CHRONIC FATIGUE SYNDROME AND ITS RELATION WITH ABSENTEEISM: ELASTIC-NET AND STEPWISE APPLIED TO BIOCHEMICAL AND ANTHROPOMETRIC CLINICAL MEASUREMENTS

Anderson Cristiano NEISSE[1]
Fernando Luiz Pereira de OLIVEIRA[2]
Anderson Castro Soares de OLIVEIRA[3]
Frederico Rodrigues Borges da CRUZ[4]
Raimundo Marques do NASCIMENTO NETO[5]

- ABSTRACT: Characterized by persistent fatigue, pain, cognitive impairment and sleep difficulties, Chronic Fatigue Syndrome (CFS) has been common in clinical practice. Studies indicate multiple factors contributing to CFS development: poor sleep, dehydration, psychological stress, hormonal dysfunction, nutrient deficiencies, among others. In risk work conditions, like the shift work of mines, CFS significantly increases the chance of fatal accidents. Work environments of mines suggest the presence of factors that increase the risk of developing CFS. Considering the severity/implications of CFS's symptoms on the social and professional lives as well as on the economy, efforts are targeting its characterization and prevention. This study aims to assess the risk of CFS by studying cross-sectional data on absenteeism of 621 shift workers, measuring 8 anthropometric and 11 biochemical variables as well as age and gender, amounting 21 variables. After imputation, logistic regression was fitted by Stepwise selection, Lasso and Elastic-Net regularization. Results suggest that the models do not discriminate very well due to noise inherent to the dependent variable. However, all models agree on the effects of Sodium and Total Cholesterol on the risk of absenteeism. The Stepwise model also indicates LDL and Triglycerides as significant factors, both Lasso and Elastic-Net show effects for LDL instead. The Elastic-Net model suggests an effect of Potassium, though inconclusive according to the literature.

- KEYWORDS: Absenteeism; biometrics; bootstrap; chronic fatigue syndrome; elastic-net.

[1] Universidade Federal de Viçosa - UFV, Departamento de Estatística, CEP: 36.570-977, Viçosa, MG, Brasil. E-mail: *a.neisse@gmail.com*
[2] Universidade Federal de Ouro Preto - UFOP, Departamento de Estatística, CEP: 35.400-000, Ouro Preto, MG, Brasil. E-mail: fernandoluiz@ufop.edu.br
[3] Universidade Federal de Mato Grosso - UFMT, Departamento de Estatística, CEP: 78.068-600, Cuiabá, MT, Brasil. E-mail: andersoncso@gmail.com
[4] Universidade Federal de Minas Gerais - UFMG, Departamento de Estatística, CEP: 31.270-901, Belo Horizonte, MG, Brasil. E-mail: fcruz@est.ufmg.br
[5] Universidade Federal de Ouro Preto - UFOP, Departamento de Clínicas Pediátricas e do Adulto, CEP: 35.400-000, Ouro Preto, MG, Brasil. E-mail: ufopraimundo@gmail.com

1 Introduction

Chronic Fatigue Syndrome (CFS), also referred to as myalgic encephalomyelitis (ME), is a disease that has been commonly present in clinical practices in the last decades (AFARI and BUCHWALD, 2003). An epidemiological study conducted by Jason *et al.* (1999) estimated a CFS prevalence of 410 cases per 100,000 individuals, which suggests 1 million cases in the United States alone. As estimated by another study conducted by Jason and Njoku (2006), 850 thousand Americans suffer from the syndrome. The disease's first case definition is attributed to the United States Center for Disease Control and Prevention (CDC) (HOLMES *et al.*, 1988) as a means of standardization for epidemiological studies. There have been some revisions of that definition since then (WILLIAMS *et al.*, 2014). According to Fukuda *et al.* (1994), a CFS case is defined by the presence of chronic (or relapsing) fatigue during 6 months or more accompanied of 4 out of the following symptoms: memory or concentration impairment, sore throat, tender lymph nodes, muscle pain, joint pain, headaches, unrefreshing sleep, and post-exertional malaise. Despite its use by most of the academic community, there is criticism to this definition case since it does not demand the presence of important symptoms, like post-exertional malaise (WILLIAMS *et al.*, 2014). The most recently used definition, proposed by the United States Centre for Disease Control and Prevention (CDC) in 2015, requires the presence of profound fatigue persisting or relapsing for more than 6 months, which should be accompanied by post-exertional malaise, unrefreshing sleep and cognitive impairment.

The factors that contribute to the development of CFS still are aimed by the academic community, amongst them are irregular sleep, psychological stress, hormonal dysfunctions, nutrient deficiency, immunological dysfunctions and infections. In a study conducted with nurses by Samaha *et al.* (2007) sleep quality was identified as a significant factor related to CFS. A mass spectroscopy study showed that patients with CFS presented altered levels of phospholipids, cholesterol, branched amino acids, vitamins and mitochondrial metabolites (NAVIAUX *et al.*, 2016). According to Litleskare *et al.* (2018), the prevalence of CFS after 10 years of the Giardia infection was 2.22 to 4.08 times the prevalence presented by the control group. Another case-control study, performed by Nagy-Szakal *et al.* (2018), indicates low levels of betaine and complex lipids as well as elevated triglycerides and phenylacetylglutamine in individuals with CFS. Evidence presented by Maloney *et al.* (2010) suggest a relation between CFS and metabolic syndrome, which includes hypertension, elevated levels of blood sugar, waist fat and abnormal levels of cholesterol. According to Bjørklund *et al.* (2019), individuals with CFS have nutrient deficiencies, like vitamin C, vitamin B, sodium, magnesium, folic acid and fatty acids which also seems to have importance in the CFS severity. A paper written by Bou-Holaigah *et al.* (1995) suggests the relations between CFS and neurally mediated hypotension and shows that its treatment, which includes moderate sodium consumption, was efficient in reducing CFS symptoms in a subgroup of individuals. The efforts to define efficient CFS markers are clearly in course, being a result of the recent recognition of CFS as an impactful disease. According to Kennedy *et al.* (2010), the life quality of children with CFS compares to the one experienced by children with type 1 diabetes mellitus or asthma.

Shift workers are naturally more susceptible to CFS development due to their unusual sleep and rest habits. According to Costa (2010), the night shift work is one of the most studied conditions since it disturbs the sleep cycle modifying rest patterns, resulting in significant stress in the biological circadian rhythms regulation of humans, naturally diurnal

beings. As shown by Shen *et al.* (2006), shift frequency has a positive correlation with the intensity of the fatigue experienced by the shift workers. The disruption of body functions' circadian rhythms is responsible for the shift-lag syndrome, which is characterized by feelings of fatigue, sleepiness, insomnia, digestive difficulties, irritability, reduced mental agility and reduced performance (COSTA, 2010). In risk work conditions, CFS development potentially increases the chance of fatal accidents. As shown by Useche *et al.* (2017), structural equation modelling (SEM) results indicate a significant relationship between fatigue and risky behaviours on bus conduction. The work on alternated shifts of the mining industry not only fits the definition of risky conditions but also includes the factors of irregular sleep and psychological stress. Regardless of the results from this paper, it is important to underline the importance of current health practices to support mining industry workers given the hard and perilous work they are exposed to.

