)	2.24	1.91	2.63
BMI	1.03	1.01	1.06
Obesity	1.27	0.99	1.63
Glucose	0.99	0.99	1.00
Triglyceride	1.005	1.003	1.01
Cholesterol	1.01	1.01	1.01
Hypertension	1.18	0.98	1.42
Smoking status	0.88	0.68	1.14
Diabetes	1.30	0.83	2.04

Scope for further study: Further investigation and research are required to demonstrate the observed connection between the most significant feature in multivariate analysis and predictive factors in the first node of subgroup analysis. In addition, we assumed independence among three outcome variables and conducted univariate subgroup analysis. A joint subgroup analysis can be performed allowing dependence among these outcomes and is left as a future research.

Conclusion: In conclusions, CRF is inversely associated hypertension, diabetes, dyslipidemia. The chance of hypertension and dyslipidemia for men is higher than women. It is interesting to note that the partitioning variables selected in the subgroup analysis are the established risk factors in the literature. The subgroup-specific results indicate that for each subgroup, the risks of hypertension, diabetes, and dyslipidemia increase with low CRF. Our study suggests that improvement of fitness level through cardiopulmonary exercise is essential to control hypertension, diabetes, dyslipidemia: three important risk factors of CVD. In addition, it is important to maintain the normal level of physiological factors such as, BMI, cholesterol, glucose, triglyceride and blood pressure.

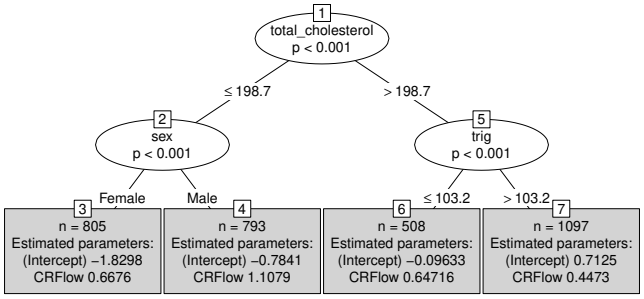

Figure 5: Logistic regression based tree for dyslipidemia. The terminal nodes of the tree plot report the intercept and estimate of CRF.

Table 10: Intercept and estimate of the CRF with 95% confidence interval(CI) for each subgroup based on GLM tree

All participants/ Subgroup	Intercept[CI]	CRF(Low)[CI]
All participants	0.41 [-0.50, - 0.32]	0.77 [0.63, 0.92]
Subgroup I (Cholesterol$<=$ 198.7& Sex= "female")	-1.82 [-2.08, -1.59]	0.67 [0.29, 1.03]
Subgroup II (Cholesterol$<=$ 198.7& Sex= "male")	-0.78 [-0.98, -0.59]	1.11 [0.81, 1.40]
Subgroup III (Cholesterol$>$ 198.7& Triglyceride $<=$ 103.2)	-0.09 [-0.29, 0.11]	0.65 [0.24, 1.06]
Subgroup IV (Cholesterol$>$ 198.7& Triglyceride $>$ 103.2)	0.71 [0.54, 0.89]	0.44 [0.18, 0.71]

References

Agostinis-Sobrinho, C., Ruiz, J. R., Moreira, C., Abreu, S., Lopes, L., Oliveira-Santos, J., Mota, J., and Santos, R. (2018), "Cardiorespiratory Fitness and Blood Pressure: A Longitudinal Analysis," *Journal of Pediatrics*, 192, 130 – 135.

Alosh, M., Fritsch, K., Huque, M., Mahjoob, K., Pennello, G., Rothmann, M., Russek-Cohen, E., Smith, F., Wilson, S., and Yue, L. (2015), "Statistical Considerations on Subgroup Analysis in Clinical Trials," *Statistics in Biopharmaceutical Research*, 7, 286–303.

Arena, R., Guazzi, M., Lianov, L., Berra, K., Lavie, C. J., Kaminsky, L., Williams, M., Hivert, M.-F., Franklin, N. C., Myers, J., Dengel, D., Lloyd-Jones, D., Pinto, F. J., Cosentino, F., Halle, M., Gielen, S., Dendale, P., Josef, N., Pelliccia, A., Giannuzzi, P., Corra, U., Piepoli, F. P., Guthrie, G., and Shurney, D. (2015), "Healthy Lifestyle Interventions to Combat Noncommunicable Disease—A Novel Nonhierarchical Connectivity Model for Key Stakeholders: A Policy Statement From the American Heart Association, European Society of Cardiology, European Association for Cardiovascular Prevention and Rehabilitation, and American College of Preventive Medicine," *Mayo Clinic Proceedings*, 90, 1082 – 1103.

Blair, S. N. (2009), "Physical inactivity: the biggest public health problem of the 21st century," *British Journal of Sports Medicine*, 43, 1 – 2.

Breneman, C. B., Polinski, K., Sarzynski, M. A., Lavie, C. J., Kokkinos, P. F., Ahmed, A., and Sui, X. (2016), "The Impact of Cardiorespiratory Fitness Levels on the Risk of Developing Atherogenic Dyslipidemia," *The American Journal of Medicine*, 129, 1060–1066.

BRUCE, R. A., BLACKMON, J. R., JONES, J. W., and STRAIT, G. (1963), "Exercising testing in adult normal subjects and cardiac patients," *Pediatrics*, 32, 742–756.

Chase, N. L., Sui, X., C., L. D., and Blair, S. N. (2009), "The Association of Cardiorespiratory Fitness and Physical Activity With Incidence of Hypertension in Men," *American Journal of Hypertension*, 22, 417 – 424.

