# A Study of Chronic Fatigue and Its Relation to Absenteeism using Stepwise and Elastic-net

Anderson Cristiano Neisse[1], Fernando Luiz Pereira de Oliveira[2], Anderson Castro Soares Oliveira[3], Frederico Rodrigues Borges da Cruz[4], Fausto Aloisio Pedrosa Pimenta[2]

[1] Universidade Federal de Viçosa (UFV), Brazil
[2] Universidade Federal de Ouro Preto (UFOP), Brazil
[3] Universidade Federal de Mato Grosso (UFMT), Brazil
[4] Universidade Federal de Minas Gerais (UFMG), Brazil

E-mail for correspondence: a.neisse@gmail.com

**Abstract:** Chronic Fatigue Syndrome (CFS) is an illness that has been commonly present in clinical practice in the last decades. It is characterized by persistent fatigue, pain, cognitive impairment, and sleep difficulties. The factors that contribute to the CFS development, as studies suggest, are: poor sleep, psychological stress, hormonal dysfunction, nutrient deficiencies, among others. Its development can increase in poor work conditions, such as in shift work in mines, therefore increasing the risk of fatal accidents. A possibly effective tool to prevent the development of CFS is predictive modeling. This study aims to assess the risk of CFS and its relation to absenteeism by means of biochemical and anthropometric variables. A cross-sectional study collected data on 621 shift workers in a mine, measuring 19 variables. After imputing missing data, logistic regression was fitted by four approaches: stepwise, lasso, ridge and elastic-net. Each model was compared between imputed and complete-cases datasets as well as with each other. The stepwise model was chosen for further exploration since the other three approaches did not show performance improvements. Results suggest a lack of discrimination power due to noise that is inherent to the dependent variable's nature. However, significative effect was observed for the LDL, total cholesterol, triglycerides, and sodium on the risk of skipping work.

**Keywords:** Chronic Fatigue; Bioinformatics; Elastic-Net; Logistic Regression.

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

# 1 Motivation

Characterized by more than six months lasting fatigue, pain, cognitive impairment and sleep difficulties, Chronic Fatigue Syndrome (CFS) is an illness that has been common in clinical practice in the last decades (Afari and Buchwald, 2003). Studies indicate factors that contribute to CFS development: poor sleep, psychological stress, hormonal disfunction, nutrient deficiencies, immunological disfunction, infections (Nozaki et al, 2009; Naviaux et al., 2018). In work conditions of risk, the development of CFS can increase the chance of fatal accidents, such as the work on shifts of mines which contains CFS factors. According to Murphy et al. (2011), predictive modelling can be an effective tool in the prevention of CFS. Therefore, this study aims to assess the risk of chronic fatigue by relating biochemical and anthropometric variables with absenteeism of shift workers of a mine.

# 2 Methodology

The data was collected from mine workers on a alternating shifts schedule, i.e., 6 hours of shift followed by 12 hours of rest. The mine's location is on the Inconfidentes region, state of Minas Gerais, Brazil. The data set is composed of 621 individuals with variables for Sex, Age, 8 anthropometric variables and 11 biochemical variables. The dependent variable on the study is a binary factor indicating whether the individual skipped work on the year of 2012. After initial exploratory data analysis, the variables with more than 15% missing cases were discarded. The remaining 17 variables were: Skipped and Sex as binary (male=1); Age, Diastolic and Sistolic Blood pressures as discrete; BMI, Waist-Hip Ratio, Total Body Fat, Visceral Fat, HDL, LDL, Triglycerides, Total Cholesterol, Calcium, Glucose, Sodium and Potassium as continuous. Remaining missing cases were imputed with the KNN algorithm. Two datasets, one with imputed data and one with only complete cases were used in the study for comparison purposes. A proportion of 0.7 was chosen to be the training set. Logistic regression was then fitted to each dataset using four different fitting approaches: (i) Stepwise; (ii) Lasso; (iii) Ridge and (iv) Elastic-Net (Zou and Hastie, 2005). All methods used the area under the receiving operating characteristic (AUROC) curve as optimization metric. The model settings that maximized cross-validated AUROC were then compared using Bootstrapping confidence intervals for AUROC, Sensitivity and specificity on the testing set. The best model then had it's effects explored by Bootstrap applied to the training set.