Considering the severity of CFS's symptoms and its implications in the social and professional lives, efforts have been made not only to its characterization but also to its effective prevention. According to Murphy *et al.* (2011), predictive modelling can be an effective tool in the prevention of the syndrome. A study conducted by Huang *et al.* (2009) compared three methods for CFS prediction using gene expression data and the Naive Bayes classifier accomplished 0.7 for the area under the ROC curve (AUC). Searching for CFS relevant factors in women suffering from breast cancer, Servaes *et al.* (2002) used linear regression to examine the contribution of physical, psychological and cognitive factors in the severity of fatigue in patients. By using decision trees, Bronikowski *et al.* (2011) obtained an accuracy of 71.88% predicting CFS based on answers to a medical questionnaire applied in a community based CFS study.

Shift workers frequently complain about irritability, anxiety and stressful work conditions. Sleep deficit and persistent circadian rhythm alteration may lead to CFS, neuroticism, chronic anxiety and/or depression, resulting in an augmented risk of absenteeism and the need of psychotropic medications (COLQUHOUN and SENN, 2000; NAKATA *et al.*, 2004). Aiming to search for evidence that contributes to the prevention of CFS as well as for related factors, this study aimed at the relation of anthropometric and biochemical variables with the risk of absenteeism in shift workers of the mining industry. The main objectives of the study are: (i) Search for a descriptive model that possibly show some level of discrimination power of absenteeism and (ii) Relate the independent variables' effects on absenteeism with the risk factors of CFS. In order to achieve the first objective Logistic regression will be fitted to the data with the Stepwise approach to variable selection. The Elastic-Net and Lasso regularization methods will also be used to adjust the Logistic model in order to verify whether the predictive performance improves due to their flexibility. The second objective will be pursued by relating the effects found by the descriptive model with the ones found as risk factors for CFS in the literature. In the next section, the dataset and all the methods used to achieve these objectives are discussed in more detail.

2 Materials and methods

A cross-sectional study was performed in 2012 on 621 shift workers of a mine located in the Inconfidentes region of the Minas Gerais state, Brazil. The study, entitled "Sindrome Metabólica em Trabalhadores da Mineração do Estado de Minas Gerais", was approved by the Ethics Committee from Universidade Federal de Ouro Preto (CAAE: 0018.0.238.000-11). The individuals work on shifts of 6 hours operating off-road trucks, which are followed by 12 hours rest periods. The data collected consists of 22 variables divided into 8 anthropometric variables, 11 biochemical variables, Sex, Age and the variable Skipped, which is the variable that indicates whether the individual was absent in any giving day in the year of 2012. The Skipped variable is the dependent variable of this study conditioned on the fact that it is the result of interviews of each individual by a medical doctor from the mining company screening for signs of Chronic Fatigue Syndrome, therefore the absenteeism in this study does not include causes such as other common diseases, or holidays, or problems with the employee's family. A summary for the variables is presented in Table 1. After the descriptive analysis all the variables were standardized, the resulting dataset was used in the rest of the analysis, starting from the missing values' imputation. The next subsections will focus on brief reviews and choices of methods for the imputation of missing cases, fitting Logistic Regression, measuring model performance, performing model selection and validation.

Table 1 - Variables present in the initial data set for the study

Description	Method	Type	Mean	SD	% Missing
Was there abseteeism?	Mining Company Files	Binary (Yes = 1)	0.2238	-	0.00
Sex of the individual	Medical file	Binary (Male = 1)	0.9646	-	0.00
Individual's age	Medical file	Discrete	36.7504	7.1336	0.00
Individual's height (m)	Platform stadiometer	Continuous	174.1032	7.3123	5.25
Individual's weight (kg)	0.1 kg precision scale	Continuous	80.3670	12.8352	2.48
Body Mass Index	Weight/Height2	Continuous	26.4756	3.6749	2.48
Avg. Waist-to-Hip Ratio (n=3)	Simple tape measure	Continuous	0.8566	0.1220	2.64
Total Body Fat (%)	Bioimpedance	Continuous	24.3923	8.1730	3.50
Visceral Fat (%)	Bioimpedance	Continuous	7.6712	3.4312	3.33
Avg. Dias. Blood Press. (n = 3)	Semi-autom. Monitor	Continuous	82.7661	9.3659	0.16
Avg. Sist. Blood Press. (n = 3)	Semi-autom. Monitor	Continuous	131.5532	13.7510	0.16
HDL Cholesterol	Enzyme-colorimetric	Continuous	55.4063	16.3538	5.79
LDL Cholesterol	Chol - (HDL + VLDL)	Continuous	109.4432	36.2181	7.63
Triglycerides	Enzyme-colorimetric	Continuous	151.5527	83.6749	5.61
Total Cholesterol	Enzyme-colorimetric	Continuous	194.5765	45.4948	5.61
Calcium	Ion selective electrode	Continuous	9.5704	1.2965	5.61
Phosphorus	Kinetic U.V. test	Continuous	3.5085	0.5965	42.43
Vitamin D	Electrochemiluminescence	Continuous	25.4991	7.7247	38.62
PTH Hormone	Chemiluminescence	Continuous	30.8635	10.8913	25.96
Fasting Glucose	Enzyme-colorimetric	Continuous	86.9475	14.5506	5.08
Sodium	Ion selective electrode	Continuous	144.2196	8.1133	12.70
Potassium	Ion selective electrode	Continuous	4.7249	0.7105	9.72

A study on similar data, measured under the same circumstances, was performed by Souza *et al.* (2015). The study evaluated the association between a lifetime of shift work on mines and blood pressure, fasting glucose, anthropometric variables, body composition and heart rate variability.

2.1 Missing data and missForest imputation

Missing values have been an issue since the beginning of field research, mostly due to the fact that the analytical procedures used, many of which were developed in the early 20th century, aimed to be used on complete datasets (GRAHAM, 2009). The missingness in biology and medicine is usually caused by sample mishandling, measurement errors or non-response which frequently lead to missing cases' removal by the researcher, which is called the complete-case analysis (STERNE *et al.*, 2009). According to Rubin (1976) in his description of the missing mechanisms, there are three classifications: Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR). Since the reason that the data is missing in this study is due to some randomly lost measurements, it is assumed that the data is at least MAR.

Many imputation methods for missing data have been proposed, a considerable group of them being based on the mean of observed data, like k-nearest neighbours (KNN), Bayesian principal component analysis (bPCA), random forests (missForest) and multiple imputations by chained equations (MICE) (SCHMITT *et al.*, 2015). The missForest method outperformed other methods in a number of comparative studies using laboratory, medical and biometric datasets (CIHAN, 2018; STEKHOVEN and BÜHLMANN, 2012; WALJEE *et al.*, 2013). The missForest method was proposed by Stekhoven and Buhlmann (2012), as a non-parametric method that copes with different types of variables simultaneously by using the random forests model to predict missing values. Basically, a random forest is fitted to the observed data and used to estimate the missing cases, the process is repeated for each of the variables with missing cases. More details on the method and the algorithm are described by Stekhoven and Bühlmann (2012). Because they can introduce significant bias both to analysis and to imputation, variables with more than 20% missing cases (Phosphorus, Vitamin D and PTH Hormone) were removed from this study at the beginning while the remaining cases were imputed with the missForest method.