Ciampi, A., Negassa, A., and Lou, Z. (1995), "Tree-structured prediction for censored survival data and the Cox model," *Journal of Clinical Epidemiology*, 48, 675–689.

Dusseldorp, E. and Mechelen, I. (2014), "Qualitative interaction trees: a tool to identify qualitative treatment-subgroup interactions," *Statistics in Medicine*, 33, 219–237.

Foster, J. C., Taylor, J. M. G., and Ruberg, S. J. (2011), "Subgroup identification from randomized clinical trial data, Statistics in Medicine," *Statistics in Medicine*, 30, 2867–2880.

Harber, M., Kaminsky, L., Arena, R., Blair, S., Franklin, B., Myers, J., and Ross, R. (2017), "Impact of Cardiorespiratory Fitness on All-Cause and Disease-Specific Mortality: Advances Since 2009." *Prog Cardiovasc Dis.*, 60(1), 11–20.

Hastie, T. and Tibshirani, R. (1993), "Varying-Coefficient Models," *Journal of Royal Statistical Society Series B (Methodological)*, 55, 757–796.

Hothorn, T., Hornik, K., and Zeileis, A. (2006b), "Unbiased recursive partitioning: A conditional inference framework," *Journal of Computational and Graphical Statistics*, 15, 651–74.

Imboden, M. T., Harber, M. P., Whaley, M. H., Finch, W. H., Bishop, D. L., and Kaminsky, L. A. (2018), "Cardiorespiratory Fitness and Mortality in Healthy Men and Women," *Journal of American College of Cardiology*, 72, 2283–92.

Kaminsky, L. A., Arena, R., and Myers, J. (2015), "Reference standards for cardiorespiratory fitness measured with cardiopulmonary exercise testing: data from the Fitness Registry and the Importance of Exercise National Database," *Mayo Clinic Proceedings*, 90, 1515–23.

Kaminsky, L. A. and Whaley, M. H. (1998), "Evaluation of a new standardized ramp protocol: the BSU/Bruce Ramp protocol," *Journal of Cardiopulmonary rehabilitation*, 18, 438–44.

Kokkinos, P. F., Holland, j. C., Pittance, A. E., Nayrayan, P., Doston, C. O., and Papademetriou, V. (1995), "Cardiorespiratory fitness and coronary heart disease risk factor association in women," *Journal of the American College of Cardiology*, 26, 358 – 364.

Lipkovich, I., Dmitrienko, A., Denne, J., and Enas, G. (2011), "Subgroup identification based on differential effect search - a recursive partitioning method for establishing response to treatment in patient subpopulations," *Statistics in Medicine*, 30, 2601–2621.

Loh, W.-Y. (2002), "Regression trees with unbiased variable selection and interaction detection," *Statistica Sinica*, 12, 361–86.

— (2009), "Improving the precision of classification trees," *Annals of Applied Statistics*, 3, 1710–37.

Loh, W.-Y., He, X., and Man, M. (2015), "A regression tree approach to identifying subgroups with differential treatment effects," *Statistics in Medicine*, 34, 1818–33.

Nugent, C., Guo, W., Müller, P., and Ji, Y. (2019), "Bayesian Approaches to Subgroup Analysis and Related Adaptive Clinical Trial Designs," *JCO Precision Oncology*.

Pollock, M. L., Foster, C., Schmidt, D., Hellman, C., Linnerud, A. C., and Ward, A. (1982), "Comparative analysis of physiologic responses to three different maximal graded exercise test protocols in healthy women," *American Heart Journal*, 103, 363–73.

Ross, R., Blair, S. N., Arena, R., Church, T. S., Després, J.-P., Franklin, B. A., Haskell, W., Kaminsky, L. A., Levine, B. D., Lavie, C. J., Myers, J., Niebauer, J., Sallis, R., Sawada, S. S., Sui, X., and Wisløff, U. (2016), "Importance of assessing cardiorespiratory fitness in clinical practice: a case for fitness as a clinical vital sign: a scientific statement from the American Heart Association," *Circulation*, 134, e653–99.

Sawada, S. S., Lee, I. M., Muto, T., Matuszaki, K., and Blair, S. N. (2003), "Cardiorespiratory fitness and the incidence of type 2 diabetes: prospective study of Japanese men," *Diabetes Care*, 26, 2918 – 2922.

Schnell, P., Tang, Q., Müller, P., Carlin, B. P., et al. (2017), "Subgroup inference for multiple treatments and multiple endpoints in an Alzheimer's disease treatment trial," *The Annals of Applied Statistics*, 11, 949–966.

Seibold, H., Zeileis, A., and Hothorn, T. (2016), "Model-Based Recursive Partitioning for Subgroup Analyses," *International Journal of Biostatistics*, 12, 45–63.

Su, X., Wang, M., and Fan, J. (2004), "Maximum Likelihood Regression Trees," *Journal of Computational and Graphical Statistics*, 13, 586–598.

Zeileis, A. and Hornik, K. (2004), "Generalized M-Fluctuation Tests for Parameter Instability," *Statistica Neerlandica*, 61, 488–508.

Zeileis, A., Hothorn, H., and Hornik, K. (2008), "Model-Based Recursive Partitioning," *Journal Computational and Graphical Statistics*, 17, 492–514.

Zeileis, A. and Hothorn, T. (2013), "A Toolbox of PermutationTests for Structural Change," *Statistical Papers*, 54, 931–954.

Received: September 1, 2020

Accepted: November 14, 2020