# 3 Results

A hosmer-lemmeshow calibration test suggests that all models fitted well the data. The confidence intervals for AUROC, Specificity and Sensitivity(FIGURE 1) overlap in all models, suggesting no difference in performance. Also, their AUROC measures included 0.5 on the confidence intervals' range suggesting lack of discrimination power.

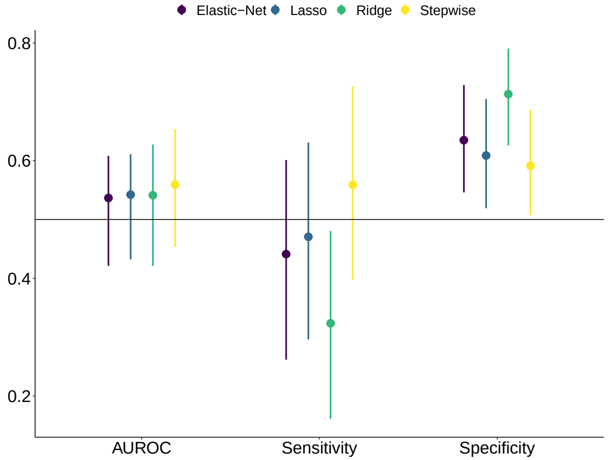

FIGURE 1. Bootstrap (n=1.000) 95% confidence intervals for AUROC, Sentitivity and Specificity.

Elastic-Net regularization approaches did not improve the discrimination. That fact summed to the fact that no model seemed to differ in therms of discrimination(FIGURE 1) led to the choice of the Logistic model for exploration of the variables' effects. The coefficients with its Bootstrapped averages and confidence intervals are shown in the TABLE 1.

TABLE 1. Original regression coefficients and Bootstrap (n = 1.000) estimates.

| | **Coefficients** | **Boot. Avg** | **Boot. CI** |
|---|---|---|---|
| Intercep | -1.3266** | -1.3610 | (-1.6330; -1.1166) |
| AvDBP | -0.1935 | -0.1881 | (-0.4710; 0.0868) |
| LDL | 0.5362. | 0.5680 | (-0.0316; 1.1882) |
| Trig | 0.4911** | 0.5028 | ( 0.1760; 0.8216) |
| Chol | -0.7375* | -0.7740 | (-1.4359; -0.1252) |
| Sodium | -0.4939* | -0.5464 | (-1.1451; -0.0911) |

Note that the data was standardized and that the coefficients are in log-odds scale. The results in TABLE 1 correspond to the KNN-imputed data, the complete case dataset's results were omitted. The results suggest significative effects on Triglyceride, Total Cholesterol and Sodium. An individual that is one standard deviation above the average of Triglyceride levels will have a 0.4911 higher log-odds of Skipping work. As for Total Cholesterol, the results suggest a decrease in log-odds for above-the-average individuals, similar to Sodium.

# 4 Conclusion

Results suggest a lack of discrimination power for all models fitted to the data. Such lack might result noise that is inherent to the dependent variable's nature. Also, all models showed statistically equal performance measures and agreed on variable effects. Exploration on effect done in the Stepwise model significative effects of Triglicerydes, Total Cholesterol and Sodium. The results also suggest the importance of LDL Cholesterol as an factor with correlation with the risk of skipping work.

**Acknowledgments:** 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.

## References

Afari, N., and Buchwald, D. (2003). Chronic fatigue syndrome: a review. *American Journal of Psychiatry*, **160(2)**, 221 – 236.

Murphy, S.M.; Castro, H.K.; Sylvia, M. (2011). Predictive modeling in practice: improving the participant identitication process for care management programs using condition-specific cut points. *Population health management*, **14(4)**, 205 – 210.

Naviaux, R. K., Naviaux, J. C., Li, K., Bright, A. T., . . . and Gordon (2016). Metabolic features of chronic fatigue syndrome. *Proceedings of the National Academy of Sciences*, **113(37)**, 5472 – 5480.

Nozaki, S., Tanaka, M., Mizuno, K., Ataka, S., . . . and Yoshida, K. (2009). Mental and physical fatigue-related biochemical alterations. *Nutrition*, **25(1)**, 51 – 57.

Vanwinckelen, G., and Blockeel, H. (2012). On estimating model accuracy with repeated cross-validation. In: *BeneLearn 2012: Proceedings of the 21st Belgian-Dutch Conference on Machine Learning*, Ghent, Belgium, 39 – 44.

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67(2)**, 301 – 320.