2.2 Logistic regression, variable selection and regularization

Since the dependent variable of absenteeism is coded as a binary outcome, the Logistic Regression is a natural choice for a descriptive model, therefore being used in this study with multiple fitting procedures. The logistic regression, a generalized linear model assuming a binomial distribution for the response variable, is frequently applied to classification problems and risk scores generation in the medical field. The parameters β are estimated with the maximum likelihood method considering the observations as independent and identically distributed (KULOTHUNGAN *et al.*, 2014). A common challenge in the regression with multiple variables is selecting the best model to describe the data and possibly predict new observations.

A common approach to the model selection is the Stepwise method, which consists of an algorithm for automatic variable selection based on a predefined performance metric (HOCKING, 1976). It was proposed as an alternative to the best-subset selection algorithm which evaluates every possible combination of subsets of size s from the p independent variables ($s = 1, \ldots, p$) and is computationally costly. All variations of Stepwise are based on the metric of choice and the two approaches to variable selection: forward selection and backwards elimination. The forward approach consists of starting on a null model (with no

variables) and increasing the number of variables by evaluating every candidate variable at each model size. The backwards elimination is very similar to the forward, only it begins with the full model and each step removes one variable. A commonly used variation is called the Bidirectional Stepwise, which starts with the null model but at each step also considers the removal of a variable (EFROYMSON, 1960). One possible limitation of Stepwise is the case where the events-per-variable ratio (EPV) is lower than 10. According to Heinze *et al.* (2018), the EPV quantifies how balanced is the information provided by the data and the number of parameters to be estimated. The dataset of this study has EPV = 8.6875 and, since the Stepwise does consider all variables in the process of choosing the final model, the Logistic Regression fitted to it might suffer from this limitation. Consequently, as shown by Pavlou *et al.* (2015), EPV < 10 might result in poorer calibration and therefore prejudice the predictive performance of the model. The authors also show that variable regularization like Ridge, Lasso and Elastic-Net might mitigate such limitations by means of performing regularization on the independent variables.

An alternative approach to the Stepwise is the Elastic-Net method, that generalizes the Ridge and Lasso methods in order to perform variable selection as well as regularization. The Ridge method, proposed by Hoerl and Kennard (1988), fits the model using an L2-type penalized residual sum of squares:

$$RSS_{Ridge} = RSS + \lambda \sum_j \beta_j^2$$

which shrinks the parameters towards zero according to λ. The hyperparameter λ is usually chosen with cross-validation in order to optimize some model-performance metric. However, the Ridge method never shrinks a parameter to zero, therefore always returning a full model and not performing variable selection. Aiming to tackle this characteristic, Tibshirani (1996) proposed the Lasso, which introduces a L1-type penalization:

$$RSS_{Lasso} = RSS + \lambda \sum_j |\beta_j|,$$

that performs regularization and eventual variable selection by effectively setting parameters to zero. Despite being more promising, the Lasso method also showed some limitations: (i) When p >> n Lasso will choose a maximum of n variables; (ii) In cases where n > p with multicollinearity then Ridge will dominate the Lasso; (iii) When there is a group of variables with high correlation the Lasso will pick just one of them, not caring which one it chooses. After considering those limitations Zou and Hastie (2005) proposed the Elastic-Net, combining the L1 and L2-type penalizations into:

$$RSS_{Elastic-Net} = RSS + \lambda [h(1 - \alpha) \sum_j \beta_j^2 + \alpha \sum_j |\beta_j|]$$

with another hyperparameter α that, conjointly with λ, is chosen in order to optimize some performance metric. As considered by Zou and Hastie (2005), the Elastic-Net is a generalization to the Lasso that performs well in the situations where it has limitations. Methods that perform both variable selection and regularization allow further improvement of performance not possible by the Stepwise approach. Despite regularization methods lacking interpretability due to their natural bias towards zero introduced by the regularization (HEINZE *et al.*, 2018), they can also be used as further evidence of whether

the results from Stepwise are consistent. The Logistic regression was fitted by using Stepwise, Lasso and Elastic-Net for comparison of predictive performance.

2.3 Model performance measurement

All the methods used in this study depend on a model selection procedure. The most used method for model selection is Cross-Validation (CV) based on a performance metric. Among the most frequent performance metrics for classification models are Accuracy, Sensitivity, Specificity and the area under the receiving operating characteristic (ROC) curve. This study used stratified repeated k-fold Cross-Validation with the AUC metric to perform model selection, both methods will be explained in more detail in subsections. Also called True Predictive Rate, the Accuracy measures the overall performance of the classifier, i.e. the proportion of predictions that matched the true class. However, it does not consider the intra-class error, if one chooses to predict every case with positive probability as the target class then the accuracy would be its proportion in the data and every class other than the target would be classified wrongly. Therefore, when used alone Accuracy may lead to misleading conclusions, which is why it is commonly accompanied by Sensitivity and Specificity (PROVOST *et al.*, 1998). Sensitivity (True Positive Rate) measures the performance of the model conditioned on the cases where the class is the target of prediction, while Specificity (True Negative Rate) conditions the performance on the cases other than the target class allowing more information for the decision making based on the model (ALTMAN and BLAND, 1994). As pointed out by Provost *et al.* (1998), there might be greater interest in optimizing the Sensitivity than in optimizing Specificity or Accuracy of a model depending on the research field and problem at hand.

For models that predict a continuous probability of the target event, there are infinite possible classifiers based on the chosen probability cut-off, each of them resulting in different measurements of performance. A method that allows choosing a good probability cut-off for a classifier is the Receiving Operating Characteristic curve (ROC curve), which is built upon the Specificity and Sensitivity measures allowing to visualize their trade-off. Also, a metric that allows to measure a classifier's discrimination power is the area under the receiving operating characteristic curve (AUC) since it is directly related to the curve's proximity to the perfect or the random classifiers (Figure 1). A perfect classifier would have AUC = 1 while the random classifier has AUC = 0.5 characterizing a random-guessing model. An important property of the AUC metric is its equivalence to the probability that a classifier ranks a positive chosen value over negative chosen value, which corresponds to the Wilcoxon rank test (Hanley and Mcneil, 1982). There is also a relation between AUC and the Gini coefficient (Gini + 1 = 2 × AUC) as shown by Hand and Till (2001).

2.4 Model selection and evaluation

Cross-validation (CV) is a method that is used to estimate the performance of a given model in predicting values for new data. The most commonly used CV methods are the k-fold and the leave-one-out, usually applied repeatedly (repeated k-fold) in order to reduce the high variability (VANWINCKELEN and BLOCKEEL, 2012). In the classification context, it is also common to perform the stratified cross-validation, which guarantees class balance like the original data's balance in each fold. Despite these improvements, however,

the k-fold methods in general result in underestimation of the true performance as a result from a bias introduced by using only a proportion of $(k\ -\ 1)/k$ from the whole data set to fit the model (VANWINCKELEN and BLOCKEEL; 2012). Despite the improvement of accuracy provided by the repeated k-fold it is still biased and therefore is suggested to model selection but not to estimate model performance (KOHAVI, 1995).

The non-parametric bootstrap, proposed by Efron (1979), is utilized to estimate the variability of any measure of interest that is a function from a representative random sample. Considering a data set with n observations, bootstrap's general idea consists of obtaining m resamples with replacement of size n from the original data and evaluating the measure of interest in these m resamples resulting in an empirical distribution of the measurement. From the empirical distribution, it is possible to obtain estimates (average), confidence intervals and standard errors for the estimator. The most frequent method for bootstrap confidence intervals is the quantile, in which a $1 - \alpha \times 100\%$ confidence interval is obtained by the $\alpha/2$ and $(1\ -\ \alpha)/2$ quantiles from the bootstrap empirical distribution. A study by Efron (1987) shows that a considerably small coefficient of variability (9%) is obtained when generating 200 bootstrap measurements which reduce to 1% when m is set to 1.000, this last value is considered as a sufficiently large number of resamples. As shown by Kohavi (1995), while the bootstrap usually results in lesser variability than k-fold ($k = 10$ and $k = 20$), neither of them dominates in terms of relative bias and whether one outperforms the other depends on the data set.

When considering the performance of prediction rules trained on the data, Efron and Tibshirani (1997) proposed the 632+ bootstrap as an attempt to correct the bias inherent to the classical bootstrap and as an improvement on cross-validation. It maintains the reduced variability in relation to k-fold and improves on the bias, therefore turning the 632+ bootstrap into a more appealing method for model performance estimation. As described by Witten *et al.* (2016), at any bootstrap resample, as n grows the proportion of cases not picked will tend to:

$$(1\ -1/n)^n\ \approx\ e^{-1} = 0.368,$$

which is the chance of a particular observation not being picked at all. The performance from the model fitted to the resample will give rather optimistically biased estimations if evaluated in the resample itself since it estimated the coefficients from the very same data set. The training set (resample) has only $0.632 \times 100\%$ of the original cases, therefore the model originated from it will result in a pessimistically biased performance estimate when evaluated in the 0.368 remaining cases (test), despite it having size n. The main idea of the 632+ bootstrap is to evaluate both the training ($\varepsilon_{training}$) and the test (ε_{test}) performance measures and to obtain the weighted performance estimate:

$$\varepsilon_{632}\ =\ 0.632\ \times\ \varepsilon_{test}\ +\ 0.368\ \times\ \varepsilon_{training}$$

by combining the pessimist test performance with the optimist training performance.

This study performs stratified repeated cross-validation for model selection in order to guarantee the target-class balance in the model selection and reduces variability in the measurements of performance. In order to obtain the empirical distributions for the parameters, classical non-parametric bootstrap is used to allow further investigation of the

selected factors' effects on the absenteeism. Finally, in order to compare the models' performance in terms of Accuracy, Sensitivity and Specificity, the 632+ bootstrap is used to obtain an estimate of the empirical distributions for model performance. Confidence intervals are obtained by using the quantile approach using the empirical distributions estimated by bootstrap, this method is used since it allows to obtain intervals for different models by the same non-parametric procedure.

2.5 Software and packages

All the analysis and plots present and discussed in this paper were produced using the R Programming language v3.5.1 (R CORE TEAM, 2020) and the RStudio IDE v1.3.125. The packages used were: *caret* for model fitting and selection (KUHN *et al.*, 2019); *ggplot2* for plots and graphics generation (WICKHAM, 2016); *missForest* for imputation of missing values (STEKHOVEN, 2013); *tibble* (MÜLLER and WICKHAM, 2019), *dplyr* (WICKHAM *et al.*, 2019) and *tidyr* (WICKHAM and HENRY, 2019) for data manipulation and cleansing; *purrr* (HENRY and WICKHAM, 2019) for efficient and readable iterations as well as *furrr* (VAUGHAN and DANCHO, 2018) for iterations' parallel processing in R. The model evaluation by non-parametric bootstrap and 632+ bootstrap methods was implemented by the author using a portion of the packages cited above.

3 Results and discussion

The variables with more than 15% missing cases were removed at the start of the analysis, namely Phosphorus, Vitamin D and the PTH hormone measurements. The missForest method was used on the standardized remaining 19 variables for missing cases imputation. Afterwards, Weight and Height were removed since BMI already accounts for most of their variability and, therefore, their removal is made to avoid multicollinearity and results in 17 variables in the whole dataset. Stepwise and grid-search of parameters for Lasso and Elastic-Net were done by using stratified 10-fold cross-validation repeated 10 times, which used the area under the ROC curve as the optimization metric. The resulting models were used to perform non-parametric bootstrap on the coefficients as well as 632+ bootstrap on the measurements of Accuracy, Sensitivity and Specificity. The subsequent subsections present the results of performance and estimated coefficients from the resulting models as well as the discussion in the context of this paper's objectives. It is important to underline that the significance level for this study is set to $\alpha = 0.10$ given the fact that it is an early, cross-sectional study that has the aim of detecting possible relations for further investigation. Also, discussion takes into consideration that the absenteeism occurrences were evaluated by a medical doctor through interviews with the individuals as a means to rule out causes clearly unrelated to Chronic Fatigue Syndrome.

3.1 Resulting models and effect sizes

The resulting models (Stepwise, Lasso with $\lambda = 0.02409091$ and Elastic-net with $\lambda = 0.5868687$ and $\alpha = 0.03787879$), selected based on the imputed data, were fitted to both the imputed and complete-cases datasets for comparison. The models from imputed and complete-case data sets were the same except for the fact that the imputed data had

more precise confidence intervals due to it having more observations (621 as opposed to 501). Also, the coefficient for potassium was shrunk to zero in the complete-cases Elastic-Net, which was not the case for the model from imputed data. Given that, the results for the complete-cases data set's models were omitted for the sake of brevity. The relative effects of each model are presented in Figure 2 as an attempt to first compare the effects of the three different methods.

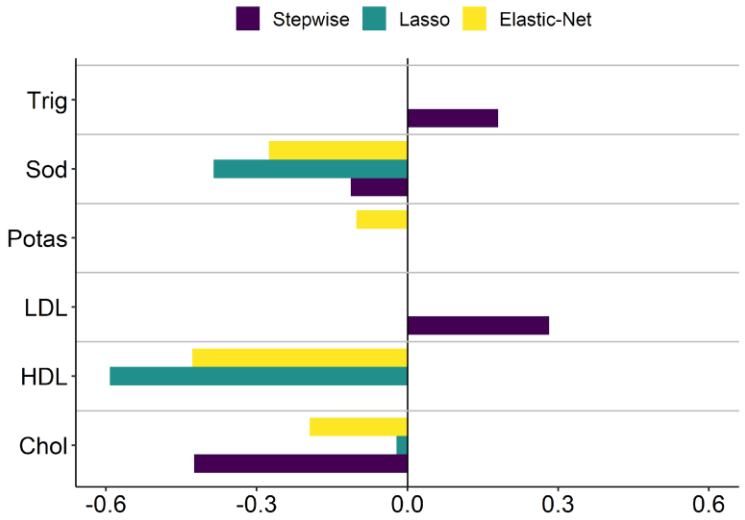

Figure 2 - Relative effects of coefficients for each model. Each model had its coefficients divided by the sum of absolute values of all their coefficients.

As shown by the relative effects in Figure 2, accounting for the fact that there is a different number of variables for each model the coefficients for HDL and Sodium had similar relevance for both Lasso and Elastic-Net. Total Cholesterol was the third place in effect size for both models though, while the Stepwise had it as the coefficient with the highest relevance. Also, the Lasso Total Cholesterol relevance showed a high disparity when compared to the other two models, also explained by the higher relevance of HDL and Sodium in the Lasso as shown in Table 2. The fact that both Triglycerides and LDL had similar effects on the Stepwise model and contrary to the effects of HDL in the regularized models might be explained by their complementary property as decompositions of the Total Cholesterol. The coefficients are presented in Table 2 with the bootstrap estimates and confidence intervals.

Despite the confidence intervals for Lasso and Elastic-Net having limits relatively close to zero, all confidence intervals indicated significant effects for the selected variables in the models. All models agreed in negative effects of Total Cholesterol and Sodium in the risk of skipping work, meaning that individuals which are 1 standard deviation above the average in Total Cholesterol and Sodium would have a decrease of -0.6458 and -0.1711 in the log-odds of skipping work respectively, according to the Stepwise coefficients. However, coefficients for LDL and Triglycerides from the stepwise indicate an increase in the log-odds of absenteeism. Despite the regularization models not suggesting significant

effects for LDL and Triglycerides, the effect of HDL is significant and inversely proportional to LDL since both are components of Total Cholesterol together with Triglycerides. The Elastic-Net fitted to the imputed data was the only model to suggest effects of Potassium and therefore, despite it being statistically significant with $\alpha = 0.10$ according to the CI, it was considered the variable with least evidence of significance. The bootstrap densities that originated the confidence intervals of Table 2 are shown in Figure 3 together with the intervals and bootstrap estimates for further understanding.

Table 2 - Coefficient estimates and bootstrapped means with 90% confidence intervals for each model. Coefficients scale is in log-odds and models were fitted to standardized variables

	Stepwise		Lasso		Elastic-Net	
Names	**Estimate**	**Bootstrap [90% C.I.]**	**Estimate**	**Bootstrap [90% C.I.]**	**Estimate**	**Bootstrap [90% C.I.]**
(Intercept)	-1.241**	-1.256 [-1.428;-1.092]	-1.236 (80.58)	-1.228 [-1.387;-1.072]	-1.241 (80.86)	-1.229 [-1.397;-1.071]
HDL	-	-	-0.064 (89.65)	-0.091 [-0.216;-0.005]	-0.017 (24.16)	-0.027 [-0.059;-0.002]
LDL	0.429.	0.442 [0.045;0.849]	-	-	-	-
Trig	0.275*	0.272 [0.067;0.465]	-	-	-	-
Chol	-0.646*	-0.661 [-1.106;-0.224]	-0.002 (0.26)	-0.071 [-0.186;-0.002]	-0.008 (0.88)	-0.021 [-0.049;-0.001]
Sodium	-0.171.	-0.183 [-0.366;-0.022]	-0.042 (34.9)	-0.075 [-0.188;-0.003]	-0.011 (9.3)	-0.019 [-0.040;-0.002]
Potassium	-	-	-	-	-0.004 (6.5)	-0.014 [-0.033;-0.001]

The densities on Figure 3 underline the main reason for the confidence intervals of regularized regression being so close to zero, namely the fact that the regularization introduces a bias towards zero in the estimates of the coefficients as stated by Heinze *et al.* (2018).

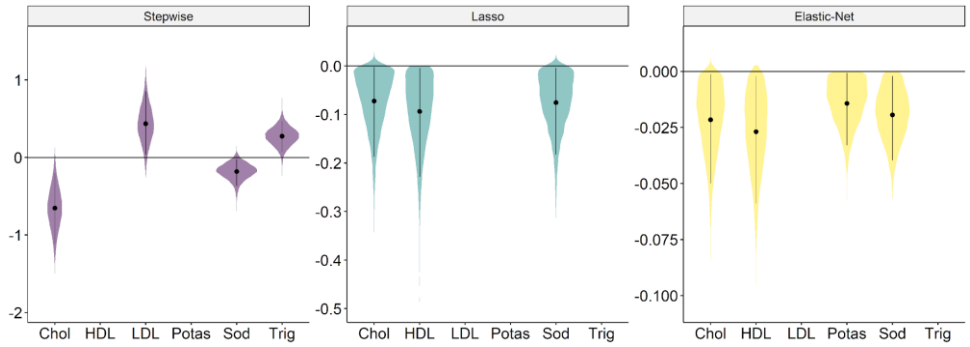

Figure 3 - Bootstrap violins (mirrored density) and 90% confidence intervals for the effects of variables for each model.

One should note the difference between regularized estimates for Total Cholesterol and their bootstrap counterparts in Table 2 meaning that, according to the bootstrap estimates, the regularization techniques (Lasso and Elastic-Net) estimated a coefficient far from the average of its empirical distribution. However, the bootstrap estimates are simply the average and the Figure 3 densities underline that the regularized estimates were closer to the highest density area of the empirical distribution which does not match the mean like

it would for symmetric distributions. This effect in the other regularized coefficients, despite softer, is present, nonetheless.

Other than being a tool for variable selection and improvement of predictions, the regularized regressions do not allow for clear interpretation of the coefficients, therefore remaining only as a concordance measurement in this context.

Despite the counter-intuitiveness of higher-than-average levels of Sodium as a beneficial factor, there are studies that agree with these results in the context of Chronic Fatigue Syndrome, therefore, suggesting the relation with absenteeism. According to Rowe and Calkins (1998), there is a substantial body of clinical evidence supporting the relationship between various forms of hypotension (including the neurally mediated) with CFS and idiopathic fatigue. Most recently, a pilot study conducted by Comhaire (2018) also suggests the benefits of sodium dichloroacetate treatment for patients with the syndrome. As for the cholesterol variables (Total, HDL, LDL and Triglycerides), there are also studies supporting the evidence found in the analysis presented in this paper. A clinical study performed by De Lorenzo *et al.* (1998) indicated that patients with CFS had higher levels of Triglycerides and lower levels of HDL when compared with patients without the syndrome. Also, the ratio HDL/Total cholesterol was significantly lower in CFS patients suggesting that higher Total Cholesterol conditioned on lower levels of Triglycerides and HDL were associated with a lower risk of CFS. More recently, a study conducted by Tomic *et al.* (2012) with a female group of patients as subjects also found higher levels of Triglycerides and lower levels of HDL in the CFS group of patients as opposed to the control group, the study found no evidence of difference for total and LDL cholesterol between groups. Lastly, for Potassium, Dechene (1993) suggests the relation of low levels of potassium with increased risk of CFS. However, studies that address this potential relation found in the literature were inconclusive: a study by Nijs *et al.* (2003) showed that while some patients with CFS had low levels of Potassium, others showed high levels, therefore, concluding that they presented abnormal levels of the mineral; another study, conducted by Lerner *et al.* (1997) found no evidence of difference in Potassium levels between CFS and control groups.

3.2 Performance evaluation

All models showed significant effects for variables relating them to the risk of absenteeism and, with support by the literature, indirectly relating absenteeism with the risk of developing Chronic Fatigue Syndrome. That considered, the possibility of discriminating between groups with high and low risk of absenteeism becomes of interest to increase the success of CFS prevention. The ROC curve for each of the three models obtained in this study are presented in Figure 4 (a) as the first assessment of the models' discrimination power. The area under the curve (AUC) obtained by the model selection was 0.5843, 0.5697 and 0.5746 for Stepwise, Lasso and Elastic-Net respectively. Not only the AUC but also the ROC curves for the models were very similar, the results slightly higher than 0.50 (Random classifier AUC) suggests poor discrimination. The fact that the AUC was also the metric of model selection leads to the necessity of measuring the variability in the performance measures. In order to assess the performance and its variability, Figure 4 (b) presents the 632+ bootstrap estimates, confidence intervals and densities for the measurements of Accuracy (ACC), Sensitivity (SNS) and Specificity (SPC) for all three models.

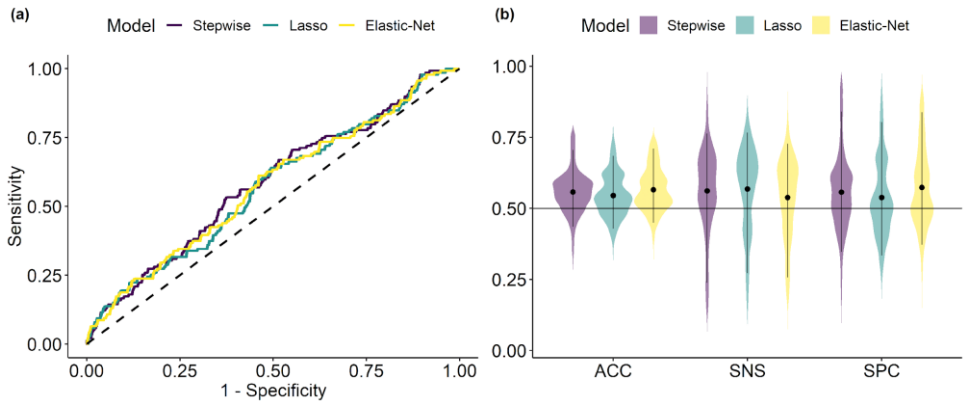

Figure 4 - **(a)** ROC curves for the models adjusted to the imputed dataset. The dashed line represents the random classifier for comparison. **(b)** Bootstrap 632+ 95%confidence intervals for the models' performance.

As suggested by the low AUC, the results in Figure 4 (b) confirms that not only the performance is poor but also it is not statistically different from a random classifier at $\alpha = 0.10$. The AUC is omitted from the plot since the fact of it being used as optimization metric biases it upwards forcing statistical significance according to the non-parametric confidence interval, therefore, inducing miss-interpretation. Despite not showing significant results in terms of performance, the empirical densities obtained by the 632+ bootstrap were very similar between the three models. This lack of discrimination power despite the detected variables might relate to the noise inherent to the dependent variables nature since it measures just whether there was at least one occurrence of absenteeism during the whole year of 2012. One might notice the lower variability in Accuracy as opposed to Sensitivity and Specificity, however, that is caused because Accuracy is nothing more than a weighted average of the latter two.

3.3 Strengths and limitations

This study is strengthened by the fact that the database has 621 observations of individuals, a size not commonly seen in most clinical trials. Also, the clinical measurements were assessed by appropriate techniques performed in a laboratory by trained professionals and researchers of the medical field. On the other hand, the study is limited by the cross-sectional design which assesses the individuals in a specific point in time and therefore not being able to confidently assume causation nor strong evidence based on the data observed. Also, the confounding or noise inherent to the measurement of the dependent variable of absenteeism (Skipped) since it is only the indicator of the occurrence of absenteeism in the whole year of 2012. Such limitation results in limited strength of detected relations and might hide additional relations due to confounding factors. Additional studies with more precise measurement of absenteeism, such as the number of

occurrences, are necessary for further investigation as well as a longitudinal framework to allow causal inference.

Conclusions

Chronic Fatigue Syndrome (CFS) or myalgic encephalomyelitis (ME) is a critical disease due to its severity being comparable to type 1 diabetes mellitus or asthma. Yet the corpus of evidence of its causes as well as its relations with other conditions is still to be consolidated as the efforts are ongoing in the clinical scientific community. This study aimed to contribute to the corpus of evidence of CFS/ME by assessing indirectly the relations between CFS and absenteeism in shift workers of the mining industry, individuals that are inserted in an environment susceptible to a higher risk of CFS than usual. The models obtained in this study had no discrimination power between individuals with a higher and lower risk of absenteeism despite showing significant effects for several variables. However, the detected effects of 5 out of 6 significant variables were found to be related to the factors present in cases of Chronic Fatigue Syndrome according to the reviewed literature. These findings amount to some evidence of a relation between absenteeism and CFS/ME and the need for further investigation. The lack of discrimination power despite the presence of significant variables might happen due to the noise which is inherent to the dependent variable's nature, being just the indicator of whether absenteeism occurred in the whole year of 2012. This study's inferences are not enough to suggest interventions aiming at the prevention of disasters related to the mining work. However, we hope to draw the attention from direct and indirect agents to important relations identified that might affect the health and life quality of these workers. Future studies with more precise measurements of absenteeism and use of longitudinal frameworks might reveal stronger effects of the selected variables as well as significant discrimination power.

Acknowledgements

This study was supported by grants from Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Universidade Federal de Ouro Preto (UFOP), Universidade Federal de Viçosa and project Prevenção da Fadiga. Authors would like to thank reviewers and editors for their comments and suggestions.

NEISSE, A. C.; OLIVEIRA, F. L. P.; OLIVEIRA, A. C. S.; CRUZ, F. R. B.; NETO, R. M. N. Síndrome da fadiga crônica e sua relação com o absenteísmo: elastic-net e stepwise aplicado a medidas clínicas bioquímicas e antropométricas. *Rev. Bras. Biom*. Lavras, v.39, n.1, p.221-239, 2021.

- *RESUMO: Caracterizada por fadiga persistente, dor, dificuldades cognitivas e de sono, a Síndrome da Fadiga Crônica (CFS) tem sido comum nas práticas clínicas. Estudos indicam múltiplos fatores que contribuem para o desenvolvimento do CFS: sono insatisfatório, desidratação, estresse psicológico, disfunção hormonal, deficiência de nutrientes, entre outros. Em condições de trabalho de risco, como o trabalho por turnos em minas, o CFS aumenta significativamente a chance de acidentes fatais. Os ambientes de trabalho das minas sugerem a presença de fatores que aumentam o risco de desenvolvimento de CFS. Considerando a gravidade/implicações dos sintomas do CFS na vida social e profissional, bem como na economia, os esforços estão direcionados à sua caracterização e prevenção. Este estudo tem como objetivo avaliar o risco de CFS por meio do estudo de dados transversais sobre absenteísmo de 621 trabalhadores em turnos, mensurando 8 variáveis antropométricas e 11 bioquímicas, além de idade e sexo, totalizando 21 variáveis. Após a imputação, a regressão logística foi ajustada por seleção Stepwise, Lasso e regularização Elastic-Net. Os resultados sugerem que os modelos não discriminam muito bem devido ao ruído inerente à variável dependente. No entanto, todos os modelos concordam quanto aos efeitos do sódio e do colesterol total no risco de absenteísmo. O modelo Stepwise também indica LDL e triglicerídeos como fatores significativos, tanto o Lasso quanto o Elastic-Net mostram efeitos para LDL. O modelo Elastic-Net sugere um efeito do Potássio, embora inconclusivo de acordo com a literatura.*
- *PALAVRAS-CHAVE: Absenteísmo; biometria; bootstrap; síndrome da fadiga crônica; elastic-Net.*

References

AFARI, N.; BUCHWALD, D. Chronic fatigue syndrome: a review. *American Journal of Psychiatry*. v.160, n.2, p.221–236, 2003.

ALTMAN, D. G.; BLAND, J. M. Diagnostic tests. 1: Sensitivity and specificity. *British Medical Journal*. v.308, n.6943, p.1552, 1994.

BJØRKLUND, G.; DADAR, M.; PEN, J. J.; CHIRUMBOLO, S.; AASETH, J. Chronic fatigue syndrome (CFS): Suggestions for a nutritional treatment in the therapeutic approach. *Biomedicine and Pharmacotherapy*. v.109, p.1000–1007, 2019.

BOU-HOLAIGAH, I.; ROWE, P. C.; KAN, J.; CALKINS, H. The relationship between neurally mediated hypotension and the chronic fatigue syndrome. *Journal of the American Medical Association*, v.274, n.12, p.961–967, 1995.

BRONIKOWSKI, C. M.; WENG, A.; FURST, J. D.; RAICU, D. S. Prediction of chronic fatigue syndrome using decision tree-based ensemble methods. *WorldComp*, v.1, p.1–6, 2011.

CIHAN, P. A Comparison of five methods for missing value imputation in data sets. *International Scientific and Vocational Studies Journal*, v.2, n.2, p.80–85, 2018.

COLQUHOUN, D.; SENN, S. Is NADH effective in the treatment of chronic fatigue syndrome? *Annals of allergy, asthma and immunology*, v.84, p.6, p.639–640, 2000.

COMHAIRE, F. Treating patients suffering from myalgic encephalopathy/chronic fatigue syndrome (ME/CFS) with sodium dichloroacetate: An open-label, proof-of-principle pilot trial. *Medical Hypotheses*, v.114, p.45–48, 2018.

COSTA, G. Shift work and health: current problems and preventive actions. *Safety and Health at Work*, v.1, n.2, p.112–123, 2010.

DECHENE, L. Chronic fatigue syndrome: influence of histamine, hormones and electrolytes. *Medical Hypotheses*. v.40, n.1, p.55–60, 1993.

DE LORENZO, F.; XIAO, H.; MUKHERJEE, M.; HARCUP, J.; SULEIMAN, S.; KADZIOLA, Z.; KAKKAR, V. V. Chronic fatigue syndrome: physical and cardiovascular deconditioning. *QJM: Monthly Journal of the Association of Physicians*, v.91, n.7, p.475–481, 1998.

EFRON, B. Computers and the theory of statistics: thinking the unthinkable. *SIAM Review*. v.21, p.4, p.460–480, 1979.

EFRON, B. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, v.82, p.397, p.171–185, 1987.

EFRON, B.; TIBSHIRANI, R. Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association*, v.92, n.438, p.548–560, 1997.

EFROYMSON, M. A. Multiple regression analysis. *Mathematical Methods for Digital Computers*, v.1, p.191–203, 1960.

FUKUDA, K.; STRAUS, S. E.; HICKIE, I.; SHARPE, M. C. DOBBINS, J. G.; KOMAROFF, A. International Chronic Fatigue Syndrome Study Group: The chronic fatigue syndrome: a comprehensive approach to its definition and study. *Annals of Internal Medicine*, v.121, n.12, p.953–959, 1994.

GRAHAM, J. W. Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, v.60, p.549–576, 2009.

HAND, D. J.; TILL, R. J. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, v.*45*, n.2, p.171-186.

HANLEY, J. A.; MCNEIL, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, v.143, n.1, p.29–36, 1982.

HEINZE, G., WALLISCH, C.; DUNKLER, D. Variable selection: a review and recommendations for the practicing statistician. *Biometrical Journal*, v.60, p.3, p.431–449, 2018.

HENRY, L.; WICKHAM, H. purrr: *Functional programming tools*. R package version 0.3.2. 2019.

HOCKING, R. R. The analysis and selection of variables in linear regression. *Biometrics*, v.32, p.1–49, 1976.

HOERL, A.; KENNARD, A. Ridge regression. *Encyclopedia of Statistical Sciences*, v.8, 1988.

HOLMES, G. P.; KAPLAN, J. E.; GANTZ, N. M.; BRUS, I. Chronic fatigue syndrome: a working case definition. *Annals of Internal Medicine*, v.108, n.3, p.387–389, 1988.

HUANG, L. C.; HSU, S. Y.; LIN, E. A comparison of classification methods for predicting Chronic Fatigue Syndrome based on genetic data. *Journal of Translational Medicine*, v.7, n.1, p.81–88, 2009.

JASON, L. A. *et al.* Testing equivalence simultaneously for location and dispersion of two normally distributed populations. *Biometrical Journal*, v.36, p.643–660, 1999.

JASON, L. A. T. H. S.; NJOKU, M. G. C. The face of CFS in the US. *CFIDS Chronicle*, v.1, p.16–21, 2006.

KENNEDY, G.; UNDERWOOD, C.; BELCH, J. J. F. Physical and functional impact of chronic fatigue syndrome/myalgic encephalomyelitis in childhood. *Pediatrics*, v.125, n.6, p.1324–1330, 2010.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai*, v.14, n.2, p.1137–1145, 1995.

KUHN, M.; WING J.; WESTON, S.; HUNT, T. *caret: Classification and regression training*. R package version 6.0-84. https://CRAN.R-project.org/package=caret, 2019.

KULOTHUNGAN, V.; RAMAKRISHNAN, R.; SUBBIAH, M.; RAMAN, R. Risk score estimation of diabetic retinopathy: statistical alternatives using multiple logistic regression. *Journal of Biometrics and Biostatistics*, v.5, n.5, p.1–6, 2014.

LERNER, A. M.; GOLDSTEIN, J.; CHANG, C. H.; O'NEILL, W. Cardiac involvement in patients with chronic fatigue syndrome as documented with Holter and biopsy data in Birmingham. *Infectious Diseases in Clinical Practice*, v.6, p.327–333, 1997.

LITLESKARE, S.; RORTVEIT, G.; EIDE, G. E.; HANEVIK, K.; LANGELAND, N.; WENSAAS, K. A. Prevalence of irritable bowel syndrome and chronic fatigue 10 years after Giardia infection. *Clinical Gastroenterology and Hepatology*, v.16, n.7, p.1064–1072, 2018.

MALONEY, E. M.; BONEVA, R. S.; LIN, J. M. S.; REEVES, W. C. Chronic fatigue syndrome is associated with metabolic syndrome: results from a case-control study in Georgia. *Metabolism*, v.59, n.9, p.1351–1357, 2010.

MULLER, K.; WICKHAM, H. *tibble: Simple data frames*. R package version 2.1.1. https://CRAN.R-project.org/package=tibble, 2019.

MURPHY, S. M.; CASTRO, H. K.; SYLVIA, M. Predictive modelling in practice: improving the participant identification process for care management programs using condition-specific cut points. *Population Health Management*, v.14, n.4, p.205–210, 2011.

NAGY-SZAKAL, D.; BARUPAL, D. K.; LEE, B.; LEVINE, S. Insights into myalgic encephalomyelitis/chronic fatigue syndrome phenotypes through comprehensive metabolomics. *Scientific Reports*, v.8, p.1, p.10056–10067, 2018.

NAKATA, A., HARATANI, T., TAKAHASHI, M.; ARAKI, S. Association of sickness absence with poor sleep and depressive symptoms in shift workers. *Chronobiology International*, v.21, p.6, p.899–912, 2004.

NAVIAUX, R. K. *et al.* Metabolic features of chronic fatigue syndrome. *Proceedings of the National Academy of Sciences*, v.113, n.37, p.5472–5480, 2016.

NIJS, J.; DE BECKER, P.; DEMANET, C.; MCGREGOR, N. R.; ENGLEBIENNE, P.; VERHAS, M.; DE MEIRLEIR, K. Monitoring a hypothetical channelopathy in chronic fatigue syndrome: preliminary observations. *Journal of Chronic Fatigue Syndrome*, v.11. n.1, p.117–133, 2003.

PAVLOU, M.; AMBLER, G.; SEAMAN, S. R.; GUTTMANN, O.; ELLIOTT, P.; KING, M.; OMAR, R. Z. How to develop a more accurate risk prediction model when there are few events. *British Medical Journal,* v.351, p.h3868, 2015.

PROVOST, F. J.; FAWCETT, T.; KOHAVI, R. The case against accuracy estimation for comparing induction algorithms. *ICML*, v.98, p.445-453, 1998.

ROWE, P.; CALKINS, H. Neurally mediated hypotension and chronic fatigue syndrome. *The American Journal of Medicine*, v.105, n.3, p.15–21, 1998.

RUBIN, D. B. Inference and missing data. *Biometrika*, v.63, n.3, p.581–592, 1976.

R CORE TEAM. *R: A language and environment for statistical computing*. R Foundation, Viena, Austria, https://www.R-project.org/, 2020.

SAMAHA, E.; LAL, S.; SAMAHA, N.; WYNDHAM, J. Psychological, lifestyle and coping contributors to chronic fatigue in shift-worker nurses. *Journal of Advanced Nursing*, v.59, n.3, p.221–232, 2007.

SCHMITT, P.; MANDEL, J.; GUEDJ, M. A comparison of six methods for missing data imputation. *Journal of Biometrics and Biostatistics*, v.6, n.1, p.1–6, 2015.

SERVAES, P.; VERHAGEN, S.; BLEIJENBERG, G. Determinants of chronic fatigue in disease-free breast cancer patients: a cross-sectional study. *Annals of Oncology*. v.13, n.4, p.589–598, 2002.

SOUZA, B. B. *et al.* Lifetime shift work exposure: association with anthropometry, body composition, blood pressure, glucose and heart rate variability. *Occupational and Environmental Medicine*, v.72, n.3, p.208-215, 2015.

SHEN, J.; BOTLY, L. C.; CHUNG, S. A.; GIBBS, A. L.; SABANADZOVIC, S.; SHAPIRO, C. M. T. Fatigue and shift work. *Journal of Sleep Research*, v.15, n.1, p.1–5, 2006.

STERNE, J. A. *et al.* Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *British Medical Journal,* v.338, p.b2393, 2009.

STEKHOVEN, D. J.; BÜHLMANN, P. MissForest - non-parametric missing value imputation for mixed-type data. *Bioinformatics*, v.28, n.1, p.112–118, 2012.

STEKHOVEN, D. J. *missForest: Nonparametric missing value imputation using random forest*. R package version 1.4. https://CRAN.R-project.org/package=missForest, 2013.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological),* v.58, n.1, p.267–2880, 1996.

TOMIC, S.; BRKIC, S.; MARIC, D.; MIKIC, A. N. Lipid and protein oxidation in female patients with chronic fatigue syndrome. *Archives of Medical Science*, v.8, n.5, p.886–892, 2012.

USECHE, S. A.; ORTIZ, V. G.; CENDALES, B. E. Stress-related psychosocial factors at work, fatigue, and risky driving behaviour in bus rapid transport (BRT) drivers. *Accident Analysis and Prevention*. v.104, p.106–114, 2017.

VANWINCKELEN, G.; BLOCKEEL, H. On estimating model accuracy with repeated cross-validation. *21st Belgian-Dutch Conference on Machine Learning*, v.21, p.39–44, 2012.

VAUGHAN, D.; MATT DANCHO, M. *furrr: Apply mapping functions in parallel using futures*. R package version 0.1.0. https://CRAN.R-project.org/package=furrr, 2018.

WALJEE, A. K. *et al*. Comparison of imputation methods for missing laboratory data in medicine. *British Medical Journal*, v.3, n.8, p.e002847, 2013.

WICKHAM, H. *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag, 2016.

WICKHAM, H.; HENRY, L. *tidyr: Easily tidy data with 'spread()' and 'gather()' Functions*. R package version 0.8.3 https://CRAN.R-project.org/package=tidyr, 2019.

WILLIAMS, Y. J.; JANTKE, R. L.; JASON, L. A. Chronic fatigue syndrome: case definitions and diagnostic assessment. *New York State Psychologist*, v.26, n.4, p.41–49, 2014.

WITTEN, I. H.; FRANK, E.; HALL, M. A.; PAL, C. J. *Data mining: Practical machine learning tools and techniques*. London: Morgan Kaufmann. 2016.

ZOU, H.; HASTIE, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (statistical methodology)*, v.67, n.2, p.301–320, 2005.

Received on 08.09.2020

Approved after revised on 08.02.2021